Vol.2019-DPS-177 No.6
2019/2/1

Catalogue Systemにおける パーソナライズサービスのための部分グラフ決定手法

上村 優介[2,a)]　金子 晋丈[1,b)]

概要：Catalogue System は，コンテンツに付与されたグローバルユニークな ID を用いながらユーザがコンテンツ間関係を Catalogue と呼ぶ意味集合のグラフとして表現し，これを自律分散的に管理する機構である．Catalogue System が保持する全体グラフを用いてユーザごとに異なる興味や知識に基づいたコンテンツサービスを構築するに当たり，ユーザにとっての意味境界を全体グラフ内に求め，部分グラフを決定することは重要である．本研究では，全体グラフのスケールフリー性と Catalogue の重複を反映した意味的距離に基づいて部分グラフを決定する手法を提案する．具体的な意味的距離の定式化においては，スケールフリー性を次数差が大きいノード間の意味的距離を短くすること，Catalogue の重複を重複数に反比例して意味的距離を短くすることによって反映させた．定式化した意味的距離を用いて，ユーザが与えた中心ノードを起点とした意味的距離に基づいたグラフ分割システムを構築し，中心ノードの変化に伴い得られた部分グラフがどのように変化するかについて評価した．評価の結果，ユーザの興味の中心となるコンテンツの位置の変化量に応じて意味集合が変化すること，部分グラフ境界における Catalogue の切断数が 30 %程度に抑えられることが明らかになった．

Subgraph Determination Method for Personalized Service in Catalogue System

Abstract: Catalogue system is an autonomous distributed mechanism that manages graphs of meaning sets called Catalogue in which users express relationships between digital contents. In constructing a content service based on different interests and knowledge for each user by using the whole graph held by Catalogue system, it is important to determine the semantic boundary for the user in the whole graph and to determine the subgraph. In this paper, we propose a method to decide subgraph based on semantic distance reflecting scale-free property of whole graph and duplication of Catalogue. In formulating the semantic distance, we shorten the semantic distance between nodes with large degree difference and in inverse proportion to the number of duplications. Using a formalized semantic distance, we construct a graph partitioning system based on the semantic distance starting from the central node given by the user, evaluate how the subgraph obtained with the change of the central node changes. We reveal that the semantic set changes according to the amount of change the position of the center content of interest of the user, and the number of cuts of Catalogue at the subgraph boundary is suppressed to about 30%.

1. はじめに

Catalogue System[1] は，コンテンツに付与されたグローバルユニークな ID を用いながらユーザがコンテンツ間関係を Catalogue と呼ぶ意味集合のグラフとして表現し，これを自律分散的に管理する機構である．Catalogue はコンテンツをノード，関係性をエッジとして意味集合を表現し，Catalogue System が管理する全体グラフの部分グラフになる．Catalogue の全体グラフを利用するユーザは，他のユーザが記述した Catalogue を辿っていくことでコンテンツの関係性からコンテンツを探索する．Catalogue の全体グラフが成熟し多数の意味集合が混在すると，ユーザが関連性を辿ることが困難になりコンテンツの探索を妨げてしまうため，クラスタリングなどの中間処理を加えることで全体グラフを整理した形でユーザに提供することが考えら

[1] 慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
[2] 慶應義塾大学大学院理工学研究科
Graduate School of Science and Technology, Keio University
[a)] simba@inl.ics.keio.ac.jp
[b)] kaneko@inl.ics.keio.ac.jp

れる．しかし，ユーザごとにコンテンツ関係の意味理解は異なるためクラスタリングのような一意に部分グラフを決定する手法は不適切である．ユーザにパーソナライズしたサービスとして，ユーザごとの興味や知識に基づいて全体グラフを分割し部分グラフを提供することが考えられる．このようなサービス実現のためには，ユーザにとっての意味境界を全体グラフ内に求め，部分グラフを決定することが必要である．

本研究では，ユーザにとっての意味境界を決定することを目的とし，問題を以下のように切り分ける．

- ユーザの興味に従って意味境界が決定される
- 得られる部分グラフが Catalogue の意味集合である

ユーザの興味はユーザごとに異なり一般性を保証した意味境界の評価が困難であるため，ユーザの興味の中心ノードを与えられたときに Catalogue の意味集合で全体グラフを分割する課題に取り組む．意味境界の必要条件を以下に示す．

- Catalogue のグラフ特徴を考慮した意味境界を得られていること
- 中心ノードの位置の変化量に応じて結果の意味境界が変化すること
- 境界で切断される Catalogue が少なくなること

これらは，グラフ特徴から意味集合の形成の仕方を抽出したい，ユーザの中心ノードごとに結果が変わってほしい，および，Catalogue の意味集合で境界が形成されるべきであるという要求に基づく条件である．Catalogue の意味集合の形成が Catalogue のグラフ特徴と関係すると仮定すると，グラフ特徴を捉え，意味境界に反映させることが必要である．また，ユーザの興味の中心ノードによって意味境界が決定されるためには少なくともユーザの中心ノードごとに意味境界の結果が変わることが必要である．さらに，得られる部分グラフは Catalogue の意味を反映するべきである．境界で切断される Catalogue の意味は無視されることになるので境界で切断される Catalogue を少なくする必要がある．

このような意味境界を取得するためにまず Catalogue のグラフ特徴をノード間の意味的距離として定式化する．そして，意味的距離に基づいて中心ノードを始点とした意味集合を形成し，意味境界を決定する．得られた意味境界が必要条件を満たしていることを確認するために，

- 中心ノードの位置の変化に対する結果の部分グラフ内のノードの変化率
- 境界で切断されるエッジの割合

を評価した．本研究の貢献は，意味的距離の定式化による，定式化に基づく意味境界の決定手法である．具体的には，Catalogue System の全体グラフがスケールフリー性をもつという特徴とエッジの重複を考慮した手法の提案，および手法が Catalogue の意味集合決定の必要条件を満たしていることを検証した点にある．

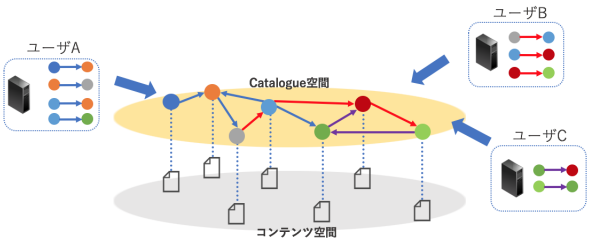

図 1 Catalogue System の全体像
Fig. 1 Overview of Catalogue System.

本稿の構成は以下の通りである．まず，第 2 章で Catalogue System の概要とグラフ特徴について述べ，第 3 章で関連研究を紹介する．第 4 章ではグラフ特徴に基づいた意味的距離の定式化と部分グラフの決定方法について述べる．第 5 章で得られた部分グラフの評価を述べ，第 6 章で本稿をまとめる．

2. Catalogue System

2.1 概要

Catalogue System の全体像を図 1 に示す．Catalogue System では画像・音声・テキストのようなコンテンツファイル空間とコンテンツ間関係を記述する空間を分離する．コンテンツファイルはグローバルユニークな識別子 Global File ID（GFID）を持つ．コンテンツ間関係は，Catalogue と呼ぶ有向グラフによって表現され，Global Catalogue ID（GCID）により一意に作成される．コンテンツの管理者だけでなく，それぞれのユーザが GFID を利用して Catalogue を作成可能であるため，ユーザ自身の視点でコンテンツ間関係が記述される．作成された Catalogue は各ユーザの Catalogue Server に自律分散的に保存管理される．

2.2 グラフ特徴

Catalogue のグラフ特徴を図 2 に示す．ユーザはコンテンツの管理者かどうかに関係なく自由に Catalogue を記述することができる．それぞれのユーザが作成した Catalogue は全体の部分グラフとなるため，様々な意味集合が混在することになる．

また，Catalogue は他のユーザが記述した Catalogue に関係なく自律的に作成・保存されるため複数の Catalogue が重なって全体グラフが構成される．Catalogue の重なりはエッジの重複として表現され全体グラフに反映される．

さらに Catalogue の全体グラフは複雑ネットワークの様相を呈する．複雑ネットワークは SNS やインターネットなどの現実世界のネットワークにみられる性質を有したネットワークモデルである．Catalogue System はユーザが自律的にネットワークを成長させていく複雑ネットワークでスケールフリー性の特徴を有する．スケールフリー性とは

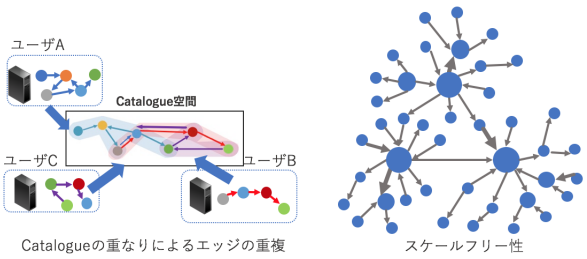

図 2 Catalogue のグラフ特徴
Fig. 2 Characteristics of Catalogue graph.

ノードの次数分布がべき乗則に従うという性質であり，一部の高次数ノードが大多数の低次数ノードに接続することを意味する．

以上より，Catalogue System のグラフ特徴は次のようにまとめられる．

- ノード間の意味の多様性がエッジの重複として表現される
- 一部の高次数ノードに低次数ノードが接続していき全体グラフが構成される

第 4 章では，グラフ特徴をもとにノード間の意味的距離を定式化し意味的境界を決定する．

3. 関連研究

3.1 クラスタリング

クラスタリングはグラフ構造から部分グラフを抽出することでグラフ解析する手法である．切り出された部分グラフをクラスタと呼び，クラスタ内の辺の数に対してクラスタ間の辺の数が少ないほど密度が高いとして，最密になるようにクラスタを決定する．

Girvan-Newman 法 [2], [3] はエッジを切断することでクラスタを決定するアルゴリズムである．全グラフを探索し媒介中心性が高いエッジを切断する．媒介中心性とは，あるノードから別のノードへ到達する経路に含まれやすいエッジの性質である．クラスタ間のエッジが疎になるように分割する場合，あるクラスタから別のクラスタに到達するために，多くのノード間の経路でクラスタ間のエッジが通過され媒介中心性が高くなる．したがって，媒介中心性が高いエッジを全グラフから決定し切断することで，クラスタ境界が得られる．しかし，媒介中心性が高いエッジを算出する計算は効率が悪くグラフの規模に対してスケールしない問題がある．

Newman 法 [4] は Girvan-Newman 法の計算効率を改善したアルゴリズムである．Girvan-Newman 法では媒介中心性の高いエッジをもとにして厳密にクラスタを最密構造にする手法である．一方で Newman 法は各ノードのみで構成されるクラスタを最小単位とし，クラスタの結合結果密度の増加量が最も高い組み合わせでクラスタを結合する．最密になるように計算するのではなく増加量が多くなるように結合することでヒューリスティックにクラスタを形成するため，精度は少し落ちるが計算量を削減することができる．

いずれの方式においても，与えられたグラフに対し密度が高いクラスタの組み合わせが一意に決定されるため，中心ノードを始点としたグラフ分割はできない．また，グラフ構造から意味を考慮して部分グラフを構成する方式は存在しない．

3.2 Local Community Detection

Local Community Detection は SNS や論文の引用関係といった複雑ネットワークにおけるクラスタの抽出が目的である．複雑ネットワークは数百万のノードが接続される巨大なネットワーク構造であり，クラスタリング手法のように全グラフを探索してクラスタを抽出することは困難である．そのため，グラフを部分的に探索することにより周辺のクラスタを解析することが課題である．一般的な Local Community Detection のアルゴリズムはランダムウォークによる Walker の遷移確率で始点ノードの周辺クラスタを抽出する [5], [6], [7], [8], [9]．ほとんどのランダムウォークに基づいたアルゴリズムは単一の Walker によりクラスタ抽出する [5], [6], [7]．Walker の遷移確率が高いノードほど始点ノードへの近接性が高いため始点ノード周辺のクラスタを構成するノードに決定される．この手法では，始点ノード周辺のグラフの密度に着目しているため，グラフ構造の意味について言及していない．また，Walker は始点ノードを中心とした遷移確率を示すとは限らず，中心ノードをもとにした部分グラフ抽出はできていない．

[8] は一定確率 α で Walker を始点ノードに遷移させる手法である．Walker が一定確率で始点ノードに戻るため，始点ノードを中心とした効率的な探索が可能となる．この手法では始点ノードの位置に大きく依存した部分グラフが得られるが，他の単一 Walker によるクラスタ検出手法と同様に，グラフの密度に着目しクラスタを得ることを目的としており，グラフの特徴から意味集合の形成の仕方を抽出することは検討していない．本研究では，グラフ特徴から Catalogue の意味に基づいたノード間距離を定義し，ノード間の意味的距離に基づき部分グラフを得ることを目的とする．

Multi Walker Chain[9] では複数の Walker を協調させながら遷移させる．それぞれの Walker はランダムに遷移しながら一定確率で他の Walker の遷移確率が高いノードに遷移する．平均的には全ての Walker はクラスタ内の媒介中心性が高いノードへの遷移確率が高くなる．このことから [8] では始点ノードを中心とした遷移確率が得られるのに対し，Multi Walker Chain ではクラスタ内の媒介中心性が高いノードを中心とした遷移確率が得られる．したがっ

て，周辺のクラスタを適切に検出することができる．この手法では，始点ノードに依存せず周辺のクラスタを得ることを目的としているため，本研究のように始点ノードをもとにした部分グラフの決定には不適である．また，他の手法と同様にグラフ特徴と意味の関係性には言及していない．

以上より，Local Community Detection の方式もクラスタリング手法と同様に始点ノードによらないクラスタ検出を目的とする．また，意味集合として部分グラフを分割することは考慮していない．本研究では，Catalogue のグラフ構造からノード間の意味的距離を定式化し，それをもとにした意味集合として部分グラフを決定することを目的とする．

3.3 ノード同士の意味的な類似度判定

グラフ上のノード同士が意味的に類似しているか判定する手法は，ノードやエッジに紐付けられた意味情報に着目する手法とグラフ構造に着目する手法に分けられる．前者の手法として Topic Sensitive PageRank[10] が挙げられる．従来の PageRank[11] ではリンク情報のみを利用して Web ページのランキングを行うため，ユーザが求めるトピックに関係があるがリンクが比較的薄い Web ページの影響が小さくなる．Topic Sensitive PageRank では，オフラインでそれぞれの Web ページをトピック分類し，それぞれのトピックごとに PageRank を計算する．最後に各トピックの計算結果を合成することで各トピックに対する Web ページの関連度が示される．これをもとに Web ページのランキングを行うことでよりトピックに関連した Web ページのスコアが高くなるように計算される．しかし，この手法は予め Web ページのトピック分類をする必要があり，数百万のノードが接続する複雑ネットワークでスケールしない．また，Catalogue ではトピックに制約されず，コンテンツ間関係を自由に記述する．ユーザ定義のコンテンツ間関係をトピック分類することは困難である．

後者の手法として SimRank[12], [13], [14] や計算量を削減した ProbeSim[15] が挙げられる．これらの手法では，類似性を比較したい 2 点を決定し，ランダムウォークにより 2 点から Walker が移動を開始する．2 人の Walker が遭遇する確率をもとに 2 点間の類似性を計算する．本研究では，グラフ構造から類似度を含むノード間の意味的な関係性を抽出することを課題とする．また，SimRank では重複なしのグラフに対して類似度を計算するが，本研究では，Catalogue の意味集合をグラフ構造から抽出するために，Catalogue のグラフ特徴を考慮してノード間関係を計算する．さらに，最終的な出力として始点ノードを中心とした部分グラフを得ることを目的とするため，ノード間関係だけでなく部分グラフの決定方法まで言及する．

3.4 パーソナライズサーチ

本研究では，Catalogue のグラフをもとにしたコンテンツ探索をパーソナライズして提供することを目的とするのに対し，パーソナライズサーチ [16] では，パーソナライズなキーワード検索を実現することを目的とする．[16] の手法では，ユーザの Web ページ訪問履歴をもとに，ユーザの興味をトピック分類する．そして，前述した Topic Sensitive PageRank における Web ページのトピック分類とユーザの興味を紐付けることで，ユーザの興味にあった Web ページをランキングする．これは，Topic Sensitive PageRank に基づいているため，3.3 節で示したとおり予め Web ページをトピック分類する必要がある．本研究では，グラフ構造を利用してノード間の意味のつながり決定する．

4. 意味的境界の決定手法

4.1 Catalogue のグラフ特徴に基づいたエッジ切断ポリシ

ノード間の意味的距離を定式化するために，Catalogue のグラフ特徴からどのエッジを切断しやすくするべきかというポリシを決定する．

4.1.1 次数を考慮したエッジ切断ポリシ

Catalogue の全体グラフはスケールフリー性より一部の高次数ノードが大多数の低次数ノードを収容してグラフが生成される．このことから次数の差が大きいノード同士がまとまり意味集合を作りやすいと考えられる．また，高次数ノードが独立した意味集合を作りやすいため，高次数ノード同士は意味境界で分割され別の意味集合に収容されやすい．以上より，次数を考慮したエッジ切断ポリシを以下のように定める．

- 次数の差が大きいノード間エッジであるほど切り離さない
- 高次数ノード間エッジであるほど切り離す

4.1.2 エッジの重複を考慮したエッジ切断ポリシ

エッジが重複するノード間は複数の Catalogue で接続されていることになる．言い換えると，ノード間に複数の意味関係が存在することを意味するため，重複エッジが多いほど意味的なつながりが強いと考えられる．このことからエッジの重複を考慮したエッジ切断ポリシは，重複が多いノード間エッジであるほど残りやすいと定める．

4.1.3 次数とエッジの重複を考慮したエッジ切断ポリシ

高次数ノード同士は独立した意味集合を作りやすい一方で，ノード間のエッジが重複しやすい．有名なコンテンツは高次数になりやすくそれ自体が独立した意味集合を作りやすいが，有名コンテンツ同士の間に Catalogue が作られやすいためである．次数のみを考慮した場合，ほとんどのエッジが有名コンテンツ同士の間に引かれている場合でも切断されやすくなる．エッジの重複のみを考慮した場合，有名コンテンツ同士とそれ以外の場合の差別化ができな

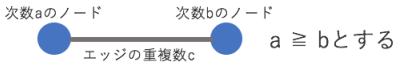

図 3 定式化するノード間エッジ
Fig. 3 Edge between nodes to be formulated.

い．そこで次数とエッジの重複両方を加味することにより重複エッジが多い有名コンテンツ同士を切断されにくくしながらスケールフリー性を考慮することができる．

4.2 エッジ切断ポリシごとの意味的距離の定式化

前節で述べた 3 つのエッジ切断ポリシごとにノード間の意味的距離を定式化する．定式化するノード間エッジを図 3 に示す．

4.2.1 次数を考慮した意味的距離

まず，次数の差が大きいほどノード間の距離を近くするために次数の差の逆数をとる．

$$\left|\frac{1}{a-b}\right| \tag{1}$$

そして，高次数ノード間であるほど距離を遠くするために次数の逆数をとってから差を求める．

$$\left|\frac{1}{\frac{1}{a}-\frac{1}{b}}\right| \tag{2}$$

4.2.2 エッジの重複を考慮した意味的距離

エッジの重複を考慮する場合，ノード間のエッジ数がそのまま距離に反映される．重複エッジ数が多いほど距離を近くするために重複エッジ数の逆数をとる．

$$\frac{1}{c} \tag{3}$$

4.2.3 次数とエッジの重複を考慮した意味的距離

次数の差とエッジの重複を同時に取り扱うために，次数を考慮した (2) 式にエッジの重複の切断ポリシを組み込むことを考える．(2) 式は次数の差が大きいほど距離が近くなるように計算されるため，エッジの重複が多いほど次数の差が大きくなるように次数のカウント方法を修正する．まずノードの次数を構成しているエッジをノード間エッジとノード間でないエッジに分離する．ノード間エッジはエッジの重複に貢献しているため，ノード間エッジを次数が高い方のノードにのみカウントすることによりエッジの重複を次数の差として計算できる．

$$\left|\frac{1}{\frac{1}{a}-\frac{1}{b-c}}\right| \tag{4}$$

この式はエッジの重複を次数の差の式に集約できることの他に次数が同じノード間での距離が計算できるという利点も有する．(2) 式では次数が同じノード間の距離は全て正の無限大になるため，高次数ノード同士ほど距離を遠くするというポリシが反映されない．ノード間エッジを片方のノードのみにカウントすることにより必ず次数差が発生するため次数を考慮した切断ポリシが正しく反映される．

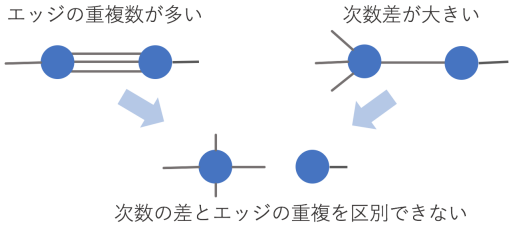

図 4 (4) 式の問題点
Fig. 4 Problem of formula(4).

しかし，(4) 式はノード間エッジとノード間でないエッジを等価に扱うため次数の差とエッジの重複を区別して距離に反映することができない．具体例を図 4 に示す．このような問題を解決するためにエッジの重複数と次数の差の重心を決定する重み α を導入する．ノード間のエッジに $(1-\alpha)$，ノード間でないエッジに α の重みをつけることでこれを実現する．

$$\left|\frac{1}{\frac{1}{\alpha(a-c)+(1-\alpha)c}-\frac{1}{\alpha(b-c)}}\right| \quad (0<\alpha<1) \tag{5}$$

4.3 部分グラフの決定

部分グラフの決定手法の計算の流れを図 5 に示す．

- **STEP1: ユーザの興味の中心ノードの決定**
  ユーザの興味や知識の中心となるノードを選出する．今回は中心ノードの選出方法については議論せず，ユーザによって中心ノードが与えられたとして部分グラフを決定する．
- **STEP2: 直接つながったエッジ間の意味的距離の計算**
  前節で定義した意味的距離の式に基づいて各ノード間のエッジについて距離を計算する．
- **STEP3: 各中心ノードからその他のノードへの最短距離の計算**
  各ノードがどの中心ノードに最も近いかを計算する．Dijkstra 法を用いて各中心ノードから他のノードへの最短距離を計算している．
- **STEP4: 部分グラフの意味境界の決定**
  STEP3 で求めた最短距離をもとに各ノードがどの中心ノードに最も近いかを判定する．一番近い中心ノードの部分グラフにノードを収容していくことで部分グラフを形成する．全てのノードを判定し得られた部分グラフの境界部分が意味境界になる．

5. 評価

5.1 評価概要

得られた部分グラフが Catalogue の意味集合になっていることを確認するために第 1 章で述べた意味境界の必要条

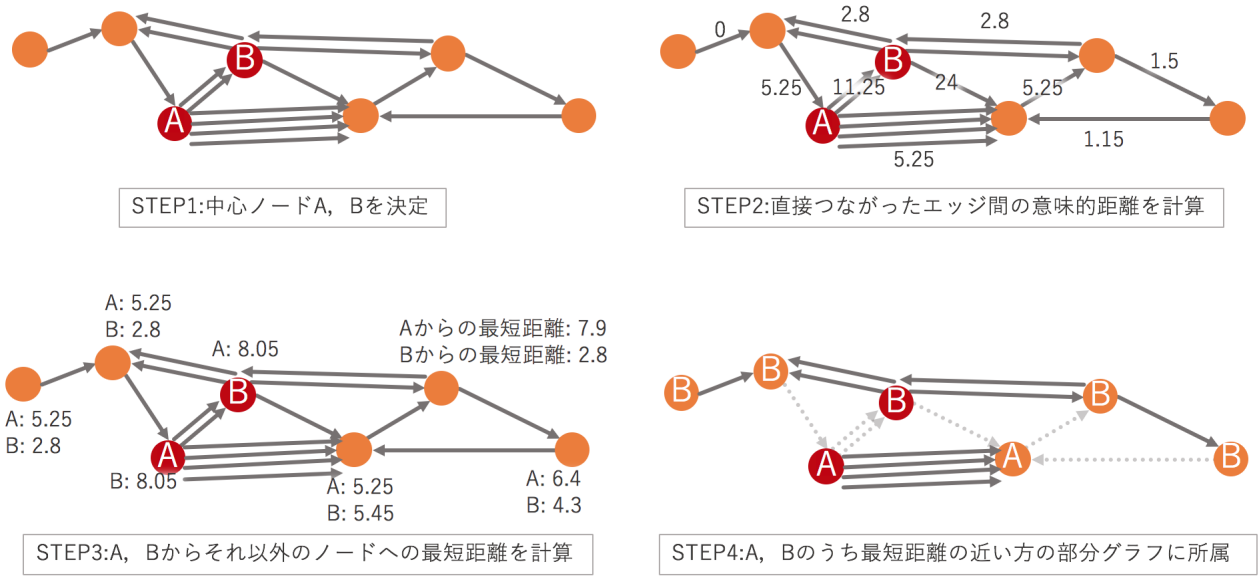

図 5 部分グラフ決定手法の計算の流れ
Fig. 5 Calculation steps of subgraph determination method.

件に基づいて評価する．まず，ユーザの中心ノードを始点とした部分グラフが得られることを評価するために，中心ノードの位置が変化したときに得られる部分グラフ内のノードがどれだけ変化するかを明らかにする．次に，ユーザが作成した Catalogue を反映しているかどうかを評価するために，意味境界で切断されているエッジの割合を算出する．1 本のエッジは 1Catalogue 内に表現されるため，境界で切断される延べ Catalogue 数の割合が算出できる．これにより得られた部分グラフが Catalogue の意味集合として形成されているかどうかを明らかにする．

5.2 評価環境

5.2.1 使用したグラフ

Catalogue のグラフはスケールフリー性を持つため，スケールフリー性を保証する Barabási-Albert モデル [17] をグラフに使用し，[18] のアルゴリズムを参考にして作成した．全ノード数はユーザが一度に閲覧するノード数として 100 個に設定した．スケールフリー性により全グラフの規模が変わっても傾向は変わらないと想定できるため全ノード数は固定して評価する．

5.2.2 比較した意味的距離の計算手法

グラフ特徴の意味的距離への反映の仕方と得られる部分グラフの関係性を明らかにするために，以下の 3 つのエッジ切断ポリシごとの意味的距離の式を用いて意味境界を決定し比較した．

次数のみ ノードの次数差のみを考慮した意味的距離
エッジのみ エッジの重複数のみを考慮した意味的距離
次数とエッジ ノードの次数とエッジの重複を考慮した意味的距離 ($\alpha = 0.3$)

5.3 中心ノードの変化に対する部分グラフの変化率

まず中心ノード数を 5 つとし，そのうち 4 つを固定した．残りの 1 つがノード A のときの部分グラフを S とし，ノード A から n hop 離れたノード A' のときの部分グラフを S' としたときの S と S' 内のノード一致率により中心ノードの変化に対する部分グラフの変化率を算出した．ノード一致率の指標として F 値を利用した．F 値は部分グラフ S，S' 内における S と S' の重複ノードの割合の調和平均であり (6) 式で計算される．

$$F(S', S) = 2 \cdot \frac{prec(S', S) \times rec(S', S)}{prec(S', S) + rec(S', S)} \tag{6}$$

$$prec(S', S) = \frac{|S' \cap S|}{|S'|} \qquad rec(S', S) = \frac{|S' \cap S|}{|S|}$$

F 値は S と S' が完全に一致する場合 1 に，全く一致しない場合 0 になり重複ノードの割合が表現される．

ノード A とノード A' の間の hop 数が 1〜5 まで変化したときの F 値の変化を図 6 に示す．横軸はノード A とノード A' の間の hop 数，縦軸は各手法における F 値を示す．中心ノード間の hop 数が 1 の部分に着目すると，次数とエッジ，次数のみの手法が 0.45 付近で hop 数の増加に従って F 値が減少しているのに対し，エッジのみの手法では hop 数 1 から 2 へ変化したときに F 値が増加している．つまり，エッジのみの手法では中心ノードの変化を反映して部分グラフが変化できていないことがわかる．中心ノード間が 1hop の場合スケールフリー性の影響が大きいことが

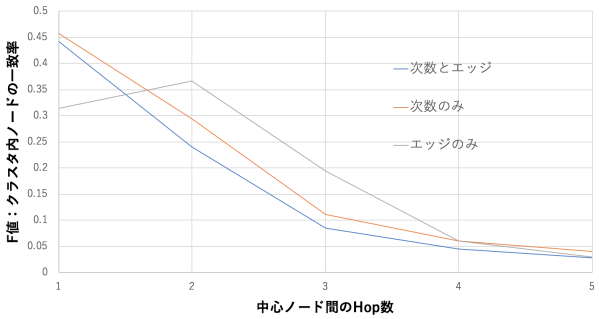

図 6 中心ノードを変えたときの F 値の変化率
Fig. 6 Change rate of F value when center node is changed.

表 1 中心ノード間の次数差
Table 1 Degree difference between center nodes.

Hop 数	1	2	3	4	5
次数差	13	3	2	1	0

原因として考えられる．Catalogue のグラフではスケールフリー性により高次数ノードと低次数ノードが 1hop でつながっている．このノードがノード A とノード A' に選ばれた場合次数差を考慮している方式では中心ノード間の距離が小さくなるため S と S' が一致しやすくなる．実際に中心ノード間の次数差を比較すると表 1 のように 1hop では次数差が 13 であるのに対し 2hop では 3 と大きく減少している．このことから 1hop の場合スケールフリー性が F 値に影響を与えている可能性が高いことがわかる．

また，次数とエッジの手法と次数のみの手法を比較すると最大 0.05 程度次数とエッジの手法のほうが F 値が低いことが読み取れるが，ユーザの興味の境界としての有意差であるかどうかはこの結果だけでは評価できず，汎用的な実データを複数用意し検証する必要がある．一方で，評価の結果がデータセットに大きく依存するため今回は評価の対象から外している．

5.4 意味境界でのエッジ切断割合

次に意味境界でのエッジ切断割合を評価する．意味境界で切断されているエッジ数と全エッジ数の商を取ることで算出した．中心ノード数が増加するとそれに応じて境界が複雑化し切断されるエッジの割合が増えることが想定されるため，中心ノード数を 2〜20 まで増加させたときにエッジの割合がどれだけ増加するかを手法ごとに評価した．結果を図 7 に示す．

次数とエッジの手法とエッジのみの手法を比較すると約 50%程度次数とエッジの手法の方が切断割合が少ないことが読み取れる．また，次数とエッジの手法と次数のみの手法を比較しても次数とエッジの手法の方が最大約 10%程度少なく，30%程度の切断割合に抑えられていることがわかる．以上より，中心ノードに対する F 値の変化率と合わせて次数とエッジを考慮した手法が Catalogue の意味集合として部分グラフを決定する手法としての必要条件を満たしていると結論付けられる．

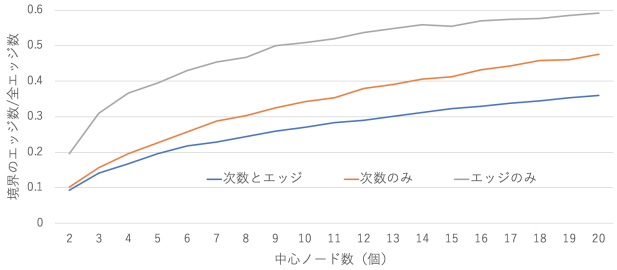

図 7 中心ノード数に対する境界のエッジ切断割合の変化
Fig. 7 Changes in the edge cutting rate of the boundary with respect to the number of center nodes.

6. おわりに

Catalogue System におけるコンテンツサービスをよりユーザにパーソナライズし提供するための部分グラフ決定手法を検討してきた．Catalogue のグラフがスケールフリー性を持つこと，Catalogue の重なりによりエッジが重複することからノード間の意味的距離を定式化し，ユーザの中心ノードを始点とした部分グラフを決定する手法を提案した．得られた部分グラフが Catalogue の意味集合として分割できているかどうかを評価するために，中心ノードの変化量に応じた部分グラフ内のノードの変化率と意味境界でのエッジ切断割合を算出した．評価の結果，次数を考慮することで中心ノードの変化に応じた部分グラフの変化が得られること，次数とエッジを考慮する方式が最も境界でのエッジ切断割合を抑えることに貢献することを明らかにした．

今後の課題として，ユーザの興味で境界が形成されていることの評価と Catalogue が有向グラフであることを反映した意味的距離の定式化が挙げられる．ユーザにとっての意味境界を決定するための十分条件として，ユーザの興味に従って意味境界が決定されることと，得られる部分グラフが Catalogue の意味集合であることを挙げた．本研究では，この内 Catalogue の意味集合で全体グラフを部分グラフに分割する課題に取り組んだ．ゆえに，十分条件を満たすためには提案した手法がユーザの興味に従って意味境界を決定することを明らかにすることが必要である．ユーザの興味で切られていることは，実データでの繰り返しの検証により複数のデータセットパターンに対し本研究の手法が有効であるとする必要がある．現状，Catalogue System を実運用するまでには至っておらず，実データのセットを検証に十分な量用意することが困難であるため今後の課題である．また，有向グラフの向きはユーザの Catalogue 作成ポリシに依存する．例えば美術作品を作成した人はモナリザのような有名なコンテンツとエッジを結ぶことが考え

られるがどちら方向にエッジを引くかはそのユーザがどうCatalogueを作成するかに依存する．その傾向を判断し意味的距離に組み込むためには実データでの傾向の分析が必要である．

参考文献

[1] Miyashita, Y., Ishikawa, H., Teraoka, F. and Kaneko, K.: Catalogue: Graph representation of file relations for a globally distributed environment, *Proceedings of the ACM Symposium on Applied Computing*, Vol. 13-17-April-2015, Association for Computing Machinery, pp. 806–809 (online), DOI: 10.1145/2695664.2696047 (2015).

[2] Girvan, M. and Newman, M. E.: Community structure in social and biological networks, *proc natl acad sci*, Vol. 99, pp. 7821–7826 (2001).

[3] Girvan, M. and Newman, M. E.: Finding and Evaluating Community Structure in Networks, *Physical review. E, Statistical, nonlinear, and soft matter physics*, Vol. 69, p. 026113 (online), DOI: 10.1103/PhysRevE.69.026113 (2004).

[4] Newman, M. E.: Fast algorithm for detecting community structure in networks, *Physical review E*, Vol. 69, No. 6, p. 066133 (2004).

[5] Andersen, R., Chung, F. and Lang, K.: Local Graph Partitioning using PageRank Vectors, *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 475–486 (online), DOI: 10.1109/FOCS.2006.44 (2006).

[6] Kloumann, I. and Kleinberg, J.: Community Membership Identification from Small Seed Sets, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, New York, NY, USA, ACM, pp. 1366–1375 (online), DOI: 10.1145/2623330.2623621 (2014).

[7] Wu, Y., Jin, R., Li, J. and Zhang, X.: Robust Local Community Detection: On Free Rider Effect and Its Elimination, *Proc. VLDB Endow.*, Vol. 8, No. 7, pp. 798–809 (online), DOI: 10.14778/2752939.2752948 (2015).

[8] Tong, H., Faloutsos, C. and Pan, J.: Fast Random Walk with Restart and Its Applications, *Sixth International Conference on Data Mining (ICDM'06)*, pp. 613–622 (online), DOI: 10.1109/ICDM.2006.70 (2006).

[9] Bian, Y., Ni, J., Cheng, W. and Zhang, X.: Many Heads are Better than One: Local Community Detection by the Multi-walker Chain, *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 21–30 (online), DOI: 10.1109/ICDM.2017.11 (2017).

[10] Haveliwala, T.: Topic-sensitive PageRank, *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, New York, NY, USA, ACM, pp. 517–526 (online), DOI: 10.1145/511446.511513 (2002).

[11] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web., Technical Report 1999-66, Stanford InfoLab (1999).

[12] Jeh, G. and Widom, J.: SimRank: A Measure of Structural-context Similarity, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, New York, NY, USA, ACM, pp. 538–543 (online), DOI: 10.1145/775047.775126 (2002).

[13] Tian, B. and Xiao, X.: SLING: A Near-Optimal Index Structure for SimRank, *CoRR*, Vol. abs/1604.04185 (online), available from ⟨http://arxiv.org/abs/1604.04185⟩ (2016).

[14] Shao, Y., Cui, B., Chen, L., Liu, M. and Xie, X.: An Efficient Similarity Search Framework for SimRank over Large Dynamic Graphs, *Proc. VLDB Endow.*, Vol. 8, No. 8, pp. 838–849 (online), DOI: 10.14778/2757807.2757809 (2015).

[15] Liu, Y., Zheng, B., He, X., Wei, Z., Xiao, X., Zheng, K. and Lu, J.: Probesim: Scalable Single-source and Top-k Simrank Computations on Dynamic Graphs, *Proc. VLDB Endow.*, Vol. 11, No. 1, pp. 14–26 (online), DOI: 10.14778/3151113.3151115 (2017).

[16] Qiu, F. and Cho, J.: Automatic Identification of User Interest for Personalized Search, *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, New York, NY, USA, ACM, pp. 727–736 (online), DOI: 10.1145/1135777.1135883 (2006).

[17] Barabási, A.-L. and Albert, R.: Emergence of Scaling in Random Networks, *Science*, Vol. 286, No. 5439, pp. 509–512 (online), DOI: 10.1126/science.286.5439.509 (1999).

[18] Ali, H., Sadegh, N., Behrooz, M. and Qiang, Q.: ROLL: Fast In-Memory Generation of Gigantic Scale-free Networks, *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, New York, NY, USA, ACM, pp. 1829–1842 (online), DOI: 10.1145/2882903.2882964 (2016).