卒業論文

不完全なサービスネットワークを用いたサービスソフトクラスタリング

指導教官　　村上 陽平 教授

立命館大学　情報理工学部
先端社会デザインコース　4回生
2600200383-9

松本　賢司

2023年度（秋学期）卒業研究3（CH）
令和6年1月31日

不完全なサービスネットワークを用いた サービスソフトクラスタリング

松本　賢司

内容梗概

複数の Web サービスを連携させて作成されたサービスのことを複合サービスという．複合サービスにより，個々に提供される Web サービスよりも拡張した機能を提供することができる．現在，多種多様な Web サービスが増えており，それらを組み合わせて新たな複合サービスが構築されている．しかし，Web サービスの利用に関する統計より，最大 85.6%の Web サービスが複合サービスに利用されていない．多くの Web サービスからユーザが必要に応じて適切な Web サービスを発見できるように，機能に応じて Web サービスをクラスタリングする必要がある．この問題に対して，Web サービス記述ファイル(以下，WSDL ドキュメントと呼ぶ)やサービス説明文をテキストマイニングし，機能ごとにクラスタリングする研究がある．しかしながら，この手法ではテキストの記述内容に大きく依存するため，Web サービス提供者の命令規則による影響を受けやすい．また，複数の機能を持つ Web サービスが単一の機能としてクラスタリングされるため，複数の機能を持つ Web サービスの多様性が適切に反映されない．

そこで，本研究では，Web サービス，提供者，利用者間の関係を表したネットワークを用いて機能ごとにソフトクラスタリングを行う．具体的には，同一の複合サービスに組み合わされた Web サービス間の Web サービス連携関係，Web サービスを提供している提供者とその Web サービスを利用している利用者の Web サービス提供利用関係をそれぞれグラフで表し，グラフ埋め込みを用いて Web サービスの特徴ベクトルを生成し，一つの Web サービスを複数の機能分類にクラスタリング可能にするソフトクラスタリングを行う．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

不完全データの補完

Web サービス間の結びつきを適切に表現するために，不完全なデータの Web サービスも考慮してネットワークを構築し，全体のネットワークを適切に評価する必要がある．

ソフトクラスタリングの適用方法

ソフトクラスタリングした各 Web サービスは全てのクラスタに所属するため，適切な機能に分類するために適切な閾値を定める必要がある．

一つ目の課題に対しては，Web サービス提供利用関係において不完全なデータの Web サービスも含めてネットワークを構築し，サンプリングを行う．具体的には，Web サービス提供利用関係では，提供者データが欠如した Web サービスに個別の ID を割り当て，提供者・利用者をノード，Web サービスの利用をエッジとした．これらのネットワークを構築したのちに探索を行い，サンプリングする．得られた各サンプリングから得られたベクトルを機能データが欠如してない Web サービスを対象に連結する．加えて，提供者データが欠如しており，個別の ID を割り当てた Web サービスについても機能データが欠損していなければ連結する．その後，連結したベクトルを用いてクラスタリングを行う．

二つ目の課題に対しては，ソフトクラスタリングとしてクラスタ中心との距離に基づいて所属度が計算される Fuzzy C-means，データを複数のガウス分布の組み合わせとして，各 Web サービスが分布の所属度として計算される Gaussian Mixture Model を使用する．実行後に各 Web サービスのクラスタ所属数は，その Web サービスに人手で付与された機能数と同様になるように閾値を設定した．

提案手法によって生成された Web サービスのクラスタと，機能ごとに人手で Web サービスを分類した正解クラスタと比較することで，生成したクラスタがサービスの機能ごとにクラスタリングされていたかを purity 値と F 値を用いて評価し，提案手法の有効性を検証した．本研究の貢献は以下のとおりである．

不完全データの補完

不完全なデータを含めた Web サービス連携関係と Web サービス提供利用関係のネットワークを構築し，サンプリングした．各サンプリングから得られたベクトルを，完全な機能データを持つサービス，また，提供者データが欠けているが機能データが完全なサービスも連結することでサービスの機能ごとのハードクラスタリングの精度がテキスト埋め込み手法と比較し，F 値が約 10%向上した．

ソフトクラスタリングの適用方法

ネットワークのサンプリング結果をクラスタリング後，人手で付与した Web サービスの機能数に基づいて閾値を設定することで，複数の機能を持つ Web サービスのうち約 32%を少なくとも 2 つの機能において正確に分類できた．

Service Software Clustering with Incomplete Service Networks

Kenji Matsumoto

Abstract

Composite services, which are created by linking multiple web services, provide richer functionality than individual services. New composite services are being formed because of the spread of diverse web services. However, statistics show that up to 85.6% of web services are left unused in composite services. Effective clustering of web services by function is essential for user discovery. Current research involves text mining of WSDL documents and service descriptions for feature clustering. However, these methods, which rely on textual descriptions, are limited by the naming conventions of service providers, and fail to capture the diversity of multi-functional web services, which are often grouped under a single feature.

Therefore, in this study, soft clustering was performed for each function using a network that represents the relationship among Web services, providers, and users. Specifically, we represented the service coordination relationship that has been combined with the same composite service and the service provider/user relationship between service providers and users who use the service in graphs, respectively, and generated feature vectors of the nodes using graph embedding to perform soft clustering. Soft clustering was performed to enable the clustering of services into multiple functions. In the implementation of this method, the following two issues needed to be addressed.

Completion of incomplete data

To properly represent the connections between services, it is necessary to construct a network that also considers services with incomplete data and to properly evaluate the overall network.

How to apply soft clustering

Soft clustered services belong to all clusters and need to be categorized into. appropriate functions by defining appropriate thresholds.

For the first task, the network was constructed and sampled, including web services with incomplete data in the web service provision-user relationship. In

the provider-user relationship, the providers and users were regarded as nodes and the service usage as edges, and IDs were assigned to the providers with missing data. After network construction, retrieval, and sampling are performed. As a result, the vectors were linked for web services with no missing functional data, and similarly for cases with missing provider data but complete functional data.

For the second task, Fuzzy C-means, where affiliation was calculated based on the distance to the cluster center as soft clustering, and the Gaussian Mixture Model, where the data was a combination of several Gaussian distributions, and each web service is calculated as a distribution affiliation. The data was then run as a cluster of web services. A threshold was set so that the number of cluster affiliations for each web service after the run was the same as the number of functions manually assigned to that web service.

The effectiveness of the proposed method was tested by comparing the generated web service clusters with the correct clusters where services were manually classified by function. The evaluation was carried out using purity and F-values. The contributions of this study are as follows.

Completion of incomplete data

A web service provision and use relationship and a web service linkage relationship network with incomplete data were constructed and sampled. By linking the vectors obtained from each sampling to services with complete functional data and services with missing provider data but complete functional data, the accuracy of hard clustering by service function was improved by approximately 10% in F value compared to the text embedding method.

How to apply soft clustering

After clustering the network sampling results, we were able to accurately classify approximately 32% of the Web services with multiple functions in at least two functions by setting a threshold based on the number of Web service functions manually assigned.

目次

第1章 はじめに

近年，Web サービスの分野において，サービスコンピューティングの分野が発達している．サービスコンピューティングとは，迅速かつ柔軟な新サービス開発を目指す技術であり，多種多様な Web サービスを目的に応じて連携させ，複合サービスを構築できる．例えば，地理情報を提供する Web サービスと天気情報を提供する Web サービスを連携させることで，地図上に各地の天気情報を表示するする複合サービスの構築が可能になる．このように，個々に提供される Web サービスでは実現できない拡張機能を提供することができる．しかし，Web サービスの利用に関する統計より，最大 85.6%の Web サービスが複合サービスに利用されていない[1]．したがって，多くの Web サービスからユーザが必要に応じて適切な Web サービスを安易に発見できるように，機能に応じて Web サービスをクラスタリングする必要がある．この問題に対して，Web サービス記述ファイル(以下，WSDL ドキュメント)やサービス説明文をテキストマイニングし，機能ごとにクラスタリングする研究がある．しかしながら，この手法ではテキストの記述内容に大きく依存するため，Web サービス提供者の命令規則による影響を受けやすい．また，複数の機能を持つ Web サービスが単一の機能としてクラスタリングされるため，複数の機能を持つ Web サービスの多様性が適切に反映されない．

そこで，本研究では，Web サービス，提供者，利用者間の関係を表したネットワークを用いて機能ごとにソフトクラスタリングを行う．具体的には，同一の複合サービスに組み合わされた Web サービス間の Web サービス連携関係，Web サービスを提供している提供者とその Web サービスを利用している利用者の Web サービス提供利用関係をそれぞれグラフで表し，グラフ埋め込みを用いて Web サービスの特徴ベクトルを生成し，一つの Web サービスを複数の機能分類にクラスタリング可能にするソフトクラスタリングを行う．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

不完全データの補完

Web サービス間の結びつきを適切に表現するために，不完全なデータの Web サービスも考慮してネットワークを構築し，全体のネットワークを適切に評価する必要がある．

ソフトクラスタリングの適用方法

ソフトクラスタリングした各 Web サービスは全てのクラスタに所属するた

め，適切な機能に分類するために適切な閾値を定める必要がある．

以下，本論文では，２章では従来の研究のクラスタリング手法として，インターフェイス情報に基づくクラスタリング，サービス説明文に基づくクラスタリングについて説明する．次に，３章では本研究で用いたネットワークの解説を行う．続いて 4 章では，ノードの分散表現に基づくクラスタリングであるグラフ埋め込み技術について説明を行い，ネットワークのサンプリング手法やグラフ埋め込みで用いる Skip-gram-model について説明する． 5 章では，ソフトクラスタリングの手法として，k-menas，Fuzzy C-means，Gaussian Mixture Model について説明する．6 章では３章と４章，５章で説明を行ったクラスタリング手法に対する評価を行い，7 章にて考察を行う．最後に，8 章にて今後の展望や課題について述べて結論とする．

第2章 サービスクラスタリング

本章では，既存の類似機能に基づく Web サービスのクラスタリング手法について説明する．サービスクラスタリングは，複合サービスを構築する際に，その構築要素の候補となるサービスを発見したいときに有用である．

複合サービスの構築事例として図 1 の Uber Eats の開発事例を用いる．Uber Eats とはユーザがレストランから食品をオンラインで注文し，その食品を配達するサービスである．このサービスはユーザが食品の注文を確定する際に，決済の API である Stripe が利用される．Stripe を通じてユーザは簡単に多数の支払い方法を選択することができる．支払い完了後，ユーザは地図と位置情報の API である Google Map を利用して，注文した商品の位置情報と到着時間をリアルタイムで追跡することが可能になる．このように複合サービスを構築する際，必要な機能のサービスを発見する必要がある．しかしながら，オンライン上に存在する膨大な数の Web サービスから，ユーザが求める機能を持つサービスを発見するのは非常に困難である．Web サービスの利用に関する統計より，最大 85.6% の Web サービスが複合サービスに利用されていないことからも，多くの有用な Web サービスが見過ごされている可能性が高い．このような場合，Web サービスを自動的に機能分類することでユーザが求める機能を持つ Web サービスの特定を簡略化することができる． Web サービスを自動的に機能分類する上で，ク

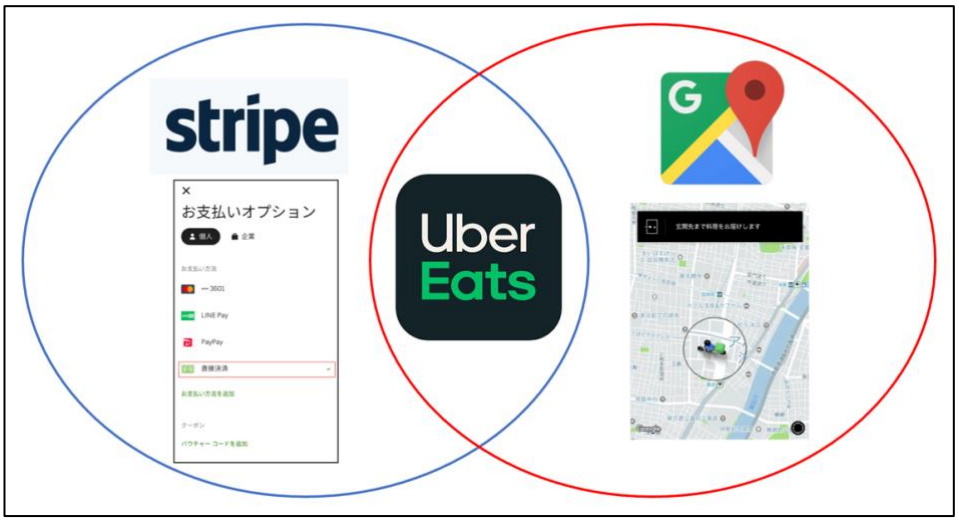

図 1 ：複合サービスの例

[1]https://www.ubereats.com/jp

ラスタリングは非常に効率的なアプローチの 1 つである．本章では，このアプローチに関する研究として，インタフェース情報に基づく手法とサービス説明文に基づく手法を説明する．

2.1 インターフェースに基づくクラスタリング

インターフェースに基づくクラスタリング手法において，ユーザが推薦される Web サービスの特徴は，ユーザが入力した Web サービスの機能を表す単語に類似するものである．そこで，Elgazzar らは，WSDL ドキュメントを基に，機能的に類似したグループにクラスター化し，ユーザのクエリに対応する Web サービスを推奨する手法を提案した[2]．具体的には，以下の 2 ステップである．

1. WSDL ドキュメントからクラスターを作成
2. クラスターからユーザの要求する Web サービスを推奨

1 つ目のステップについて説明する．はじめに，WSDL ドキュメントの WSDL コンテンツから，その Web サービスの機能性を表す単語を抽出する．次に，WSDL タイプ，WSDL メッセージ，WSDL ポート，WSDL サービス名から特徴を抽出する．これら 5 つの特徴を結合して，Web サービスを機能的に類似したグループにクラスター化する．

2 つ目のステップについて説明する．はじめにユーザがサービス検索エンジンに入力された目的の単語でクエリを行う．次に，1 つ目のステップで作成されたクラスターとクエリを意味的に一致させる．例えば，「Cheap」と「Inexpensive」のように意味が同じだが，単語が違うものを同一に一致させるための処理である．最後に，入力された目的の単語を満たす最も慣例性の高い Web サービスをユーザに返す．

2.2 サービス説明文に基づくクラスタリング

Min らは，Web サービスを Word2vec によって拡張された LDA モデルを用いて，機能別にクラスタリングする手法を提案した[3]．LDA モデルとは文書のトピックを推定するモデルで，文章の分類に利用される手法である．従来の WSDL ドキュメントから特徴ベクトルを抽出し，機能別にクラスタリングする手法では，WSDL ドキュメントに記述されている単語が限られていることから，意味的な関係を捉えることが困難であるため，クラスタリング精度が低い問題がある．そこで，サービス説明文と LDA モデルを用いて，Web サービスを機能

別にクラスタリングを行う．具体的には，Word2vec からサービス説明文から単語ベクトルを生成する．次に k-means++アルゴリズムなどを用いて，単語をベクトルの類似性に基づいてクラスタリングする．その後，クラスタ情報を利用して LDA モデルの訓練を訓練する．最後に，k-means を用いて Web サービスを異なる機能にクラスタリングするか，潜在的なトピックに従って，同じトピックに属する Web サービスを一緒にクラスタリングすることができる．

第3章 ネットワークモデルの実装

本章では，全体のネットワークモデルからモデルを部分化し，具体的なネットワークの実装について紹介する．

3.1 全体のネットワークモデル

本研究で用いる複合サービスのモデルと提供者・利用者のモデルを説明する．本研究で用いるデータモデルを図 2 に示す．Web サービスは，1 つの提供者が存在し，複数の利用者がいる．また，複合サービスに連携している場合がある．各提供者は 1 つ以上の Web サービスを提供しており，複数の Web サービスを提供している場合がある．複合サービスは複数の Web サービスを連携している．このデータモデルから，複合サービスと Web サービス，Web サービスの提供者・利用者のモデルを抜き出し，各ネットワークを構築する．

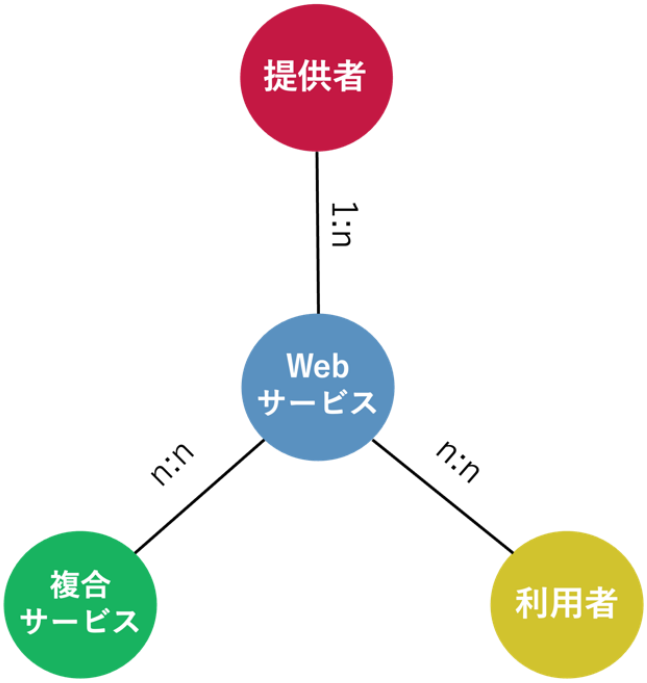

図 2：全体のデータモデル

3.1.1 複合サービスのモデル

複合サービスのモデルとして，複合サービスと Web サービスの関係を用いる．複合サービスは複数の Web サービスが連携されて構築されている．よって，複

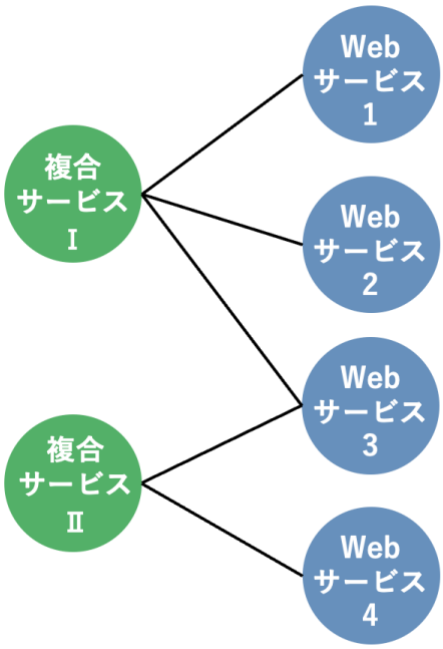

図 3：複合サービスと Web サービスデータモデル

合サービスとその複合サービスに連携された Web サービスを連携関係で繋げることで，図 3 のモデルとなる．このデータモデルを用いて，後述のネットワークを作成する．

3.1.2 提供者・利用者のモデル

提供者・利用者のモデルとして，Web サービスの提供者と利用者の関係を用いる．提供者とは，Web サービスを配布している提供者である．また，利用者とは，提供者が配布している Web サービスを利用している利用者である．よって，提供者とその提供者から配布された Web サービスを提供関係，Web サービスとその Web サービスを利用している利用者を利用関係で繋げることで，図 4 のモデルになる．このデータモデルを用いて，後述のネットワークを作成する．

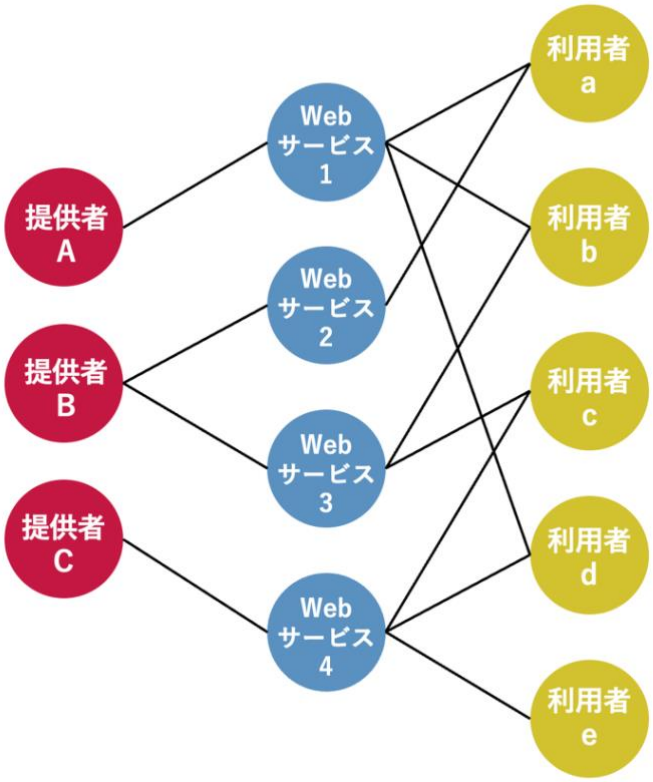

図 4：提供者と利用者データモデル

3.2 Web サービス連携関係ネットワーク

複合サービスに組み合わされている Web サービスを用いて，複合サービスによる連携の依存グラフを作成する．複合サービスの例として，図 3(3.1.1 節参照)の複合サービス Ⅰ，Ⅱを示す．複合サービス Ⅰ には，Web サービス 1，2，3 が連携されている．複合サービスⅡには，Web サービス 3，4 が連携されている．このような複合サービスから，図 5 のグラフを作成する．このグラフでは，複合サービス Ⅰ は，Web サービス 1，2，3 が連携されたことから，3 つのノードを生成し，エッジで繋いでいる．また，複合サービスⅡは，Web サービス 3，4 が連携されていたことから，新しく Web サービスノード 4 を生成し，Web サービスノード 3 とエッジで繋ぐ．

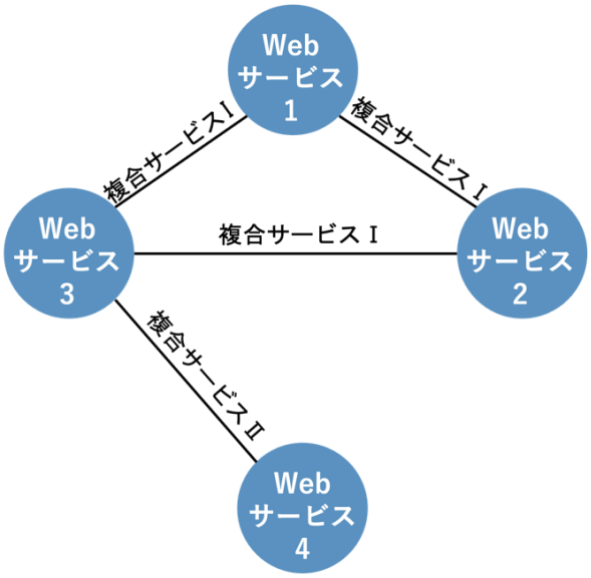

図 5：Web サービス連携関係の例

3.3 Web サービス提供利用関係ネットワーク

Web サービス提供している提供者と，提供している Web サービスを利用している利用者の情報を用いて，提供者・利用者の依存グラフを作成する．Web サービスの例として，図 4(3.1.2 節参照)の Web サービス 1，2，3，4 を提供者との Web サービス提供関係，利用者との Web サービス利用関係を表 1 に示す．このような Web サービスの提供者・利用者の情報から図 6 のグラフを作成する．具体的には，Web サービス 1 は提供者 A と提供関係があり，利用者 a，b，d とは利用関係があるので， 提供者ノード A，利用者ノード a，b，d を作成し，提供者ノード A と利用者ノード a，b，d の間にエッジを繋ぐ．また，Web サービス 2 は提供者 B と提供関係があり，利用者 a とは利用関係があるので，新しく提供者ノード B を生成し，利用者ノード a とエッジで繋ぐ．これを Web サービス 3，4 でも同様に行う．これにより，提供者 A，B の近傍は利用者 a，b で同じであるため，同じ Web サービスを提供していると推測できるネットワークを構築できる．

表 1 ：Web サービスの提供・利用関係

	Web サービス提供者	Web サービス利用者
Web サービス 1	A	a，b，d
Web サービス 2	B	a
Web サービス 3	B	b，c
Web サービス 4	C	c，d，e

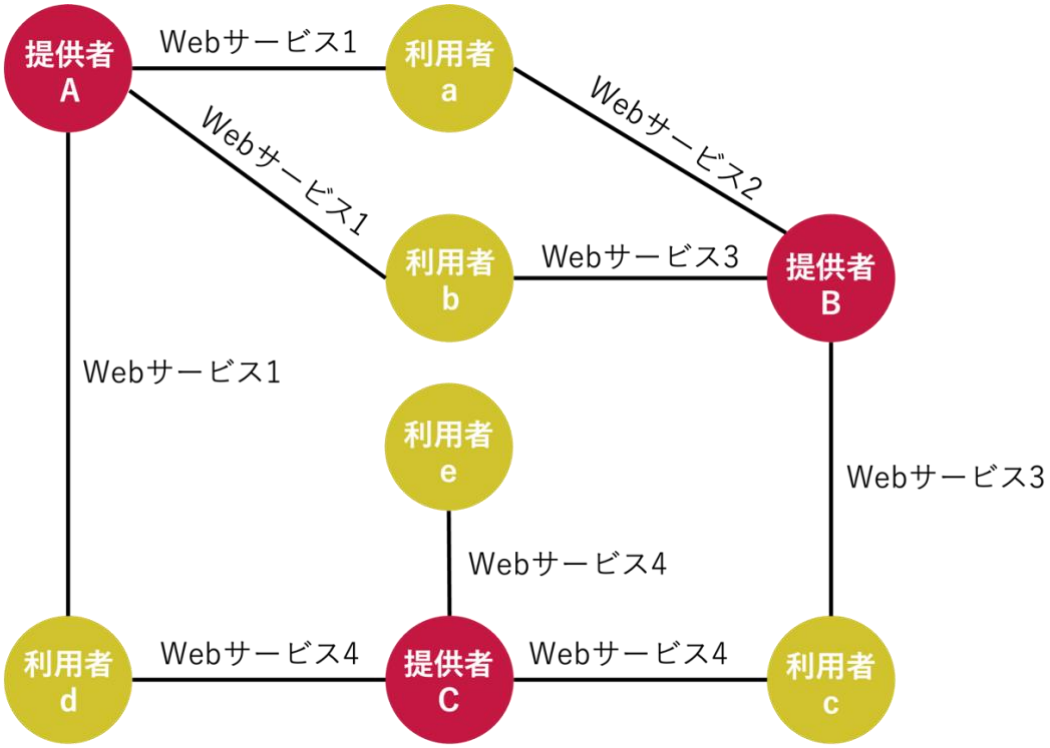

図 6：Web サービス提供利用関係の例

3.4 不完全データを用いたネットワーク構築

3.2，3.3 で述べた Web サービス連携関係ネットワーク，Web サービス提供利用ネットワークを構築する際に，それぞれのネットワークには必要な情報が不足しており，ネットワーク構築に悪影響を与える問題がある．そのため，不足している情報をそれぞれ補充し，ネットワークを構築した．

3.4.1 連携関係ネットワークでの不完全データの取り扱い

Web サービス連携関係の構築の際に，7.1 で後述するデータの中には，いくつかの Web サービスで正解データが欠損していることがある．しかし，正解データがない Web サービスを取り除いて構築を行うと，複合サービス間の関係性が損失することがある．例えば，図 7 に示される通り，複合サービス I と II が Web サービス 4 で繋がっているとする．この時，Web サービス 4 の正解データがないため，Web サービス 4 を取り除いて構築した場合，複合サービス I と II の隣接関係が失われてしまう．そのため，Web サービス連携関係ネットワークの構築においては，正解データがない Web サービスも含めて構築する．

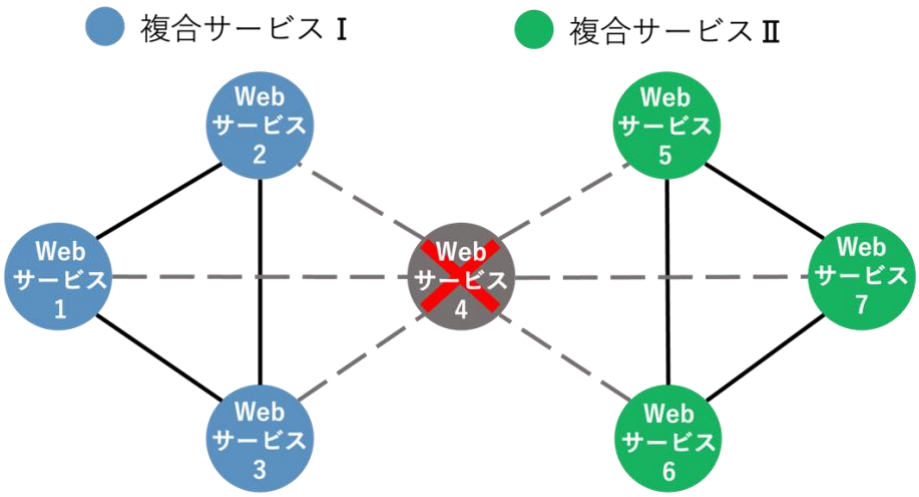

図 7：正解データのない Web サービス例

3.4.2 提供利用関係ネットワークでの不完全データの取り扱い

Web サービス提供利用関係ネットワークでは，7.1 で後述するデータの中には，いくつかの Web サービスで提供者データが欠損していることがある．このネットワークは二分グラフであり，提供者・利用者のノード間にのみエッジが存在する．そのため，提供者データが欠損している場合，その Web サービスはネットワーク上に表現できない．例えば，3.3 の図 6 から提供者 A，提供者 B が欠損していた場合を図 8 に示す．この時，Web サービス 1 の利用者 a，b，d と Web サービス 2 の利用者 a，Web サービス 3 の利用者 b 間のエッジを形成できない．このように，提供者のノードが欠損すると，利用者のノードは互いに接続されなくなる．このことから，欠損している提供者には個別に ID を割り当て，ネットワークを構築する．例として，Web サービス 1 の提供者を ID:1，Web サービス 2 の提供者を ID:2, web サービス 3 の提供者を ID:3 とする(図 9)．これにより，Web サービス 1，2，3 の提供者・利用者のノード間にエッジを追加することができ，結果として全体のネットワークの関係性をある程度維持して表現できるようになる．

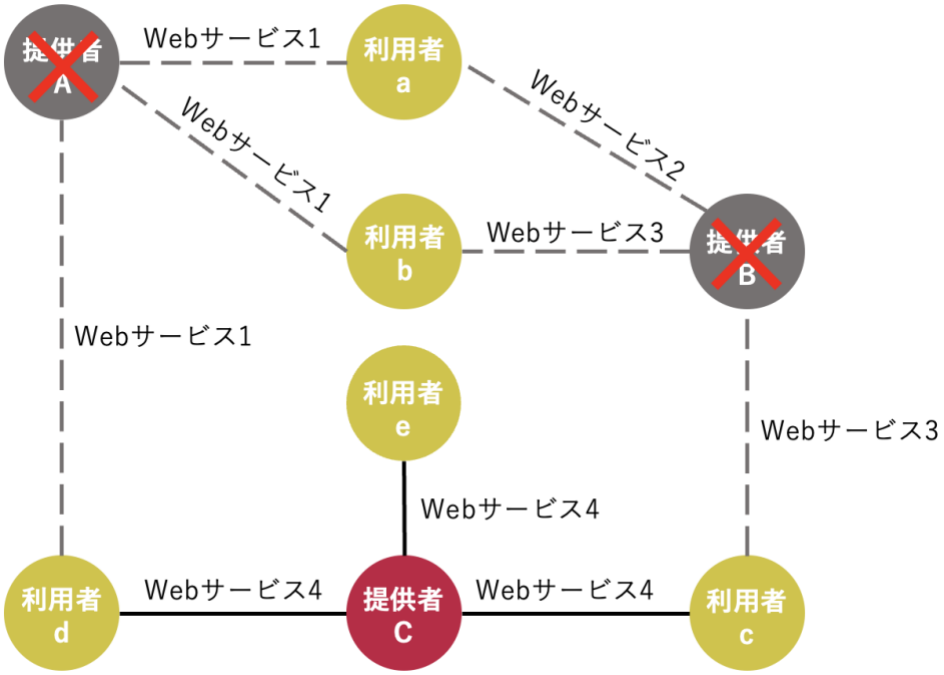

図 8：提供者情報が欠損した Web サービス提供利用関係ネットワーク

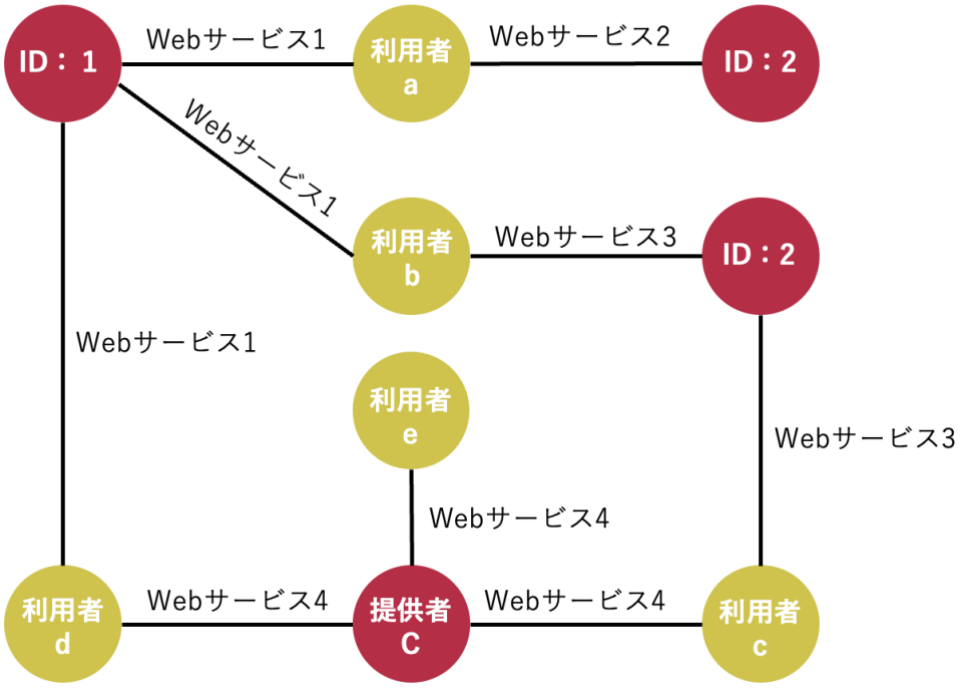

図 9：欠損した提供者情報に ID を振った提供利用関係ネットワーク

第4章 グラフ埋め込み

グラフ埋め込みとはグラフネットワークのノードやエッジ，サブグラフなどを一つのベクトル空間で表現する手法のことであり，特に skip-gram model をネットワーク構造に対して拡張した手法がいくつか提案されている．本章では，グラフネットワークのサンプリング手法として同一複合サービス優先のサンプリングについて説明する．また，最後に異種グラフ結合方法として，Concat について説明する．

4.1 同一複合サービス優先のサンプリング

Web サービス連携関係ネットワークでは，訪問中のノードが含まれている同一複合サービス内の別のノードを優先して探索する同一複合サービス優先のサンプリング手法を提案する．Web サービス連携関係ネットワークは，複合サービスにより連携されたことのある Web サービスを連携関係によりネットワークを構築しているので，それぞれの Web サービスがどの複合サービスによって連携されているかがわからない．また，同一複合サービス内のノード同士は完全ネットワークとなるので，各ノードは密接になり，ランダムウォーク[4]によるサンプリング手法では，各ノードの特徴を取得するのは非常に難しい．そこで，サンプリング結果に複合サービス情報を反映させるため，同一複合サービス優先サンプリングを行う．このサンプリング手法のアルゴリズムを図 10 に示す．はじめに，任意のデータで構築する必要がある．構築したネットワークからサンプリングを開始するノードを選択し，選択したノードが含まれる同一複合サービスをランダムに選択する．

その後，選択した同一複合サービス内のノードで指定の長さまで，未訪問のノードをランダムで探索を行う．指定の長さまで達せずに同一複合サービス内のノードをサンプリングし終えた場合，現在訪問中のノードが他の複合サービスにも含まれているなら，ランダムに未訪問の複合サービスを選択し，その複合サービス内の未訪問のノードをランダムで探索する．また，同一複合サービス内のノードをサンプリングし終えたとき，現在訪問中のノードに他の複合サービスが存在しない場合，現在訪問中の同一複合サービス内から他の複合サービスにも含まれているノードが発見されるまでランダムに探索を繰り返す．これを指定の長さまで繰り返し，全てのノードのサンプリングが完了すると，得られた各

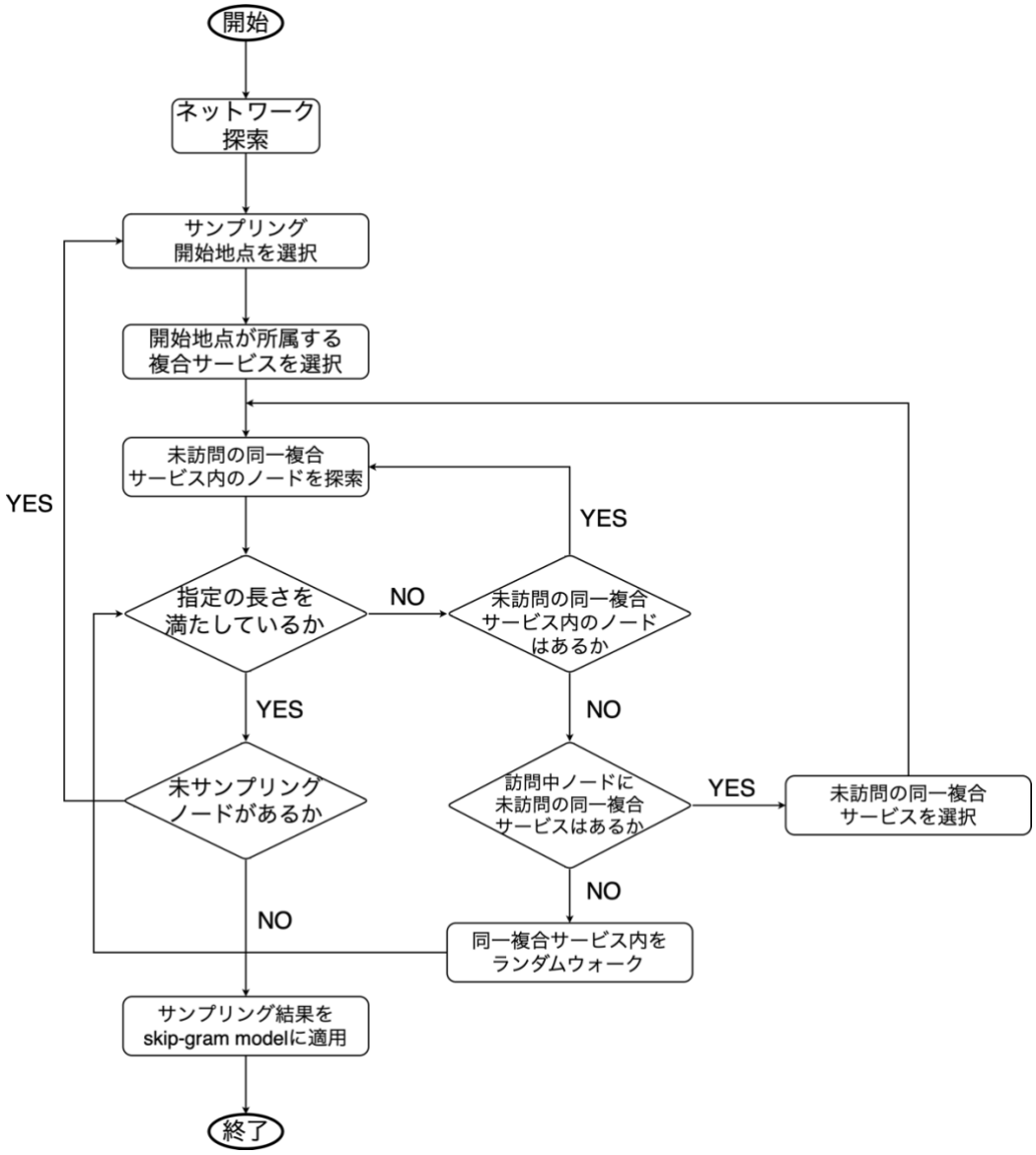

図 10：同一複合サービス優先のサンプリングのフローチャート

ノードのシーケンスは空白で区切ったテキストデータとして扱い，skip-gram model を用いて各ノードの分散表現を獲得する．

次に，図 11，図 12 を用いて，同一複合サービス優先探索のサンプリング手法の特徴を，ランダムウォークの手法と比較し，説明する．これは，複合サービスⅠ[サービス 1，サービス，サービス 3，サービス 4]と複合サービスⅡ[サービス 4，サービス 5，サービス 6，サービス 7]で構築されたネットワークである．Web サービス 1 のノードから 5 つ目のノードまでサンプリングを行うとする．図 11 はランダムウォークの行った場合のサンプリング結果例であり，サンプリング

結果は以下のようになる．

サービス 1 → サービス 2 → サービス 4 → サービス 5 → サービス 6

ランダムウォークを用いた場合，同一複合サービス I 内のノードをサンプリングし終える前に，同一複合サービスⅡをサンプリングすることがある．また，図 12 は同一複合サービス優先探索を行った場合のサンプリング結果例であり，サンプリング結果は以下のようになる．

サービス 1 → サービス 2 → サービス 3 → サービス 4 → サービス 5

同一複合サービス優先探索を用いた場合，同一複合サービス I 内のノードをサンプリング後，同一複合サービスⅡをサンプリングするようになる．今回，得られた各ノードのサンプリング結果は空白で区切ったテキストデータとして扱い，skip-gram model を用いて各ノードの分散表現を獲得する．つまり，ネットワーク探索の場合，周辺ノードが似ているノード同士は，ベクトルが近くなると考えられる．よって，同一複合サービス優先探索のサンプリング手法では，周辺ノードは同一複合サービスに連携されたノード群になるため，似た構成を持つ複合サービスに連携されたノード同士のベクトルは近くなることが期待できる．

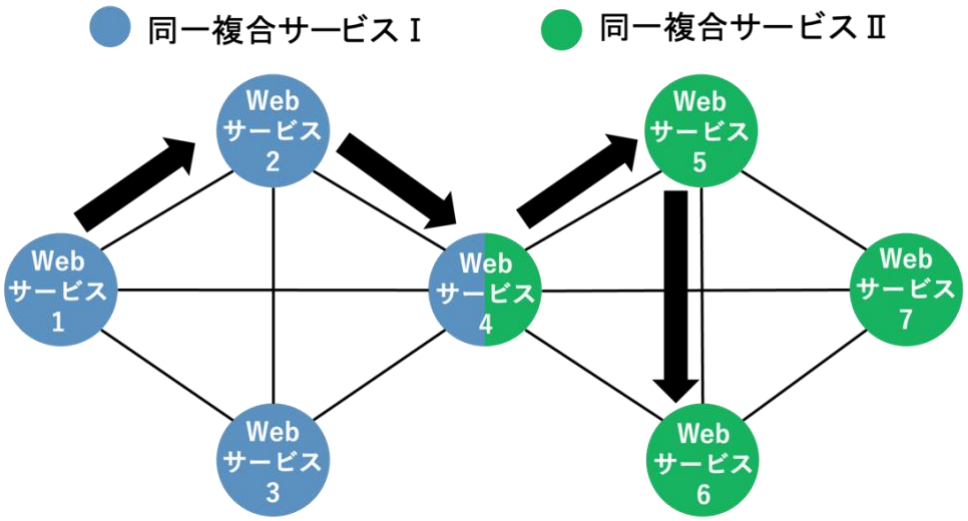

図 11：ランダムウォークのサンプリング結果

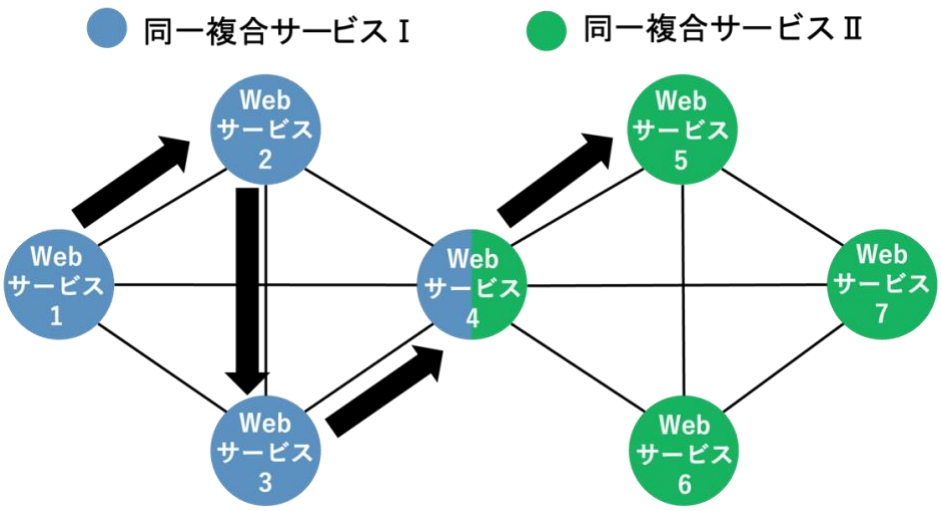

図 12：同一複合サービス優先探索のサンプリング結果

4.2 異種グラフの結合

本研究では Web サービス連携関係ネットワークと，Web サービス提供利用関係ネットワークの二種類を利用し，それぞれでサンプリングを行いグラフ埋め込みによって分散表現を獲得している．その分散表現を結合するため，本研究では Concat の手法を用いた．具体的には，Web サービス連携関係ネットワークで得た Web サービスのベクトルと Web サービス提供利用関係ネットワークで得た Web サービスの提供者のベクトルを結合する．例を表 2 に示す．

表 2 は，Web サービスである”Yahoo maps”のベクトルと，”Yahoo maps”の提供者である”Yahoo”のベクトルを表示している．これを Concat によって”Yahoo maps”のベクトル[1，2，3]と”Yahoo”のベクトル[4，5，6]を結合すると以下のようになる．

$$[1，\ 2，\ 3] + [4，\ 5，\ 6] = [1，\ 2，\ 3，\ 4，\ 5，\ 6]$$

Concat に似た手法として，2 つのベクトルの要素ごとに加算する Add2，2 つのベクトルを要素ごとに乗算させる Mul2 がある．従来手法において，Add2，Mul2 より Concat の方が予測精度は高いので，本研究においても用いる．

表 2：ハイパーパラメータの一覧

	タグ	ベクトル
Web サービス	Yahoo maps	[1，2，3]
Web サービスの提供者	Yahoo	[4，5，6]

第5章 クラスタリング

本研究では 4 章で説明した埋め込み手法から得られた分散表現をクラスタリングする必要がある．本章では，複数あるクラスタリング手法のうち，使用したハードクラスタリングの k-means 法とソフトクラスタリング の Fuzzy-C-means 法，Gaussian Mixture model 法について説明する．

5.1 k-means

まず, 従来手法で使用されていた k-means について説明する. k-means 法は，各 Web サービスは一つのクラスタに属するというハードクラスタリングである[7]．具体的なアルゴリズムを数式(1)に示す.

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{ik}\|x_i - \mu_k\|^2 \tag{1}$$

数式(1)より，最初に K 個のクラスタ中心μ_kをランダムに選択する．次に，各データ $\mathrm{x_i}$を，最も近いクラスタ中心μ_kに基づいてクラスタを割り当てる．この割り当ては，各データとクラスタ中心間の距離の二乗$\|\mathrm{x_i} - \mu_\mathrm{k}\|^2$を最小化に基づいている．その後，クラスタ中心$\mathrm{q_{ik}}$を，そのクラスタに割り当てられたデータの平均に更新する．このプロセスを繰り返すことで各データに唯一のクラスタが割り当てられる．各データの割り当ては，q_{ik}により表現され，割り当てられたクラスタがkであれば$q_{ik} = 1$，そうでなければ$q_{ik} = 0$になる．

5.2 Fuzzy-c-means

Fuzzy-c-means(以下，FCM)は k-means 法のアルゴリズムをもとに，複数のクラスタへの所属を許可するように改変したソフトクラスタリングである[8]．具体的なアルゴリズムを数式(2)に示す.

$$J = \sum_{n=1}^{N}\sum_{k=1}^{C} (u_{ik})^m\|x_i - \mu_k\|^2 \tag{2}$$

数式(2)より，各データとクラスタ中心間の距離の二乗$\|\mathrm{x_i} - \mu_\mathrm{k}\|^2$に基づいて，クラスタを割り当てる．各データの割り当ては，u_{ik} で計算され，0~1 の範囲で各クラスタ中心にどれだけ近いかに基づいている．その後，各クラスタ中心を，

所属度に基づいて重み付けされたデータポイントの平均に更新する．のプロセスを繰り返すことで各データに複数のクラスタが割り当てられる．

このアルゴリズムの特徴としてハイパーパラメータ m>1 の数値を増加させるほど，各クラスタへの所属が曖昧になることが挙げられる．つまり，m>1 の数値を増加させるほど，クラスタの境界線が曖昧になるため，各データポインタはより多くのクラスタに所属することになる．

5.3 Gaussian Mixture model

Gaussian Mixture Model(GMM)は混合ガウス分布に基づいたソフトクラスタリング手法である．このモデルは各データがいくつかの異なるガウス分布の一つから生成されたと考え，それぞれの分布をクラスタとしている[9]．具体的なアルゴリズムを数式(3)に示す．

$$p(x) = \sum_{i=1}^{n} \pi_k N(x|\mu_i, \sum_k) \tag{3}$$

数式(4)より，確率密度関数$p(x)$を使用してデータがx複数のガウス分布の一つから生成される確率を計算する．各ガウス分布は$N(x|\mu_i, \sum_k)$で表され，平均μ_iと共分散$\sum_k$を持つ．混合係数π_kは，各分布のデータ全体における存在割合を示す．これにより，各データがそれぞれのガウス分布に属する確率を推定し，データポインタが複数のクラスタに属する可能性を表現する．

第6章 評価

第 3 章で説明した，Web サービス連携関係と Web サービス提供利用関係の二種類のネットワークを構築し，第 4 章で説明したグラフ埋め込み手法を用いて，Web サービスを類似機能のグループにクラスタリングする．本章では，はじめに実験データを説明する．次に，それぞれの手法を比較する際に使用する評価指標について説明する．最後に，ハードクラスタリングの結果とソフトクラスタリングの結果とその比較方法について説明する．

6.1 評価手法

6.1.1 実験データ

実験データは programmableweb[2] に記載されている複合サービスデータ 2890 件，全 Web サービスデータ 15,278 件を用いる．取得した複合サービスデータには，複合サービス名とその複合サービスに連携されている Web サービス名が記述されている．また，取得した全 Web サービスデータには，Web サービス名，機能名，Web サービスの説明文，提供者，利用者が記載されている．取得した複合サービスデータから Web サービス連携関係ネットワーク，全 Web サービスデータから Web サービス提供利用関係ネットワークを構築する．

6.1.2 評価指標

各手法により生成されるクラスタリング結果を比較する際の評価指標として，Purity 値(数式 4)と F 値(数式 5)を求めた．また，F 値を求める際に使用した Recall(数式 6)と Precision(数式 7)の定義も示す．

$$Purity = \frac{1}{N}\sum_i \max_j \left|sc_i \cap cc_j\right| \tag{4}$$

$$F = \frac{2 * Recall * Precision}{Recall + Precision} \tag{5}$$

$$Recall = \frac{1}{C}\max_j \sum_i \left(\frac{sc_i \cap cc_j}{cc_j}\right) \tag{6}$$

$$Precision = \frac{1}{C}\max_j \sum_i \left(\frac{\left|sc_i \cap cc_j\right|}{sc_i}\right) \tag{7}$$

---

[2]https://www.programmableweb.com

数式(2)の Purity 値では, 各手法において生成したクラスタ sc と正解クラスタ cc の要素を比較し，最も共通項の多い組み合わせを作り，共通項の総和を Web サービス総数 N で割っている．これにより，対象となる全 Web サービスの何割が正しいクラスタに所属しているか数値的に判断することができる．また, 数式(4)の F 値は，Recall と Precision の調和平均である．Recall は正しいクラスタに属する Web サービスのうち，そのクラスタに実際に含まれている Web サービスの割合を示す．Precision は特定のクラスタに分類された Web サービスのうち，正しく分類された Web サービスの割合を示す．よって，F 値はクラスタリングがどれだけ効果的にデータを正確かつ完全に分類しているかを数値で表す．ここでの, 正解クラスタは Programmableweb に記載されている人手で付与された各サービスのカテゴリ情報を基に作成している．

6.1.3 生成クラスタと正解クラスタの対応付けのパターン

6.1.2 の purity 値を計算する際の生成クラスタと正解クラスタの対応付けにおいて，以下の 4 つのパターンが想定される．

・pattern1：正解クラスタをメイン機能のみで作成し，生成クラスタと正解クラスタが 1 対 1 の対応付けを行う．

・pattern2：正解クラスタを全部の機能で作成し，生成クラスタと正解クラスタが 1 対 1 の対応付けを行う．

・pattern3：正解クラスタをメイン機能のみで作成し，生成クラスタと正解クラスタが 1 対多の対応付けを行う．

・pattern4：正解クラスタを全部の機能で作成し，生成クラスタと正解クラスタが 1 対多の対応付けを行う．

これら 4 つのパターンに分類される理由は以下の二つである．

・各 Web サービスは複数の機能を持つ場合がある

・複数の生成クラスタが共通の正解クラスタと対応付される場合があるため

まず，各 Web サービスは複数の機能を持つ場合について説明する．6.1.1 で述べたように正解クラスタは ProgrammableWeb から取得した人手で付与された機能情報を基に作成する．ProgrammableWeb に機能情報を登録する際に，メイン機能とサブ機能を登録することができる．以上の理由により，メイン機能のみで正解クラスタを作成するパターンと，全部の機能で正解クラスタを作成するパターンが存在することがわかる．

次に，複数の生成クラスタが共通の正解クラスタと対応付される場合があ

る場合について説明する．Purity 値と F 値の計算を行う過程で，最も共通する正解クラスタと生成クラスタの組み合わせを見つける工程がある．各類似機能でクラスタ化されている時，正解クラスタは唯一に決定する．しかし，同一類似機能のクラスタが複数生成された場合，共通の正解クラスタと組み合わせる場合が生じる．よって，同一類似機能のクラスタが複数生成され，組み合わされる正解クラスタを重複しないようにした場合，その類似機能で組み合わされる要素数が最も多いクラスタが選ばれ，2 つ目以降の同一類似機能クラスタは要素数が次に多い機能で組み合わされることが考えられる．以上の理由により，正解クラスタと生成クラスタが 1 対多の対応付けを行う場合，正解クラスタと生成クラスタが 1 対 1 の対応付けを行うパターンが存在することがわかる．この 4 つのパターンを用いてクラスタリング精度を計測した．

6.1.4 ソフトクラスタリング時の閾値の設定

5 章で説明したソフトクラスタリングを，Web サービスの分散表現に適用すると，Web サービスはほぼ全てのクラスタに所属するため，評価が困難になる．そこで，各 Web サービスのクラスタ所属数は，Programmableweb に記載のある機能数と同様になるように閾値を設定し，ソフトクラスタリングを行った．例えば，”google maps”という Web サービスをソフトクラスタリングする時，Programmableweb に記載されている機能数が 2 であるなら，所属するクラスタ数は 2 になる．また，ハードクラスタリングの結果をソフトクラスタリングと比較する場合，各 Web サービスは Programmableweb に記載されている機能数分 Web サービスを複製し，ハードクラスタリングを行った．

6.1.5 評価するネットワーク

今回評価するネットワークを表 3 に示す．はじめに，Web サービス連携関係ネットワークで生成した Web サービスベクトルと Web サービス提供利用関係ネットワークで生成した提供者ベクトルを 4.2 で説明した Concat で結合し，サービスの機能ごとにクラスタリングを行った．この際，Web サービス連携関係ネットワークは，ランダムウォークと，4.1 で説明した同一複合サービス優先探索の両方を調査した．最後に，グラフ埋め込みの手法の有効性を示す基準として，付録にて説明をしているテキスト埋め込みを用いて，サービス説明文を入力とした BERT での Web サービスのクラスタリング行った．

表 3：評価するネットワーク

	連携関係		提供利用関係	テキスト埋め込み
	ランダムウォーク	同一複合サービス優先	ランダムウォーク	
①	○		○	
②		○	○	
③				○

6.2 ハードクラスタリングの結果

6.1 で説明した評価指標のうち，メイン機能で正解クラスタを作成している pattern1，pattern3 で評価した．各手法を用いて Web サービスを類似機能にクラスタリングした結果をそれぞれ表 4，表 5 に示す．振られている番号のネットワークは 6.1.5 の表 3 と同じである．表 4~5 より，最も Purity 値が高くなったのは値が 0.6325 の pattern3 の②である．次に最も F 値が高くなったのは値が 0.2513 の pattern3 の①である．このことから，ハードクラスタリングでは，グラフ埋め込みの方がテキスト埋め込みよりクラスタリング精度が高いことがわかる．

表 4：pattern1 における各ネットワークのクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.2483	0.2029	0.2220	0.2523
②	0.2539	0.1915	0.2250	0.2167
③	0.2406	0.1814	0.2057	0.2207

表 5：pattern3 における各ネットワークのクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.3448	0.2513	0.2261	0.5345
②	0.3681	0.2393	0.2046	0.5270
③	0.3647	0.2159	0.1750	0.4448

6.3 ソフトクラスタリングの結果

6.1 で説明した評価指標のうち，全部の機能で正解クラスタを作成している pattern2，pattern4 で評価した．その際，FCM は 5.2 で説明したように m のパラメータによってクラスタ境界線が曖昧になるため，パラメータ m を変化させた結果を図 13 に示す．図 13 の結果から，どちらの patten においても m= 1.05

の時に精度が一番高くなることがわかる.

続いて，各ネットワークを用いた場合のクラスタリング精度を，FCM を使用した場合と， GMM を使用した場合，ハードクラスタリングの場合を表 6~11 に示す．振られている番号は 6.1.5 の表 4 と同じである．また，各 Web サービスの閾値は 6.1.5 に説明したもので評価している.

表 6~表 11 より，最も purity 値が高くなったのは値が 0.4371 の FCM の pattaern4，③である．次に最も F 値が高くなったのは値が 0.2231 の GMM の pattern2，①である．また，最も purity 値が高くなった FCM の pattaern4，③では，複数の機能を持つ Web サービス 784 個のうち 250 個を少なくとも 2 つの機能において正確に分類できた.

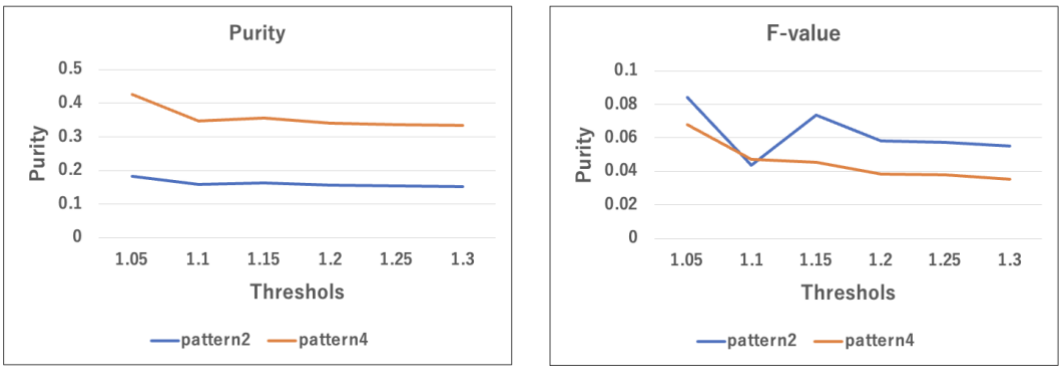

図 13：FCM(m=1.05，1.1，1.15，1.2，1.25，1.3)の精度

表 6：pattern2 における FCM の各ネットワークのクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.1830	0.0840	0.0969	0.1253
②	0.2171	0.1427	0.1480	0.2154
③	0.2601	0.2470	0.3111	0.2733

表 7：pattern4 における FCM の各ネットワークのクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.2433	0.0679	0.0579	0.1770
②	0.3200	0.1156	0.0913	0.3984
③	0.4371	0.1986	0.1488	0.5312

表 8：pattern2 における GMM の各ネットワークのクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.2148	0.2231	0.1864	0.4232
②	0.2107	0.2091	0.1792	0.3634
③	0.2081	0.2125	0.1852	0.3718

表 9：pattern4 における GMM の各ネットワークのクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.3118	0.1508	0.1019	0.7979
②	0.3054	0.1440	0.0957	0.7551
③	0.3061	0.1385	0.0937	0.7227

表 10：pattern2 における各ネットワークのハードクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.1613	0.1501	0.1937	0.1524
②	0.1602	0.1529	0.2090	0.1574
③	0.1523	0.1542	0.1568	0.2063

表 11：pattern4 における各ネットワークのハードクラスタリング精度

	Purity 値	F 値	Recall	Precision
①	0.2268	0.1211	0.1085	0.2611
②	0.2260	0.1143	0.0991	0.2574
③	0.2272	0.1264	0.1178	0.2547

第7章 考察

本章では，6.2，6.3 で提示した結果を受けて各ネットワークのクラスタリング精度について考察のソフトとハードの結果から考察する．また，グラフベースとテキストベースの手法のクラスタリング精度について考察を行う．

7.1 ハードとソフトクラスタリング結果の比較

従来手法であるハードクラスタリングと本手法であるソフトクラスタリングを利用した場合のクラスタリング結果を比較し，考察する．表 12 では，ハードクラスタリングを用いた場合と，ソフトクラスタリングを用いた場合の生成クラスタに割り当てられた機能の数と正しく機能別に分類された Web サービスの総数を比較している．Web サービスの総数は，各 Web サービスが持つ機能の内，一つでも正しく機能に分類された Web サービスの数としている．マッチした機能の数はハードクラスタリングが多いのに対し，正しく分類された数はソフトクラスタリングの方が多いことがわかる．このことから，ハードクラスタリングはデータの幅広い特徴を捉えるのに適している可能性があるが，ソフトクラスタリングの方が特定のデータポイントの細かい特徴や関係性をより正確に表現できると考える．

表 12：マッチした機能の数と機能別分類 Web サービスの総数

	マッチした機能の数	正しく機能別に分類された Web サービスの総数
ハードクラスタリング	75	597
ソフトクラスタリング	67	833

7.2 BERT と異種ネットワーク結果の比較

本研究で使用したグラフ埋め込みを用いたクラスタリング手法と，テキスト埋め込みを用いたクラスタリング手法のクラスタリングを比較する．具体的には，6.2 の pattern3 の評価において，テキスト埋め込みでクラスタできない Web サービスの内，提案手法で正しくクラスタリングできた Web サービスのベン図を図 14 に示す．図 14 より，本手法で正しく分類できた Web サービスの内，約 45%がテキスト埋め込み手法で分類できない Web サービスということがわかった．次に，テキスト埋め込み手法のクラスタリング結果の一例を表 13 に示す．

表 13 より，このクラスタでは Web サービス名に Google が付くものが多く所属していることがわかる．しかし，Web サービスの持つ機能に関連性がないため，サービス説明文内の Google をキーワードとして，同じクラスタに分類された可能性が高いと考えられる．また，全 Web サービス 902 個の内，776 個の Web サービスには，“API”という文字がサービス説明文に記載されていた．このことから，サービス説明文に”API”と記載されている Web サービス同士は類似したベクトルを生成し，更に特定の固有名詞もサービス説明文に多く記載されていることで同一クラスタ内に分類されたのだと考えられ，グラフ埋め込みではこのようなサービスを正しく分類できるようになったのではないかと考えられる．

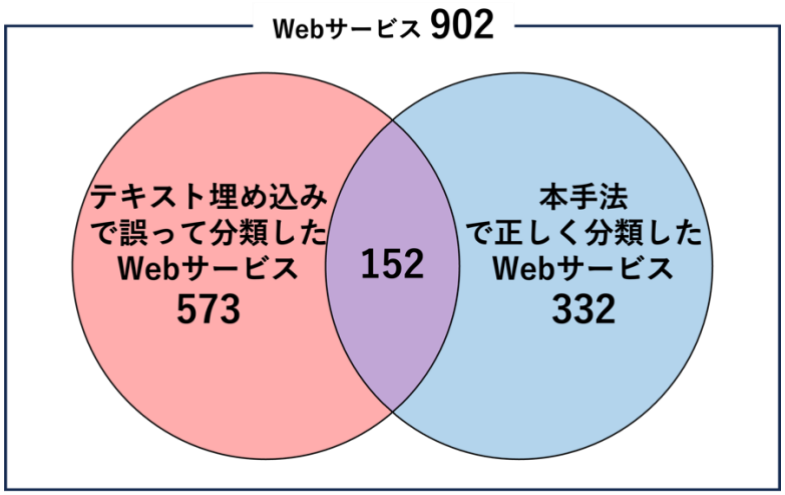

図 14：分類結果のベン図

表 13：テキスト埋め込み手法のクラスタの一例

Web サービス名	メイン機能
Google plus	social
Google waze	travel
Google openid	security
Google cloud print	office
Google apps email migration	email
Google web authentication	security
Google mirror	cloud
cronofy	calenders
bart	transportation

第8章 おわりに

本研究では，従来の WSDL やサービス説明文に基づくテキストベースの代わりに提案された Web サービス依存関係・提供利用関係に基づいて，不完全データを補完するクラスタリング手法を提案した．そして，ソフトクラスタリングを用いた手法では，従来のハードクラスタリングを行う手法よりも有効であることを示した．

本研究の貢献は以下の通りである．

不完全データの補完

Web サービス間の結びつきをある程度維持するために，Web サービス連携関係ネットワークでは機能情報の無いネットワークも含めて構築した．また，Web サービス提供利用関係ネットワークでは，提供者情報が欠損している Web サービスには個別に ID を振ってネットワークを構築した．提供者データが欠けているが機能データが完全な Web サービスも評価が可能になったことで，Web サービスの機能ごとのハードクラスタリングの精度が向上した．

ソフトクラスタリングの適用方法

ネットワークのサンプリング結果をクラスタリング後，人手で付与した Web サービスの機能数に基づいて閾値を設定することで，既存手法と比較し，Web サービスの多様性を考慮したクラスタリング結果を得られた．

今後，実世界において複合サービスを構築する際，必要な機能を持つ Web サービスをユーザが検索するのではなく，システムが自動で Web サービスの候補を挙げることが望ましいと考えられる．これを実現するためには，各クラスタが類似機能なクラスタであるだけでなく，クラスタ内の Web サービスでユーザの要求に最適な Web サービスを候補に上げる必要がある．

本研究では Web サービスの連携関係・提供利用関係に着目したクラスタリング手法であるため，複合サービスの連携回数・提供利用回数などを考慮したネットワークを構築することで，より精度の高いクラスタリング結果を得られると考えられる．また，本手法のグラフベースのクラスタリングとテキストベースのクラスタリングは正しく分類できるサービスが異なる可能性があることから，グラフベースとテキストベースを混合させたクラスタリング手法を行うことで，より高いクラスタリング精度を得られると考えられる．

謝辞

本研究を行うにあたり，熱心なご指導，ご助言を賜りました指導教官の村上陽平教授並びに大久保弘基先輩，大井也史先輩に深謝申し上げます．また普段からお世話になっている社会知能研究室の皆様にも感謝の意を表します．

参考文献

1. Li, C., Zhang, R., Huai, J., & Sun, H: A Novel Approach for API Recommendation in Mashup Development, In Proceedings of the 2014 IEEE International Conference on Web Services, pp.289-296 (2014).
2. KhalidElgazzar, AhmedE.HassanandPatrickMartin: ClusteringWSDL Documents to Bootstrap the Discovery of Web Service, Proceedings of IEEE International Conference on Web Services, pp. 147-154 (2010).
3. Min Shi, Jianxun Liu, Dong Zhou, Mingdong Tang, and Buqing Cao: WE-LDA: a word embeddings augmented LDA model for web services clustering, in *IEEE International Conference on Web Services (ICWS)*, pp. 9–16 (2017).
4. Aditya Grover, Jure Laskovec: node2vec: Scalable Fearure Learning for Networks, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855-864(2016).
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
6. Kemas Muslim Lhaksmana, Yohei Murakami, Toru Ishida: Analysis of Large-Scale Service Network Tolerance to Cascading Failure, IEEE Internet of Things Journal, Vol. 3, No. 6, pp. 1159-1170(2016).
7. Shi Na, Liu Xumin, Guan Yong: Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, In 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 63 – 67(2010).
8. James C. Bezdek, Robert Ehrlich, William Full: FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences vol. 10, No. 2-3, pp. 191-203(1984).
9. Yi Zhang, Miaomiao Li, Siwei Wang, Sisi Dai, Lei Luo, En Zhu, Huiying Xu, Xinzhong Zhu, Chaoyun Yao, Haoran Zhou: Gaussian Mixture Model Clustering with Incomplete Data, ACM Transactions on Multimedia

Computing, Communications, and Applications, Vol. 17, No. 15, pp. 1-14(2021).

付録：テキスト埋め込み

第４章で説明したグラフ埋め込みはグラフネットワークに適用するのに対し，テキスト埋め込みは，自然言語のテキストをベクトル空間で表現する手法のことである．本章では，テキスト埋め込み手法の一つである BERT を説明した後，BERT の Web サービスの説明文への適用について説明する．

A.1 BERT

BERT とは，2018 年 10 月に Google の Jacob Devlin らによって発表された自然言語処理モデルである[5]．　BERT は文全体を考慮し，各単語が文脈内でどのような役割を果たしているかを理解できる．例えば，”Japan traveler to USA”　という文章があった時，従来の自然言語処理では”to”が文章中のどの単語にかかるか理解できず，”日本へのアメリカ旅行者”と誤って解釈されることがある．しかし，BERT では”to”が”USA”にかかることが理解できる．これにより，文章の文脈を考慮した分散表現を獲得できるようになる．

A.2 サービス説明文への適用

本研究では，BERT の事前学習モデルを用いて，各 Web サービスの説明文の分散表現を獲得した．事前学習モデルは，bert-base-uncased[1] を用いる．このモデルは 110 万の単語から構成された，主に Wikipedia の英語版のテキストをデータセットとして学習されている．上記の事前学習モデルを使うことで，各 Web サービスの説明文を入力とし，768 次元の分散表現を獲得することができる．

[1] https://huggingface.co/bert-base-uncased