QA データから構築した共起グラフを用いた関連症状の発見

本白水健輔† 湯本 高行†

† 兵庫県立大学 情報科学研究科 〒 651–2197 兵庫県神戸市西区学園西町 8 丁目 2-1
E-mail: †{ad21d049,yumoto}@gsis.u-hyogo.ac.jp

あらまし 本論文では、特定の症状から関連のある症状の発見を行う。QA サイトには実際に患者が医師に聞くようなテキストが含まれていることから、QA サイトの質問データを用いる。質問データから症状に関する用語を抽出し、相関ルールを用いて共起を求める。求めた症状の共起を共起ネットワークで表す。得られた共起ネットワークに入力となる自覚症状の周辺のノードを抽出し、各ノードのスコアリングを行う。このスコアリングでは、HITS アルゴリズムの概念を応用した固有ベクトルによるスコアリングを行うことで、症状の関連パターンを複数抽出する。

キーワード 医療・ヘルスケア，相関ルール,HITS アルゴリズム

1 はじめに

病気には様々な症状があり、病気によって起こり得る症状がある。病気の中には死に至るような病もあり、病気に対応した症状を把握しておくことは重要である。また、些細な症状が重病の症状である可能性もあるため、このような症状をできるだけ早くに気づき、正しい治療が速やかに行うことができれば、重症化を防ぐことや軽症で済む場合もある。しかし診察において、このような重要な症状を見逃してしまうことが考えられる。例えば、診察で患者自身の判断で申告しない場合である。例えば、重要な症状がその患者にとって主でない症状であった場合である。具体例として “腹痛”という自覚症状を挙げる。この時、一般的には胃腸炎などのお腹の風邪と予想することが多いと考えられる。ここに “肩の痛み”という症状があるとしても、直接 “腹痛”と関連付けることなく、肩こりと捉えて、患者自身の判断で申告しない可能性がある。しかし、“腹痛”や “肩の痛み”は心筋梗塞の症状であるため、見逃してしまうと病気の発見が遅れ、重症化してしまう恐れがある。また、その症状が患者にとって表現しづらいような症状であった場合も申告しないことが予想できる。例えば、患者が症状や病名の名称を知らない場合や医師に指摘された症状とニュアンスが違って、症状の認識に乖離があり申告しない場合などである。

これらの課題に対して、システムが関連症状を推薦することによって、重要な症状の見逃しを防ぐことができると考えた。そこで本研究では病気における関連症状を推薦するにあたっての関連症状の発見手法を示す。

まず、関連症状を発見するためのテキストデータとして QA サイトにおける質問データを用いる。QA サイトの質問データ、特にヘルスケア関連の質問データには実際に患者が医師に聞くようなテキストが含まれている。また、QA サイトの質問データは、基本的に１つの質問が１人に対応している。これらの点に着目し、QA サイトの質問データを用いて症状の関連を分析し、関連症状の発見することができると考えた。QA サイトの質問データは、関連症状発見においての症状に関する用語の抽出を行うためのテキストデータとして用いる。抽出した症状の用語を相関ルールを用いて共起を検出をする。求めた症状の共起情報を共起ネットワークで表し、HITS アルゴリズムの概念を用いてグラフの分析を行うことで関連症状を推薦するために必要な症状の組み合わせを発見する。

本論文の構成を以下に示す。まず 2 章では本研究に用いる相関ルールおよび HITS アルゴリズムの概念を述べる。3 章では、関連症状の発見において、使用するデータおよび辞書について説明する。4 章では関連症状の発見の手法として病名表現の抽出、相関ルールを用いた共起の検出手法、固有ベクトルを用いたスコアリング手法について述べる。5 章では、実際の症状を入力し、その出力を評価するための評価データの作成、評価方法、評価結果を述べる。6 章でまとめを述べる。

2 関連研究

2.1 医療系テキストマイニング

以前からテキストマイニングを医療に用いた研究が存在する。岡部らの研究 [8] では、医療事故における事故の種類や発生場所、担当看護師などのテキストデータを用いた共起ネットワークの分析が行われている。医療現場においてヒヤリとしたことやハッとすることをレポート形式でまとめたインシデントレポートと呼ばれるテキストデータが用いられ、テキストに含まれる語句の共起情報を用いた重要語句や関連語句の抽出手法と実際のインシデントレポートに対する解析が行われた。また、Ishii らの研究 [2] では、ネットワークを医療分野に活用した研究が行われている。近年平均寿命が上がり、死因が慢性疾患へと以降していることから、多死因分析という分野に注目され、研究が行われた。死亡診断書に含まれる死因の共起を無効グラフで表現し、死因について分析されている。Fast-Greedy 法によるコミュニティ分析が行われ、死因のコミュニティが抽出された。次に本研究の先行研究となる研究について述べる。新本らの研究 [1] では、QA データを用いた関連症状について研究されている。患者自身で問診票などを入力し医療従事者の電子カルテ入力などの作業の時間を削減することを目的とする

ために研究が行われた。新本らの研究では、Q & A サイトにおける質問データから、「症状+部位」に限定して抽出を行う。抽出した症状に対して、モダリティ解析と呼ばれるイベントに対して真偽判断の解析手法を用いて、それらが成立しているかどうかの判定を行い、得られた「症状+部位」の組み合わせの共起関係から相関ルールを抽出する。また、部位の抽出には部位の概念ツリーを定義し、関連部位は表記揺れの緩和が試された。関連症状の推薦においては、「症状+部位」の組み合わせでの出力となっており、相関ルールにおける確信度及び帰結部の最大確信度を用いたスコアが定義され、そのスコアによってランキングを行い、症状が推薦される。関連症状の推薦の評価では、7 割程度の適合率で推薦された。

3　本研究の対象データおよび辞書

3.1　データセット

本研究では、関連症状を発見するための対象データとして QA サイトにおける質問データを用いる。使用する QA データは Yahoo!データセット第 3 版 [7] を使用する。Yahoo データセット第 3 版は Yahoo!株式会社のサービスの一つである Yahoo!知恵袋におけるデータセットである。本研究では、病気および症状に関連する質問データのみが必要であるため、上記のデータセット内における “病気、症状”カテゴリを利用する。本研究で実際に使用する Yahoo!知恵袋における質問例を表 1 に示す。本研究では表 1 に示したテキストのように、実際に患者が医師に聞くようなテキストを想定している。

表 1　本研究で用いる Yahoo!知恵袋の質問例

1	頭痛、吐き気、焦点が合わない、ということがあり、4 時間くらいで頭痛のみになりました。これは脳震盪が起きていたということでいいんでしょうか？
2	最近、腹痛が続きます。腹痛といっても、我慢できないほどではなく、へそ周辺がチクチクする感じです。そのほかにも、背中が弱いズキっとした痛み、肩甲骨周辺もチクチクする感じの痛みがあります。これは大腸がんなどが考えられるのでしょうか？自分は猫背で姿勢が悪いのでそういうのも影響してるのでしょうか？
3	頭が痛くて、目が重くて、ちゃんと呼吸してるか不安になります。熱っぽいのに熱はありません。日々の疲れはなかなか取れません。

3.2　症状に関する用語の辞書

関連症状を発見するにあたって、症状に関連する用語を定義する必要がある。そこで本研究では、奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室が公開している患者表現辞書 [3] を使用する。患者表現辞書とは、患者が実際に用いる病名表現を網羅した辞書である。データは約 9000 件あり、データの属性としては、出現形、部位、表現例、方言地方、ICD10 コード、エビデンスがある。しかし、すべてのデータに対して各属性すべてが存在するわけではない。この患者表現辞書における出現系は、1 つの症状に対して複数の病名表現を表す用語が存在する。本研究では、症状を表す用語として、患者表現辞書における出現形を用いる。

4　関連症状の発見手法

4.1　関連症状の発見の概要

本研究では共起グラフを用いて関連症状の発見を行う。Q & A サイトにおける質問データ上から症状に関連した語彙を抽出し、相関ルールによって共起情報を検出し、得られた共起情報をグラフ構造で表現する。このグラフ構造に対し、患者の自覚症状である入力症状に対応したノードの周辺のみを切り出したサブグラフを作成する。このサブグラフ内のノードに対して、スコアリングを行い、入力症状との関連を見出す。関連症状の発見における一連の処理の流れを図 1、と以下の箇条書きに示す。

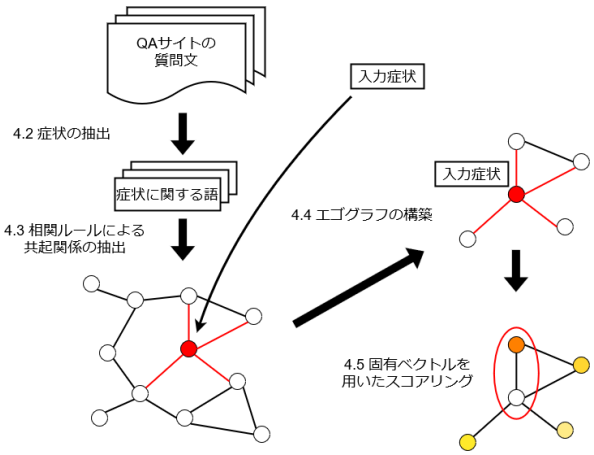

図 1　関連症状の発見における一連の処理

（a）質問データから病名表現を抽出
（b）相関ルールによる共起関係の検出
（c）共起グラフの構築
（d）固有ベクトルを用いたスコアリング

4.2　病名表現の抽出

本研究では、症状の共起情報をもとに関連症状を発見する。そのため、QA サイトにおける質問データから病名表現を抽出する必要がある。病名表現を抽出するための手法を以下に示す。

初めに、質問テキストデータに対しての処理を示す。1 つの質問データに対して形態素解析を行い、形態素に分解する。形態素に分解したテキストを集合とする。次に、患者表現辞書の出現系に対しての処理を示す。まず、患者表現辞書には、病気における症状に直接関連のない病名表現が含まれているため、患者表現辞書から直接症状と関係のないと判断した約 80 件の病名表現を削除する。本研究で使用する病名表現と使用しない病名表現の一部を表 2 に示す。本研究では、表 2 での “やし”、“なし”、“はい”のように症状以外の単語として含まれてしま

う病名表現や、“何をしているのだろう”、“普通でない感じ”、“どうでもいい”などの、症状といえない表現や症状以外にも使用されるような表現を削除する。

これらに類似した直接症状と関係のない病名表現を削除し、残った患者表現辞書の各出現系に対して形態素解析を行い、形態素に分解する。分解した形態素に対して品詞の割り当てを行い、集合とする。作成した集合から助詞、助動詞を削除する。これらの処理を行った質問データと患者表現辞書を用いて、症状の抽出を行う。この時、患者表現辞書の処理を行う病名表現の優先順位として、患者表現辞書の各集合における単語数の多い病名表現から判定を行う。質問データから症状を抽出する処理のフロー以下に示す。また、データの前処理および（1）から（3）の処理を図 1 に示す。

（1）質問データの単語集合に出現形の単語集合がすべて含まれているかを判定する。

（2）含まれている場合、出現系と質問 ID を質問と語の対応関係として記録し質問データの単語集合から出現系の単語集合を削除する。

（3）すべての出現形を判定するまで（1）、（2）を繰り返す。

（4）これらの処理をすべての質問データに対して行う。

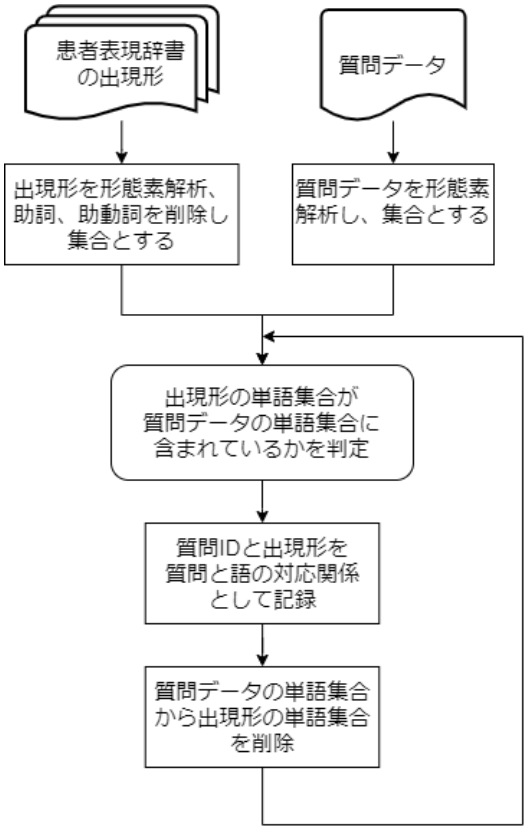

図 2　症状抽出の処理

4.3　相関ルールを用いた共起関係の検出

本研究では、症状の共起を表現する手法として相関ルールを用いる。4.2 節で得た質問と語の対応関係に対して、質問 ID でグループ化を行う。グループ化とは、指定したカラム内の値ごとにデータをまとめる処理である。グループ化を行った各 ID を相関ルールにおけるトランザクションとする。実際に質問データから症状の抽出を行い、得られたデータと質問 ID でグループ化を行ったトランザクションデータの一部を表 2 に示す。

表 2　グループ化して得られたトランザクションデータの一部

質問 ID	アイテム集合
250	{吐き気や吐くこと, 遺産の逆流, 歯の痛さ, 血が出た}
252	{胸の痛みと発作, 食欲が起きない, 咳が止まらない, めまいがする}
297	{痛痒い, 内出血}
329	{筋肉の痛み}
372	{息苦しさ, 黄色い痰, 気管支喘息}
377	{頭の後ろが痛くなる, 目が痛い}

4.4　病名表現の集約とグラフの構築

本研究で使用する患者表現辞書は、8862 種類もの出現形があり、本研究で使用する出現形だけでも 6928 種類もの出現形がある。加えて、同じ症状を表す表現が複数存在しする。このようなデータを用いて質問データから、4.2 の手法で病名表現の抽出を行い、グラフ構造で表現した場合、ほとんど同じ症状にもかかわらず、同義の語に共起していても、直接共起していないため実際のグラフ構造では、関連しないような表現となってしまう。また、患者表現辞書に含まれる標準病名と呼ばれる病気の名称はそれぞれの病気にまとめるためには、1 つの標準病名に対応した出現形の表現の幅が広い点と 1 つの出現形が複数の標準病名に該当するという点から利用できないと判断した。例えば、“痺れる”や “麻痺する” という出現形は “しびれ感” という標準病名は同じ、もしくは似ている意味になっているが、“苦しい” という出現形は “しびれ感” という標準病名とは直接同じ意味とは言えない。しかし患者表現辞書では、それぞれ、“痺れる”、“麻痺する”、“苦しい” がすべて “しびれ感” という標準病名に該当する。このように、1 つのノードにはできるだけ違った症状が含まれずに、患者表現辞書の出現形における表記ゆれの緩和及び意味的に近い出現形の集約を行う必要がある。このような症状の集約が必要となる例を図 3 に示す。図 3 より、“頭

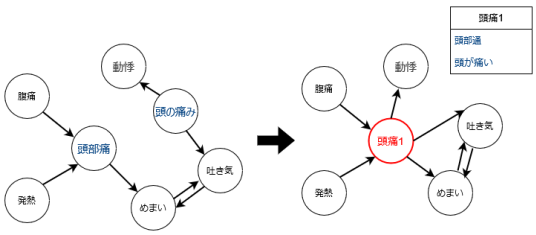

図 3　症状の集約例

部痛” というノードと “頭が痛い” というノードは意味的に近い

表現である。しかし、図 3 左では、それぞれ別のノードとして表現され、それぞれのノードが接続されている症状は全く違うものとなっている。そこで、“頭部痛”と “頭が痛い”というノードを 1 つのノードに集約し、1 つにまとめることで、それぞれのとしか接続されていなかった “動悸”や “発熱”、“めまい”などの症状をそれぞれのノードと接続されていない方のノードとも共起していることをグラフ構造上で表現することができる。

このようなノードの集約を実現させるために、語の分散表現とクラスタリングを用いて、意味的に近い出現形の集約を行うことができると考えた。患者表現辞書における出現形に対して BERT モデルを用いてベクトル化を行い、語の分散表現を行ったベクトルのデータに対してクラスタリングを行い、意味的に近い出現形のクラスターを作成する。作成したクラスター内の出現形をグラフ構造における 1 つのノードとして扱うために、相関ルールにおけるトランザクションデータ内の抽出した出現形をクラスター名で置き換える。これによって、似ている出現形を 1 つに集約し、それぞれの共起情報をまとめることができる。クラスタリング手法には階層的クラスタリングにおける凝集的手法を用いる。理由としては、本研究では患者表現辞書の出現形における表記ゆれや意味的に高い出現形のみを集約したいと考えているため、近いものから集約し、任意の閾値をもとにクラスタリングを行うことができる階層的クラスタリングにおける凝集的手法を利用する。階層的クラスタリングにおけるパラメータは、距離の閾値が 0.25、リンケージアルゴリズムは平均法を使用する。また、クラスター間の非類似度には、コサイン類似度を使用し、設定した閾値以上の類似度かを判定する。使用する BERT モデルは、東北大学の乾研究室による日本語の事前学習モデル [4] を使用する。BERT モデルを用いたベクトル化では、[PAD] トークン以外のトークンの平均をとる。各パラメータは、トークン列の最大を 256 とし、最大トークン列の 256 でパディングする。

このトランザクションから支持度、確信度、リフト値を計算する。また、本研究では、相関ルールにおける頻出アイテムを求る。アイテム集合を作成する手法として Apriori アルゴリズム [5] を用いる。Apriori アルゴリズムを用いる際の支持度の閾値を下限として設定し、アイテム集合を作成する。本研究では、Apriori アルゴリズムを用いる際の支持度の閾値を 0.0001 とする。また、本研究で作成するアイテムセットは、前提部、結論部に 1 つずつ症状を対応させた共起のみ使用し、前提部、結論部に 2 つ以上症状が対応した共起は使用しない。理由としては、このアイテムセットから得られた情報をグラフ構造で表現するからである。この処理によって得られた前提部、結論部及び、支持度、確信度、リフト値から共起ネットワークを構築する。共起ネットワークとは、テキストデータの中の語の共起関係を表現したグラフ構造である。本研究の場合は、同一の質問に含まれる語を共起する語とみなす。また、グラフ構造における 1 つのノードが前提部、結論部に当たり、前提部から結論部に向けてエッジを張る。相関ルールにおける確信度をエッジの重みとして設定し共起ネットワークを構築する。

4.5 共起グラフの構築

4.5.1 エゴグラフの構築

本研究では、1 つの症状のから関連する症状を出力することが目的である。しかし、共起ネットワーク全体からでは、1 つの症状に対して関連する症状を抽出することができない。そのため、関連する症状を抽出するため、入力値となる症状からエゴグラフを構築し、構築したエゴグラフ内でスコアリングを行うことで、症状のランキングを行う。エゴグラフ (ego graph) とは、エゴセントリックネットワーク (Ego Centric Network) とも呼ばれ、中心ノード (ego) と、中心ノードの周辺のノード (alter) で構築されるサブグラフである。エゴグラフの構築には、入力値となる中心ノードに対する半径を設定することができる。エゴグラフ構築における半径とは、中心ノードからいくつのエッジを介してそのノードに接続されるかをパラメータとして設定する値である。半径が 1 では、中心ノード (ego) と直接接続していることを意味し、中心ノードと中心ノードに直接接続しているノードのみで構築されたサブグラフとなる。。また、本研究ではエゴグラフ構築において有向グラフとして捉え、中心ノードと接続しているエッジに対し、中心ノードからのエッジのみで探索しエゴグラフを構築する。しかし、逆方向のエッジがある場合はエゴグラフに含めるとする。また、本研究でのエゴグラフ構築においては、エッジの重みは考慮せず、エッジ数のみでエゴグラフを構築する。

4.5.2 ダミーノードを追加する手法

本研究では、4.4 で作成したグラフ構造を全体グラフ、入力症状となる自覚症状を中心ノードとして、入力症状のエゴグラフを構築する。

エゴグラフを構成するノードに対応する症状は一般的には中心ノードに対応する症状との関係が強い。しかし、全体グラフで次数中心性が高いノードに対応した症状はどの症状とも関連しているが、エゴグラフにおいて、とりわけ中心ノードに対応する症状とだけ関係が強いわけではない。次数中心性 (degree centrality) とは、各ノードの次数 (そのノードに接続しているエッジの本数) を用いた、そのノードがどの程度他のノードと接続しているかを示す指標である。

このような問題に対して、全体グラフで次数中心性が高いノードの緩和を考えたエゴグラフの改善案として、ダミーのノードを追加する手法、入力症状にエッジを張る手法を提案する。

全体グラフで次数中心性が高いノードの緩和を考えたエゴグラフの改善案の 1 つとして、ダミーのノードを追加する手法について述べる。入力症状を中心ノードとしたエゴグラフに対して、中心ノードに対応する症状と同じ症状に対応するノードを新たに追加し、これをダミーノードとする。このダミーノードに対して中心ノード以外のノード（alter）からエッジを張る。張ったエッジに対して、そのダミーノードに接続された alter の全体グラフでの次数中心性の値で重みをつける。これによって、全体グラフでの次数中心性の高いノードほどダミーノードとの接続が強くなるようなグラフ構造を構築することができ

る。ダミーノードの追加及び、エッジの接続例を図 4 に示す。

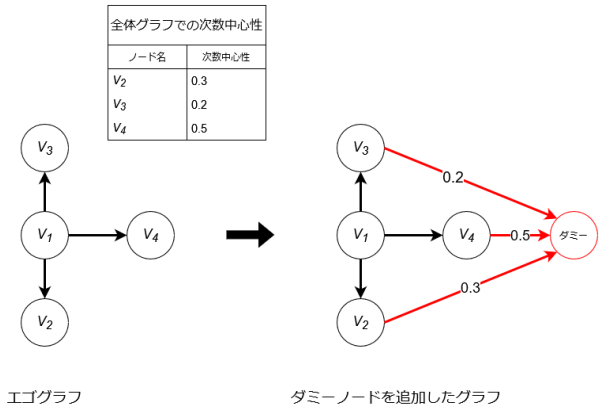

図 4　ダミーノードを追加する手法

4.5.3　入力症状にエッジを張る手法

次に、全体グラフで次数中心性が高いノードの緩和を考えたエゴグラフの改善案の 1 つの、入力症状にエッジを張る手法について述べる。入力症状を中心ノードとしたエゴグラフに対して、中心ノード以外のノード（alter）から、中心ノードに向けてエッジを張る。張ったエッジに対して、そのダミーノードに接続された alter の全体グラフでの次数中心性の値の逆数で重みをつける。また、元から中心ノードに向けられたエッジがある場合の重みは次数中心性の値の逆数で上書きする。これによって、全体グラフでの次数中心性の高いノードほど中心ノードとの接続が強くなるようなグラフ構造を構築することができる。入力症状にエッジを張る手法におけるエッジの接続例を図 5 に示す。

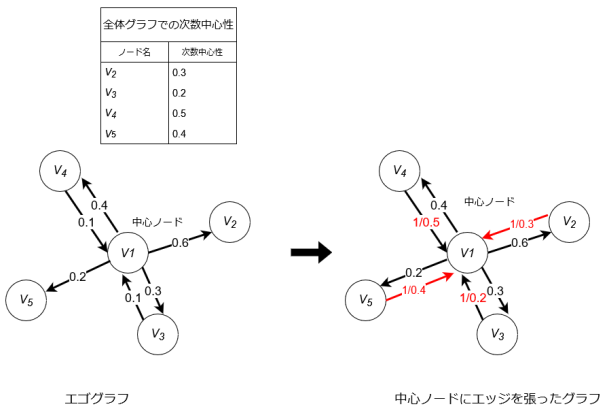

図 5　入力症状にエッジを張る手法

4.6　固有ベクトルを用いたスコアリング

4.5.1 で作成したエゴグラフはグラフ構造であるため、そのエゴグラフ内の症状の関連を特定の病気と対応付ける。病気と対応付けにおいて、エゴグラフ内のノードに対してどの症状をピックアップするかを決める必要がある。そこで、エゴグラフ内のノードに対して、スコアリングを行いランキング形式で出力する。ランキング形式で出力された各ノードに対応した症状とそのスコアを利用して病気と対応付けを行う。

グラフ構造及びリンク構造のランキング手法には様々な手法がある。本研究では、グラフ構造のランキング手法として HITS アルゴリズムを用いる。HITS アルゴリズムによってエゴグラフ内のノードに対してスコアリングを行う。しかし、病気には様々な症状があり、エゴグラフ内のノードに対するスコアリングでは、必ず順位が低くなる症状が存在する。そのような症状やその症状と接続しているノードに対しても、対応した病気が存在する可能性もある。そのため、1 つのグラフ構造で複数種類のスコアリングができないかを考えた。そこで、HITS アルゴリズムによって求まるスコアがハブ行列の絶対値最大の固有値に対応した固有ベクトルの値と一致することに着目した。処理としては、エゴグラフ内のノードを重み付きの隣接行列として表現し、その重み付きの隣接行列を HITS アルゴリズムにおけるハブ行列に変換する。このハブ行列の固有ベクトルを求め、固有ベクトルの各要素を対応した各症状のスコアとする。ここでグラフ構造の画像とその固有ベクトルをヒートマップとして可視化した画像を図 6 に示す。図 6 のヒートマップは各行が固有ベクトルに対応し、各列が、症状に対応した固有ベクトルの要素となっている。また、各固有ベクトルは固有値の絶対値に準じて上から降順で並び替えている。

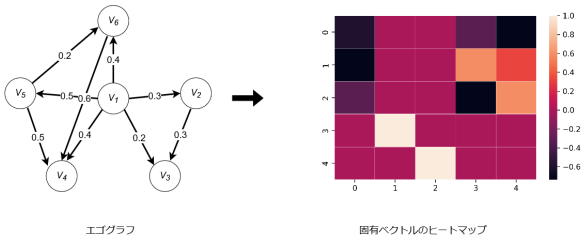

図 6　固有ベクトルのヒートマップ

図 6 のヒートマップより、行番号 0 と行番号 3 を例にとると、行番号 0 と行番号 3 で値の大小関係が大きく変わっていることがわかる。また、固有ベクトル内の要素の分布としても、行番号 3 では、1.0 付近の値がに対し、行番号 0 では、0.0 付近の値が多くなっており、1.0 付近の要素がない。本研究では、この固有ベクトルの分布の違いや大小関係の違いを利用し、グラフ構造におけるノードに対応した症状のスコアリングを行う。また、固有ベクトルは、非主固有値に対応した固有ベクトルの要素に負の値が含まれる場合がある。HITS アルゴリズムにおける固有ベクトルの性質として、同一の固有ベクトル内で正に大きな値と負に大きな値では、それぞれの値に対応したノードが離れている傾向にあると示されている [6]。そこで本研究では、固有ベクトル内で正に大きな値を持つノードと負に大きな値を持つノードで関連のある症状を別々で捉え出力する。この固有ベクトル内の正と負の要素に分けて、対応したノードの出力を考慮することで、グラフ上のノード間の隣接の近いノードごと

に関連する病気に対応付けることができる。また、4.5.2 節、4.5.3 節で述べた手法では、ダミーのノードを追加したことで、固有ベクトルの要素数及び固有値の数が 1 増える。そのため、ダミーのノードを追加する手法では、ダミーノードに対応した行と列を削除し、既存の症状に対応した値のみを出力する。

5　結果・評価

5.1　評価データの作成

本研究では、Q & A サイトにおける質問データを利用し、病気の関連症状の出力するための手法を提案した。それに伴って、関連症状の出力が実際の病気に対する症状が出力できるかの評価を行う。本節では、本研究で使用する評価データの作成及び詳細について示す。

関連症状の評価において、関連症状の出力が実際の病気に対応した症状が出力できているかを検証するためには、評価データは医学的根拠のあるデータを使用する必要がある。そこで本研究では、Web サービスである病院検索 i タウンの病気辞典 [家庭の医学] [9] を用いる。病気辞典 [家庭の医学] の “症状から病気を調べる”という項目のページに記載される内容から一部利用し評価データを作成する。このページは、大きく分けられた症状のある部位のカテゴリから、具体的な症状を選択し、その症状から予想される病気が記載されており、症状の関連を利用して病院の検索を行うことのできる Web サービスである。また、この病気辞典 [家庭の医学] の病気と症状の関連は法研六訂版家庭医学大全科 [10] を用いて作成されている。病気辞書では大きく 7 つのカテゴリ分けがされており、本研究では、“ 子どもの症状 ”、“ 女性特有の症状 ” を除いた 5 つのカテゴリを用いる。本研究で使用する 5 つのカテゴリは、“ 頭・顔の症状 ”、“ 全身の症状 ”、“ 皮膚の症状 ”、“ 上半身の症状 ”、“ 下半身の症状 ” である。また、各カテゴリの中にベースとなる症状が複数存在し、その症状と関連症状から疑われる病気が記載されている。“ 頭・顔の症状 ” カテゴリにおける “ 発熱 ” というベースとなる症状を例に出し、病気辞書の表記例を表 3 に示す。

表 3　病気辞書の表記例

症状 1	症状 2	症状 3	疑われる病気
高熱	咳	胸痛、呼吸困難	細菌性肺炎
		夜中に頑固な激しい咳	マイコプラズマ肺炎
		頭痛、血淡	オウム病
	のどの痛み、全身倦怠感		急性扁桃炎
	全身倦怠感、筋肉痛		インフルエンザ

しかし表 3 のような表記のままでは検証を行う際の症状の検索時に手間がかかってしまうことや、症状の組み合わせによっては、重要な症状がない場合でもその病気に該当してしまう可能性がある。そこで、本研究では、表 3 のようなデータを必須症状と選択できる症状の 2 つに分けて、すべての症状の組み合わせを列挙するような形に変換し、この症状＋病気（単数, 複数）をまとめて評価データを作成する。例えば表 3 症状 1 の “ 高熱 ” という症状は、表 3 に含まれる 5 つの病気すべてに必要な病気である。しかし、症状 3 の “ のどの痛み、全身倦怠感 ” という症状はどちらかを満たせばその疑われる病気に該当する。このように本研究では、同じマスに記述される症状を選択できる症状とし、すべての組み合わせをリストにする。これによって、評価時にどの症状が含まれていれば該当するかという評価を行うことが出来るデータとなる。表 3 から作成した評価データの例を表 4 に示す。

表 4　評価データの例

症状 1	症状 2	症状 3	疑われる病名
高熱	咳	胸痛	細菌性肺炎
高熱	咳	呼吸困難	細菌性肺炎
高熱	咳	頭痛	オウム病
高熱	咳	血淡	オウム病
高熱	のどの痛み		急性扁桃炎
高熱	全身倦怠感		急性扁桃炎
高熱	全身倦怠感		インフルエンザ
高熱	筋肉痛		インフルエンザ

表 4 のように病気＋症状（単数, 複数）の形式でまとめ、評価データとする。また、表 4 でもあるように、病気に対応した症状のリストは複数存在することになるため、どの症状のリストを満たしてもその症状のリストがどの病気に対応したかが評価できる形式となっている。

病気辞書 [家庭の医学] における 5 つのジャンル、31 のベース症状を症状パターンのリストにまとめた。結果としては、病気辞書から合計 1949 件の症状リストを作成することが出来た。病気の種類数は 617 種となった。各症状は検索しやすいように、細かすぎる表記は避けるように記載を行った。また、多くの症状には具体的な症状名があり、検索することで病気を絞り込むことができるように表記した。

5.2　評価方法, 実験設定

4.5.1 節で述べたエゴグラフ及び応用のサブグラフを用いて 4.6 節で述べた手法でスコアリングを行い、各ノードに対応した症状のスコアをと 5.1 節で作成した評価データを用いて、関連症状の評価を行う。評価方法として定量的評価では、MRR という評価指標と 1 つでも病気を出力できた入力症状の割合によって評価を行う。MRR（Mean Reciprocal Rank）とは、ランキングにおいて、初めて正解を出力した順位の平均値である。また、1 つでも病気を出力できた入力症状の割合をカバー率と呼ぶ。また、本研究では固有値の絶対値の降順をランキングに当たる。そのため、絶対値最大の固有値に対応した固有ベクトルから降順で、その固有ベクトルの各要素に対応した症状とそのスコアを用いて評価データとの一致を行う。評価データとの一致における一連の処理を以下の図 7 に示す。また、図 7 における評価データの病気に対応した症状の集合に含まれるかの判定では、入力症状を必ず含むものに絞る。次に本研究の評価における実験設定について述べる。まず、入力症状から作成するエゴグラフは半径が 1、エゴグラフの形状は、標準のエゴグラ

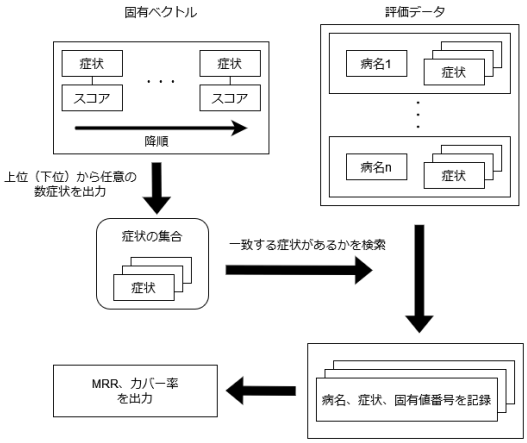

図 7　評価データとの一致

フに加えて、4.5.1 節で述べた 3 つの手法から得られるサブグラフの 4 つのグラフを用いる。標準のエゴグラフを手法 1、ダミーノードを追加する手法を手法 2、入力症状にエッジを張る手法を手法 3、入力ノードを削除する手法を手法 4 とする。入力症状は、全体グラフ内で出現数の多い症状を利用する。出現数は相関ルールにおける前提部、結論部をノードごとにカウントしたものである。この出現数が 100 以上の 109 件のノードを入力値とした。また、評価データとの一致において、入力症状を必ず含むものに絞るように制限しているため、入力値となる症状が評価データ自体にない場合、そもそも対応した病気が存在しないため、入力値となるノードは、評価データに含む症状のみを使用する。固有ベクトル内の要素から上位 3、5、7、9 件出力する。上位からの要素数は固有ベクトル内の要素について、正の要素、負の要素に分けて評価を行う。また、それぞれの符号の要素が上位 (下位) からの要素数以下の場合はその要素数を使用する。ある固有ベクトル内の入力症状に対応した要素が正の場合その固有ベクトルは上位からの出力となり、入力症状に対応した要素が負の場合は下位からの出力となる。また、入力症状が固有ベクトル内の要素において、指定した要素数以下の要素番号となった場合はその固有値での固有ベクトルは MRR の出力にはカウントしないものとする。

5.3　評 価 結 果

まず、手法 1 の評価結果について述べる。正の要素を用いた評価結果を表 5、負の要素を用いた評価結果を表 6 に示す。

表 5　カバー率と MRR(手法 1, 正)

上位からの要素数	カバー率	MRR
3	0.495	0.378
5	0.688	0.457
7	0.752	0.467
9	0.807	0.471

固有ベクトルを用いたランキングにおいて最初に病気を該当できた固有値の番号を集計し、ヒストグラムで可視化する。手法 1 を用いて、上位 (下位)5 件の場合での最初に病気を該当できた固有値番号のヒストグラムを、正、負に分けて図 8、図 9 に示す。

表 6　カバー率と MRR(手法 1, 負)

下位からの要素数	カバー率	MRR
3	0.660	0.607
5	0.816	0.695
7	0.899	0.774
9	0.917	0.755

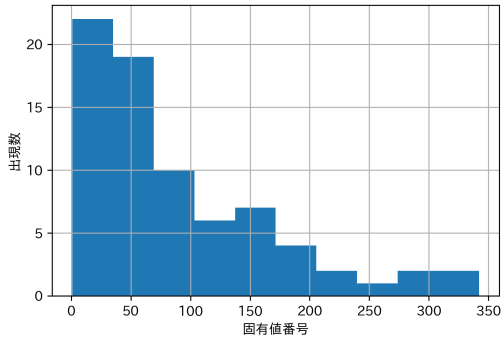

図 8　固有値のヒストグラム (手法 1, 正)

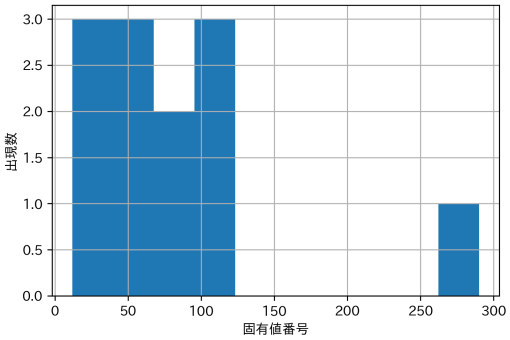

図 9　固有値のヒストグラム (手法 1, 負)

表 5 より、正の要素、負の要素の上位下位から任意の数出力した場合、すべて要素数において負の要素から出力した方がカバー率が高くなった。これは [6] でも示されるように、固有ベクトル内で正に大きな値を持つ要素は繰り返し現れる傾向にあるため、同じ症状に対応した症状が固定されてしまい、対応した病気との一致が減ったと考えられる。それに対し負の要素を用いた場合のカバー率が高くなっていることから、負に大きな値を持つ要素は固有値ごとに様々な症状が低い値になっていることが予想される。上位からの要素数の違いでは、正と負どちらに関しても、要素数を増やすごとによるカバー率が高くなるが、要素数を増やすごとに緩やかになる。

次に、4.5.2 節及び 4.5.3 節で述べたエゴグラフの応用である 2 つの手法を用いた評価結果について述べる。手法 2、手法 3 によるカバー率と MRR を正、負に分けて表 7 から表 10 に

表 7　カバー率と MRR(手法 2, 正)

上位からの要素数	カバー率	MRR
3	0.559	0.413
5	0.706	0.452
7	0.779	0.477
9	0.807	0.473

表 8　カバー率と MRR(手法 2, 負)

下位からの要素数	カバー率	MRR
3	0.633	0.574
5	0.853	0.689
7	0.899	0.751
9	0.926	0.755

表 9　カバー率と MRR(手法 3, 正)

上位からの要素数	カバー率	MRR
3	0.504	0.376
5	0.688	0.427
7	0.752	0.465
9	0.816	0.462

表 10　カバー率と MRR(手法 3, 負)

下位からの要素数	カバー率	MRR
3	0.587	0.391
5	0.770	0.472
7	0.880	0.514
9	0.917	0.524

示す。

手法 1 から手法 3 において、正の要素を用いた場合では、上位からの要素数 3 件を例に出すと、手法 2 が 0.559 と最もカバー率が高く、病気に対応した症状が上位に上がりやすいことがわかる。次に手法 3 となっており、エゴグラフをそのまま使用するよりは病気に対応した症状が上位に上がりやすい。また、負の要素を用いた場合でも、手法 2 が 0.633 と最もカバー率が高く、標準のエゴグラフを用いた場合が最も病気に対応した症状が上位に上がりやすい結果となった。MRR では、正方向と負方向で差があり、上位からの要素数が 3 で考えた場合、正方向では手法 2、負方向では手法 1 の MRR が最も高くなった。

6 ま と め

本研究では、質問データ上の症状の共起を、相関ルールで数値化し、グラフ構造として表現した。そして、患者の自覚症状である入力値の周辺のグラフを抽出し、HITS アルゴリズムの考えの基、固有ベクトルでスコアリングすることで、入力値周辺の症状に対して、複数パターンのスコアリングを行った。また、入力値周辺のグラフに対して、ダミーノードの追加や入力値に向けたエッジの追加など複数の手法を提案し関連症状の評価を行った。カバー率では、手法 2 が最も高く、病気に対応した症状が固有ベクトル内の上位 (下位) に集まりやすく、より症状間の関連性を表現することができ、病気に対応した症状の出力が行える共起グラフを構築できた。また、固有ベクトル内の要素を正と負で分けて捉えることで、固有ベクトルの要素がグラフ構造のエッジの有無やエッジの重みから症状間の関連性を表現し、エゴグラフの中でも病気ごとのコミュニティを発見することができた。

謝　　辞

本研究では、国立情報学研究所の IDR データセット提供サービスによりヤフー株式会社から提供を受けた「Yahoo! 知恵袋データ（第 3 版)」を利用した。

文　　献

[1] 新本拓也, 湯本高行, 金子周司, 礒川悌次郎, 松井伸之, 上浦尚武 “QA サイトでの共起に基づく患者の自覚症状入力支援”, 情報処理学会研究報告データベースシステム (DBS), 2019.16, pp.1-6, 2019.

[2] Futoshi Ishii, Reiko Hayashi, Emiko Shinohara, Motomi Beppu, “Application of Network Analysis to Multiple Causes of Death Data in Japan”, The 29th International Population Conference ,2021.

[3] 奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室, “患者表現辞書”, https://sociocom.naist.jp/patient-dic/

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” ,Proc. NAACL 2019, pp.4171–4186.

[5] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami,“Mining Association Rules Between Sets of Items in Large Databases”, Proc. ACM SIGMOD, pp. 207-216, 1993.

[6] Jon Michael Kleinberg, “Authoritative sources in a hyperlinked environment” , Journal of the ACM, Vol 46, pp.604—632, 1999.

[7] 国立情報学研究所, “ 情報学研究データリポジトリ Yahoo!知恵袋 (第 3 版) ”, https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html

[8] 岡部貴博, 吉川大弘, 古橋武, “メタデータと語句の共起情報を利用したインシデントレポート解析システムの提案”, 知能と情報 (日本知能情報ファジィ学会誌)vol.18,No.5,pp.689-700,2006

[9] NTT タウンページ株式会社, “病気事典 [家庭の医学]”, https://medical.itp.ne.jp/byouki/shoujou/

[10] 株式会社法研関西, “法研 六訂版 家庭医学大全科”, 株式会社法研関西