Vol.2011-BIO-26 No.1
2011/9/13

生物医学要素関係間の意味的類似度に基づく仮説の順位付け

宮西大樹[†1]　関　和広[†1]　上原邦昭[†1]

近年，生物医学分野で発表される文献の数は急増しており，その結果，研究者個々人が自身の専門分野のすべての情報を理解し統合することは困難になっている．そのため，未だ発見されていない知識（仮説）が大量の文献の中に埋もれていると考えられる．事実，そのような仮説が複数報告され，一部は臨床的にも証明されている．本研究では，意味的に類似したイベントから生成された仮説はより妥当であると仮定し，仮説を構成するイベント間の意味的な類似度を用いて仮説の妥当性を定義する．そして，この妥当性に基づいて仮説を順位づけることで，真に重要な仮説の効率的な発見を目指す．従来用いられてきた頻度に基づく手法と提案手法を比較し，意味的な類似性が妥当な仮説を同定する際に有効であることを示す．

Hypothesis Ranking Based on Semantic Event Similarities

Taiki Miyanishi ,[†1] Kazuhiro Seki [†1]
and Kuniaki Uehara [†1]

Accelerated by the technological advances in the biomedical domain, the size of its literature has been growing very rapidly. As a consequence, it is not feasible for individual researchers to comprehend and synthesize all the information related to their interests. Therefore, it is conceivable to discover hidden knowledge, or *hypotheses*, by linking fragments of information independently described in the literature. In fact, such hypotheses have been reported in the literature mining community; some of which have even been corroborated by experiments. This paper mainly focuses on hypothesis ranking and investigates an approach to identifying reasonable ones based on semantic similarities between events which lead to respective hypotheses. Our assumption is that hypotheses generated from semantically similar events are more reasonable. We developed a prototype system called, *Hypothesis Explorer*, and conducted evaluative experiments through which the validity of our approach is demonstrated in comparison with those based on term frequencies, often adopted in the previous work.

1. はじめに

生物医学分野で発表される文献の数は週に数千にのぼり，研究者個々人が自身の専門分野のすべての情報を理解することは難しい．ましてそれらの情報を有機的に統合し，新たな知識を包括的に導き出すことは事実上不可能である．[1)–3)]．そのため，未だ発見されていない潜在的な知識（これを「仮説」と呼ぶ）が大量の文献の中に埋もれていると考えられる[4),5)]．

このような仮説を発見する先駆的な研究の例として，Swanson が同定した魚油とレイノー病の関係がある．Swanson は，文献を個々に手作業で調べ上げ，抽出した知識を組み合わせることにより，魚油がレイノー病の治癒に効果的であることを予測した．この関係は，後年実験的に証明されている[6)]．Swanson の発見以降，Swanson が手作業で行った仮説生成を自動化することで，大量の文献の中から科学的発見につながるような未知の知識を自動的に同定する試みが複数の研究グループによって行われている．しかし，これらの手法は仮説を生成するために人手を要したり，頻度を基にした手法であるため低頻度の概念に対処できないといった問題がある．例えば，後者については，低頻度の概念や関係によって導出された仮説は（不当に）低く評価されてしまう傾向がある．そのため，仮説発見を行う枠組みは，概念および関係の頻度に依存しない方法で構成されるべきである．

これらの問題に対処するため，本研究では，文献から抽出した概念間の関係（イベント）から仮説を自動的に生成し，イベント間の意味的な類似度を用いることで，頻度に依存しない仮説の順位付けを行う．これにより，低頻度の概念および関係から導出される重要な仮説が提示されにくくなることを防ぐ．

以降では，2 節で関連する研究と我々の提案手法の位置づけについて説明し，3 節で仮説を生成する手法とイベントの類似度に基づく順位付けの方法について述べる．4 節では，提案手法である意味に基づく手法と従来の頻度に基づく手法との比較実験を行い，仮説の順位付けについて評価する．最後に，5 節で結論と今後の課題について述べる．

2. 関連研究

前述の Swanson による仮説生成の手法は ABC モデルと呼ばれている．このモデルは，概念 A と概念 B，概念 B と概念 C の関係が文献中に明示的に示されており，かつ概念 A

†1 神戸大学 システム情報学研究科
Graduate School of Systems Informatics, Kobe University

と概念 C の関係を示す文献が存在しない場合，概念 A と概念 C の間には潜在的な関係があるとして仮説を生成するモデルである．このモデルは単純ではあるものの，2 つの異なる専門分野の論文から抽出した関係を仮説生成に使用する場合や，両概念の関係が 1 つの分野中にあるにもかかわらず，文献の数が膨大なために両者の関係を調べることが困難な場合など，人手による作業が容易でない場合に有効である．大量の文献が存在し，専門分野が多岐にわたる生物医学分野においては，上記どちらの状況も十分に考えられる．

ABC モデルを基に， Swanson を含むいくつかの研究グループが仮説生成の自動化を試みている[7)–17)]．この内，本研究に最も関連する Weeber と Srinivasan の 2 つの研究について紹介する．

Weeber は，自然言語処理の技法を用いて仮説発見を支援するための DAD システムを開発した[15)]．Weeber のシステムと他のシステムとの違いは，仮説の提示やフィルタリングを行う際に，UMLS（Unified Medical Language System）メタシソーラス[⋆1]を用いた点である．Swanson が考案した MEDLINE のレコードにある文から語彙や句を抽出する手法[13)]とは違い，Weeber は MetaMap[18)] というツールを用いた．MetaMap は NLM（National Library of Medicine）で開発されている生物医学関連の文章から生物医学要素を同定するためのツールである．Weeber らはこのツールを用いて MEDLINE のレコード上に記述された生物医学要素と UMLS のシソーラス中にある概念との対応付けを行った．例えば，「Platelet aggregation is known to be high in patients with Raynaud's syndrome.」という文は以下の 5 つの概念，「Platelet aggregation」，「Known」， 「High」，「Patients」，「Raynaud's disease」に対応づけられる．MetaMap を用いれば，同義語，複数形，語尾変化などの語の変異形を 1 つの概念に対応付けることができる．また，UMLS のメタシソーラス中において，各概念は「Body location or region」，「Vitamin」，「Physiologic function」などの意味タイプと呼ばれる意味カテゴリー 1 つ以上に属し，概念が MetaMap により特定された後，この意味タイプを用いてフィルタリングを行う．このフィルタリングの段階で，特定の意味タイプだけを用いることで，探索する仮説の数を劇的に減らすことができ，生成される仮説数を制限することができる．

Srinivasan は Manjal[⋆2] と呼ばれる仮説発見のシステムを開発した[12)]．このシステムは，MEDLINE のレコード中の文から概念抽出を行う Weeber の研究とは違い，UMLS の意味タイプと共に，MeSH（Medical Subject Headings）語と呼ばれる MEDLINE の各レコードに付与された索引語を用いている．Srinivasan はこの MeSH 語を使うことで潜在的な仮説の発見を行っている．Manjal では，概念 A を与えることで MEDLINE に対して検索を行い，検索された MEDLINE レコードから MeSH 語を概念 B として抽出し，あらかじめ用意した UMLS の意味タイプに応じてグループ化を行う．そして，Weeber と同様に UMLS の意味タイプを用いてフィルタリングを行い，特定の意味タイプに属する概念だけに限定して仮説を生成する．また，Srinivasan はユーザーに対して潜在的に重要な関係を同定し易くするために，情報検索の分野で文書を特定する目的で用いられる TFIDF（Term Frequency-Inverse Document Frequency）値[19)] を使用し，仮説生成の段階において概念 B と C の各 MeSH 語に対して順位付けを行っている．

本論文における仮説生成の手法は，Weeber の手法を改良したものであり，さらに意味的類似度に基づく妥当性を用いて，生成した仮説の順位付けを行う．また，関連研究との比較のため，意味的類似度に基づく妥当性（提案手法）と TFIDF や関係の出現回数などの頻度に基づく妥当性との比較を行う．

3. 提 案 手 法

この章では，文献から抽出した概念間の関係から生物医学要素ネットワークを構築して仮説の生成を行う枠組みについて述べる．次に，有望な仮説を同定するために仮説の妥当性という尺度を導入し，イベントの意味的類似度に基づく妥当性と意味を考慮しない頻度に基づく妥当性の 2 種類の尺度について定義する．

3.1 仮 説 生 成

仮説を生成するため，まず Weeber と同様の手法で，生物医学文献から固有表現と関係の抽出を行い，生物医学要素ネットワークを構築する[15)]．Weeber は MEDLINE のタイトルとアブストラクトから MetaMap[18)] を用いて UMLS の概念を抽出し，文中の概念の共起を関係として抽出した．Weeber が MetaMap の出力した概念すべてを使用したのに対し，本研究では固有表現抽出の精度を考慮し，MetaMap が概念と共に出力するスコアを基に，最もスコアの高い概念だけを使用する．さらに，意味のない仮説が無数に生成されることを防ぐため，仮説に使う関係は文献のタイトルだけから抽出する．なお，生物医学分野においては，タイトルは非常に記述的であり，文献の内容を高精度に要約していることが報告されている[20)]．また，文献のタイトル中に記述された関係は経験的に肯定的な関係が大部分であるため，概念の共起による関係抽出を使用しても否定と肯定の関係が混在する仮説が生成さ

⋆1 http://www.nlm.nih.gov/research/umls/
⋆2 http://sulu.info-science.uiowa.edu/Manjal.html

れにくいと考えられる．

Weeber は仮説生成の段階で UMLS の意味タイプによって使用する概念を制限し，専門家の知識によって中間の概念と終端の概念に対して異なる意味タイプを用いている．一方，本手法は一貫して同一の意味タイプを使用しており，生物医学関連の知識に乏しくても仮説を生成することができる．

以降，文献のタイトル中にある生物医学要素の 2 項関係をイベントと呼ぶことにする．本研究では，仮説生成の簡素化のため，イベントがどのような関係であるかについては考慮しないことにする．共通の概念を基に，抽出したイベントを結合することにより構築した生物医学要素ネットワークを図 1 に示す．この例では，文献 d_1 から概念 c_1 と c_2 に基づくイベント c_1-c_2，文献 d_2 からはイベント c_3-c_5，文献 d_3 からはイベント c_1-c_3 をそれぞれ抽出している．抽出したイベントに ABC モデルを適用すると，イベント c_1-c_3 と c_3-c_5 より，潜在的な関係 c_1-c_5 を新たに発見することができる．このような関係 c_1-c_5 は，個々の文献 d_2，d_3 に記述された関係を統合することでしか発見できない．

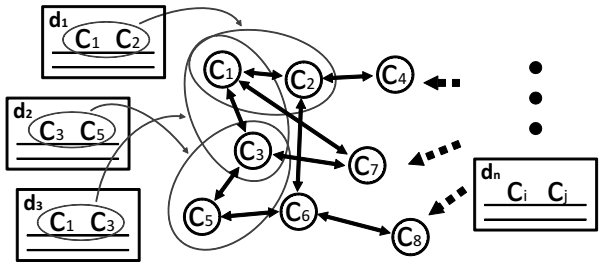

図 1　文献から抽出した関係を用いて構成した生物医学要素ネットワーク

仮説の生成は，所望の生物医学要素を始点として，生物医学要素ネットワークを探索することで行う．探索を終了したノードを終点とし，始点と終点を結ぶ関係が仮説となる．ただし，始点と終点の間には直接的なつながりはないものとする．このとき，始点となるノードを A-term，始点と終点を結ぶ中間のノードを B-term，終点となるノードを C-term と呼ぶことにする．図 2 に，図 1 の概念 c_1 を始点とした探索により取得できるパスを A，B，C-terms の順番に示す．

3.2　仮説の妥当性

生成された仮説は，実験と検証による過程を経て正当と認められることにより新たな知識

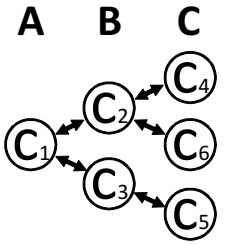

図 2　A，B，C-terms の例

となる．本研究の目的は，自動的に導出した無数の仮説から妥当な仮説を見つけ出すことである．上述のように，本論文では生物医学要素の 2 項関係をイベントと呼び，共通する要素を持つ 2 つ以上のイベントから導出される新たなイベントを仮説とした．ここで，仮説の妥当性を測るために，意味的に類似するイベントから導出される仮説はより妥当であるという仮定を置く．逆に，類似していないイベントから導出される仮説には意味的な飛躍があるため，その仮説の論理的解釈が難しいと考えられる．この仮定を基に，仮説の妥当性の尺度としてイベント間の類似度を定義する．イベント間の類似度には，人手で体系化された MeSH と呼ばれるシソーラスを使用する．

このイベントを特徴付けるために使用する MeSH 語は，MEDLINE の各レコード（論文）を索引付けするため，人手により通常十数個ほど付与されている．2009 年の時点では，25,186 個の定義語が存在し，意味的な構造を持つ 11 の階層に整理されている．この MeSH の階層構造の中では，上位にある語ほど一般的な意味を持ち，下位に行くほど厳密な意味を持つ語となる[21]．

なお，MeSH は文献を特徴付けるための索引語であり，必ずしも文献から抽出されたイベントを特徴付けるものではない．しかしながら本研究では，文献の内容を最も簡潔に表現したタイトルだけからイベントを抽出することで，文献に付与された MeSH 語をイベントの特徴と見なす．そして，イベントの特徴として MeSH 語間の類似度を定義し，さらにこれをイベント間の類似度へと拡張する．

3.3　概念間の類似度

提案手法では，イベントに対応する MeSH 語同士の類似度を測るために，各 MeSH 語同士の類似度を求め，それらの類似度の総計を最終的な類似度とする．現在まで，MeSH のようなシソーラスにおいて，概念間の類似度を測定する方法が数多く提案されている[22)–25)]．

この内，我々は概念間の類似度を測る手法として，Seco ら[25] が提案した手法を用いる．この手法の特徴は，シソーラスの構造を用いることで概念間の類似度を測ることにある．よって，頻度に依存することなく，概念間の類似度を計算することができる．

Seco らは，概念が持つ厳密さを概念が持つ情報量により定義した．シソーラス上では，上位の概念は一般的な語彙であるため情報量が少なく，下位の概念は厳密な語彙であるため情報量が多くなる特性がある．彼らはこの特性を利用して，シソーラス上の位置に応じた情報量を定め，概念間の共通祖先の情報量を概念間の類似度とした．これは，互いに類似している概念同士ほど下位に共通の祖先を持ち，類似していないときほど上位に共通の祖先を持つことを意味する．以下に概念間の情報量 (IC) を用いて MeSH 語 m_1，m_2 間の意味的類似度 $\mathrm{sim}(m_1, m_2)$ を定義する．

$$\mathrm{sim}(m_1, m_2) = \max_{m \in S(m_1, m_2)} \mathrm{IC}(m) \tag{1}$$

$$\mathrm{IC}(m) = 1 - \frac{\log(\mathrm{hypo}(m) + 1)}{\log(\mathrm{N_s})} \tag{2}$$

ここで，$\mathrm{IC}(m)$ は MeSH 語 m のシソーラス上における情報量であり，$S(m_1, m_2)$ は m_1 と m_2 間の共通の祖先となる概念の集合を表している．また，$\mathrm{hypo}(m)$ はシソーラス上での語 m 以下の下位語の数であり，$\mathrm{N_s}$ はシソーラス中にある語の総数である．式 2 の分母は最小の情報量をもつ概念の値と等しく，IC の値が 0 から 1 の間になるように正規化する．また IC は概念の一般性に従い単調に減少する．なお，シソーラスの最上位にある仮想的な概念の情報量は 0 である．

3.4 イベント間の類似度

式 1 で定義した語間の類似度をイベント間の類似度へと拡張し，この類似度を仮説の妥当性と見なす．仮説の妥当性とイベント間の類似度の関係を図 3 に示す．

図 3 において，イベント c_1-c_3 と c_3-c_5 の類似度はイベント c_1-c_2 と c_2-c_6 よりも高いものとする．この場合，前者のイベントから生成された仮説 c_1-c_5 の妥当性は，後者のイベントから生成された仮説 c_1-c_6 よりも高くなる．次項で概念間の類似度を拡張することで得られるイベント間の類似度について説明する．

3.4.1 類似度平均による仮説の妥当性

以下に示す類似度の定義は，イベントに対応する MeSH 語間の類似度をすべて求め，その平均をイベント間の類似度としたものである．このイベント間の類似度を当該イベントによって導出される仮説の妥当性（reasonability）と見なす．

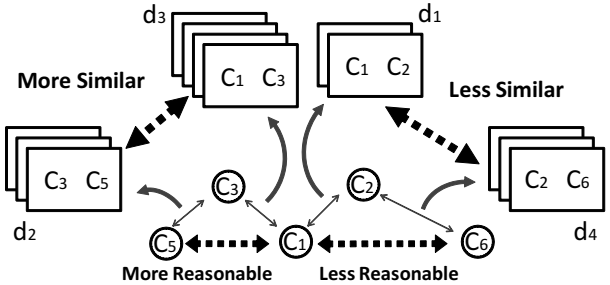

図 3 イベントの類似度と仮説の妥当性との関係

$$\mathrm{R_{avg}}(e_i, e_j) = \frac{1}{|\mathbf{M}_i||\mathbf{M}_j|} \sum_{m_k \in \mathbf{M}_i} \sum_{m_l \in \mathbf{M}_j} \mathrm{sim}(m_k, m_l) \tag{3}$$

ここで，$\mathbf{M}_i$ と $\mathbf{M}_j$ はそれぞれイベント e_i，e_j に対応する MeSH 語の集合を表している．MeSH 語の集合 $\mathbf{M}$ は，イベントの抽出元である文献に付与された MeSH 語を重複しないように集めたものである．この妥当性 $\mathrm{R_{avg}}$ の定義の欠点は，似ていない概念同士の類似度が影響を持ち易いことである．この点については図 4 で議論する．

3.4.2 類似概念に注目した仮説の妥当性

次の妥当性 $\mathrm{R_{max}}$ の定義は，最も類似する概念だけに着目することで，上述した $\mathrm{R_{avg}}$ の問題に対処できるように定義している．イベント e_i に対応する各概念に対して，もう一方のイベント e_j にある概念と最も類似する概念を選び，互いの類似度の平均をもとめる．そして，イベントに対して対称になるように，e_i と e_j を入れ替えて同様の計算を行い，最終的な e_i と e_j の類似度をもとめる．

$$\mathrm{R_{max}}(e_i, e_j) = \frac{1}{|\mathbf{M}_i|} \sum_{m_k \in \mathbf{M}_i} \max_{m_l \in \mathbf{M}_j} \mathrm{sim}(m_k, m_l) + \frac{1}{|\mathbf{M}_j|} \sum_{m_l \in \mathbf{M}_j} \max_{m_k \in \mathbf{M}_i} \mathrm{sim}(m_k, m_l) \tag{4}$$

3.4.3 TFIDF 値に基づく仮説の妥当性

意味的類似度に基づく妥当性と比較するため，テキスト間の類似度を測るためによく使われる TFIDF とコサイン類似度を用いた妥当性を定義する．この定義は，Srinivasan の手法[12] に相当する．イベントに対応する MeSH 語を文献から抽出する際，イベントを複数の

文献から抽出することができる場合がある．そこで，重複した MeSH 語の数をその MeSH 語の TF（単語頻度）値とし，ある MeSH 語が付与された文献の数をその MeSH 語の DF（文書頻度）値とする．そして，この TF 値と DF 値を用いて MeSH 語の TFIDF 値を求める．イベント e_i に対応する TFIDF による重み付けした MeSH 語ベクトル $\mathbf{TFIDF}(e_i)$ を以下に示す．

$$\mathbf{TFIDF}(e_i) = (w_{i1}, w_{i2}, \cdots, w_{in})$$

ここで，$w_{ij} = n_{ij} \times \log(N/n_j)$ であり，イベント e_i に対応する j 番目の MeSH 語 m_{ij} の重み（TFIDF）を表している．N はデータベース中にある文献の総数，n_j は MeSH 語 m_j を付与した文献の数であり，これが DF 値となる．n_{ij} はイベント e_i の抽出元となった文献中で MeSH 語 m_j を含む文献の数であり，これが TF 値となる．この MeSH ベクトルを用いることで，イベント e_i と e_j とのコサイン類似度をもとめ，イベントから導きだされた仮説の妥当性 $\mathrm{R_{tfidf}}$ とする．

$$\mathrm{R_{tfidf}}(e_i, e_j) = \frac{\mathbf{TFIDF}(e_i) \cdot \mathbf{TFIDF}(e_j)}{|\mathbf{TFIDF}(e_i)||\mathbf{TFIDF}(e_j)|} \quad (5)$$

3.4.4 頻度による仮説の妥当性

頻度に基づくもう一つの定義として，イベントの頻度を利用した妥当性を次のように定義する．

$$\mathrm{R_{freq}}(e_i, e_j) = \sqrt{freq(e_i) \times freq(e_j)} \quad (6)$$

ここで，$freq(e)$ はイベントの頻度を表す．この定義の直感的解釈は，頻繁に言及されるイベントから構成される仮説は妥当である，というものである．

4. 評　　価

4.1 実験の設定

1986 年に Swanson が発見した「レイノー病の症状改善に魚油の摂取が有効である」という仮説を用いて，評価実験を行なう．Swanson は，レイノー病患者には，高い血液粘性，強い血小板凝集作用，および血管収縮などの血液反射に関する特徴がみられること，また魚油が血液粘性，および血小板凝集作用を下げる働きがあることを人手で文献から調べ上げ，魚油とレイノー病の関係を予測した．Swanson によって発見されたこの仮説を提案手法によって生成し，妥当な仮説を優先的に同定できるかを検証することにより，仮説の妥当性の尺度を比較評価する．

Biologic Function, Cell Function, Disease or Syndrome, Lipid, Molecular Function, Organ or Tissue Function, Organism Function, Pathologic Function, Physiologic Function

図 4　実験に利用した UMLS 意味タイプ

Swanson が発見した仮説を再現するため，1960 年から 1985 年の期間に発表された生物医学文献を基に生物医学要素ネットワークを構築し，そこから魚油を所与の概念として仮説を生成する．そして，3.4 節で定義した妥当性の尺度に基づき仮説の順位付けを行う．この順位付けにおいて，意味的類似度に基づく尺度が頻度に基づく尺度より妥当な仮説（B-term として血液粘性・血小板凝集作用・血管収縮などを持つ仮説）を高く順位付けすることができれば，意味的な類似度を考慮することが重要な仮説の同定に効果的であることを示すことができる．

本実験では Weeber が用いた手法[15] にならい，図 4 に示す UMLS の意味タイプを関係抽出の段階で使用した．また，Weeber が用いた意味タイプの内，「Laboratory or Test Result」は今回の実験において関連が薄いことから，概念 Blood Viscosity [Laboratory or Test Result] を Blood Viscosity [Physiologic Function] に置き換えた．これらの条件の下で仮説生成を行い，ノード数が 15,774 個，エッジ数が 193,165 個の生物医学要素ネットワークを得た．

4.2 結果と考察

前節で取得したネットワークを用いて，A-term を魚油とし，深さ 2 で探索を行ったところ，合計で 13,677 個の仮説を得ることができた．その内，C-term がレイノー病である仮説は 8 つ存在した．表 1 に，魚油とレイノー病の関係を表す仮説を A，B，C-terms の順に B-term のアルファベット昇順に並べて示す．ここで，「Primary Raynaud's」と「Paroxysmal digital cyanosis」は「Raynaud's disease（レイノー病）」と同義である．表 1 中にある「Blood Viscosity（血液粘性）」は，上述したように魚油とレイノー病の関係を正当に説明する B-term である．また，「Atheromatosis」と「Peripheral vascular disease」は血小板凝集作用や血液粘性と関係のある語である．よって，$\mathrm{H_1}$，$\mathrm{H_6}$，$\mathrm{H_7}$ は妥当な仮説であると考えられる．一方，「Development」と「Suppression」は魚油とレイノー病の関係を説明するには一般的過ぎる概念であるため，仮説 $\mathrm{H_5}$ と $\mathrm{H_8}$ は有用とは言えない．

つづいて，3.4 節で定義した妥当性の尺度を用いて生成された 13,677 個の仮説に対して順位付けを行った．図 5 に，妥当と判断された各仮説が全体の上位何%に含まれているのかを妥当性尺度ごとに示す．妥当な仮説は高く順位付けされるべきなので，小さい値を持っ

Vol.2011-BIO-26 No.1
2011/9/13

表 1　魚油を所与として生成された仮説

ID	Reasonable	*A*-term	*B*-term	*C*-term
H_1	Yes	Fish Oil	Atheromatosis	Raynaud Disease
H_2	Yes	Fish Oil	Blood Viscosity	Paroxysmal digital cyanosis
H_3	Yes	Fish Oil	Blood Viscosity	Primary Raynaud's
H_4	Yes	Fish Oil	Blood Viscosity	Raynaud Disease
H_5	No	Fish Oil	Development	Paroxysmal digital cyanosis
H_6	Yes	Fish Oil	Peripheral vascular disease	Paroxysmal digital cyanosis
H_7	Yes	Fish Oil	Peripheral vascular disease	Raynaud Disease
H_8	No	Fish Oil	Suppression	Paroxysmal digital cyanosis

ているほど，良い妥当性尺度であるといえる．逆に図 6 は，非妥当な仮説だけを示したグラフであり，仮説が低く順位付けられている（大きな値を持つ）ほど，良い妥当性尺度であるといえる．図 5 と図 6 の最右に示されている Avg は，各々の妥当性尺度によって得られた仮説の平均順位を表している．R_{avg}，R_{max}，R_{tfidf}，R_{freq} は各々図 3.4 で定義した仮説の妥当性尺度である．R_{avg} は仮説を構成するイベントに対して概念間の類似度の平均をもとめたものであり，R_{max} は最も似ている概念間の類似度の平均，R_{tfidf} はイベントに対応する概念を TFIDF で重み付けしたときのコサイン類似度，R_{freq} はイベントの頻度の相乗平均である．

まず，妥当・非妥当すべての仮説について，妥当性尺度の持つ平均的な振る舞いについて議論する．図 5 における平均順位（Avg）から，最も低い値を持つ R_{max} が一番良い結果を示していることが分かる．次に図 6 の平均順位を見ると，意味的類似度を用いた R_{avg} と R_{max} が非妥当な仮説を同程度に低く順位付けており，頻度に基づく R_{tfidf} や R_{freq} と比べて同等かそれ以上の結果を示している．次に，個々の仮説の順位付けについて考察する．Blood Viscosity（血液粘性）に関係する仮説 H_2，H_3，H_4 は妥当な仮説であり，意味的類似度に基づく妥当性尺度 R_{max} と R_{avg} は，このような妥当な仮説を頻度に基づく妥当性尺度 R_{freq} よりも適切に高く順位付けできている．R_{freq} が有効に機能しない原因は，文献から抽出できた魚油と血液粘性の関係と，血液粘性とレイノー病の関係の数がごくわずか（それぞれ 1 と 4 個）であることによる．一方，もう一つの頻度に基づく妥当性尺度 R_{tfidf} は，これら低頻度の概念や低頻度の関係から導出された仮説であっても比較的高い順位を与えることができている．これは IDF の項が低頻度の概念に対して有効に働いたためだと考えられる．また，仮説 H_1，H_6，H_7 については，仮説 H_2，H_3，H_4 と比べて低く順位付けされていることがわかる．これは上述したように仮説 H_1，H_6，H_7 が血小板凝集作用や血液粘

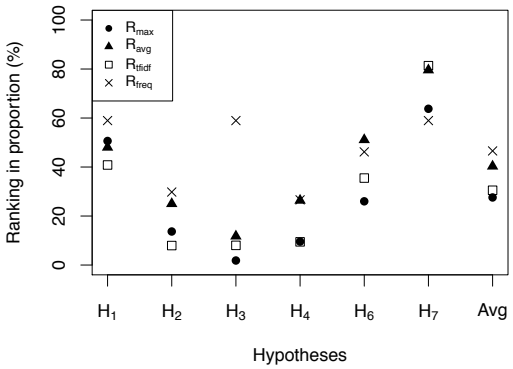

図 5　妥当な仮説の順位（魚油とレイノー病）

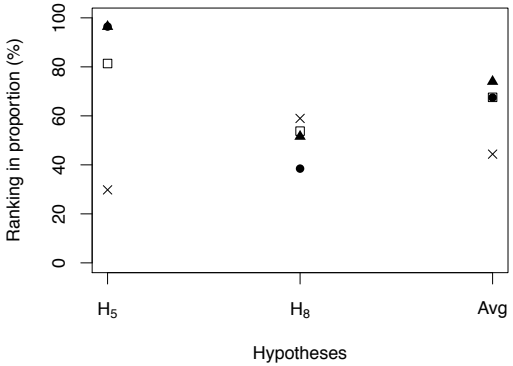

図 6　非妥当な仮説の順位（魚油とレイノー病）

性と関連のある語を含む仮説であるためと考えられる．

非妥当な仮説 H_5 については R_{avg}，R_{max}，R_{tfidf} が仮説を適切に低く順位付けできており（グラフの上側に表れる），R_{freq} の結果は著しく悪いものとなっている．その理由としては，*B*-term が一般的な概念「Development」であるためにイベントの頻度が高くなり，その結果 R_{freq} の性質上，誤って高い妥当性を示すことになったと考えられる．仮説 H_8 については，妥当性尺度の違いによる大きな開きは見られなかった．

次に，意味的類似度に基づく 2 つの妥当性尺度 R_{max} と R_{avg} について比較を行う．非妥当な仮説において両者には明確な違いが観測されなかった．一方，妥当な仮説について

Vol.2011-BIO-26 No.1
2011/9/13

R_{max} は R_{avg} より総じて有効であった．この結果から，R_{max} は類似していない概念を考慮しないことで妥当な仮説を適切に評価できたものと考えられる（3.4.2 節参照）．次に，なぜ R_{max} は R_{avg} と比べて良い結果が得られるのかを仮想的な例によって示す．ある 2 つのイベント e_a と e_b があると仮定する．各イベントは次に示す概念の組み合わせ，{Blood Viscosity, Fish Oil} と {Blood Viscosity, Platelet Aggregation} によって表現されるものとする．式 1 に従うと，イベントを表現する概念のすべての組み合わせに対する類似度は，sim(Blood Viscosity, Blood Viscosity)=1 （Blood Viscosity は下位語を持たないため），sim(Blood Viscosity, Platelet Aggregation)=0.64，sim(Fish Oil, Blood Viscosity)，sim(Fish Oil, Platelet Aggregation)=0 のように計算できる．この場合，イベント e_a と e_b によって導かれた仮説の R_{max} と R_{avg} は，それぞれ $(1+0)/2+(1+0.64)/2=1.32$ と $(1+0.64+0+0)/2\cdot 2=0.41$ になる．さらに，概念「Vascular Disease」を e_a に，概念「Raynaud Disease」を e_b に加えた新たなイベント e'_a と e'_b を仮定する．新たに加えた概念は意味的に類似しているので，イベント e'_a と e'_b の類似度はイベント e_a と e_b の類似度と比較して，大きく変化すべきではない．概念「Vascular Disease」と「Raynaud Disease」を含む概念の類似度は sim(Raynaud Disease, Vascular Diseases)=0.47 であり，イベント e'_a と e'_b の類似度はそれぞれ R_{max}=$(1+0+0.47)/3+(1+0.64+0.47)/3\simeq 1.19$，$R_{avg}=(1+0.64+0+0+0+0+0+0+0.47)/(3\cdot 3)\simeq 0.234$ となる．この例から，R_{avg} は e_a と e_b の類似度に比べて半減してしまっているのに対し，R_{max} はわずかしか減少していないことが分かる．このような R_{avg} の望ましくない振る舞いは，類似していない概念間の類似度が 0 になる場合が多く存在するためである．一方，R_{max} の定義では，類似の概念だけに注目することで，類似していない概念の影響を抑制することができる．

4.3 追加実験

この実験では，Swanson が 1988 年に発見した偏頭痛とマグネシウムの関係[5] を利用し，意味的類似度に基づく手法の有効性をさらに検証する．

4.1 節の実験と同様に，はじめに 1966 年から 1987 年のデータを用いて生物医学要素ネットワークを構築した（偏頭痛とマグネシウムの関係は 1988 年に発見されたため）．また，Weeber と同様の UMLS の意味タイプ[15] を用いて概念を絞り込むことで，29,915 個のノードと 260,562 個の関係を得た．構築したネットワークを基に，偏頭痛を A-term として与えると，69,972 個の仮説が生成され，その内，マグネシウムやマグネシウム欠乏症を C-term とする仮説は 90 個存在した．つづいて，3.4 節で定義した妥当性尺度に基づき，69,972 個の仮説すべてに対して順位付けを行った．

表 2　妥当・非妥当な仮説の順位（偏頭痛とマグネシウム）

R_{max} (%)	R_{avg} (%)	R_{tfidf} (%)	R_{freq} (%)	
妥当	36.6	47.6	36.3	22.6
非妥当	53.4	54.1	42.0	45.9

偏頭痛とマグネシウムの仮説に関して，各妥当性尺度を用い，妥当および非妥当な仮説を全体の上位何%に順位付けできたかを表 2 に示す．表 2 の妥当な仮説について見ると，R_{freq} が妥当な仮説を最も高く順位付けしたことを除き，前述のレイノー病に関する実験とほぼ同様に，各妥当性尺度は妥当な仮説を順位付けしている．R_{freq} が良い結果を出した理由は，妥当な仮説を説明するイベントの頻度が高かったためである．一方，表 2 の非妥当な仮説について見ると，非妥当な仮説に対しては R_{max} と R_{avg} が比較的良い結果を示している（仮説を低く順位付けしている）ことがわかる．

以上の結果をまとめると，仮説に関連するイベントの数が十分な場合，R_{freq} は妥当な仮説を適切に順位付けることができる．しかし，一般的な語を B-term として持つ仮説に対して，不当に高く順位付けてしまうという欠点があった．一方，意味的類似度に基づく尺度 R_{max} を用いると，イベントの頻度に左右されず，妥当・非妥当どちらの仮説に対しても，比較的適切な順位付けを得られることが確認できた．

5. おわりに

本研究では，意味的に類似したイベントに注目することで，妥当な仮説，特に低頻度の概念またはイベントによって導かれる重要な仮説の同定を行うことを目的とした．そして，MeSH シソーラスに基づく概念間の類似度を拡張してイベント間の類似度を定義し，これを仮説の妥当性として利用した．次に，仮説発見の過去の事例を用いることで検証実験を行い，意味的類似度に基づく妥当性の指標である R_{max} と R_{avg}，頻度に基づく指標である R_{tfidf}，R_{freq} が妥当・非妥当な仮説を適切に順位付けできるかどうかを比較した．その結果，ほとんどの場合において，仮説を説明するイベントの頻度に関わらず R_{max} は安定かつ適切な順位付けを行うことができた．一方，頻度に基づく妥当性は，概念またはイベントの頻度に大きく左右されるため，不適切な結果を示す場合があった．

今後は，イベントと MeSH 語の関連性を考慮したり，MeSH の他に UMLS メタシソーラスや WordNet などを知識資源とすることで，文献のタイトルだけでなくアブストラクトも活用し，取得できる妥当な仮説の被覆率を上げていく必要がある．

参 考 文 献

1) Ananiadou, S., Kell, D.B. and Tsujii, J.: Text mining and its potential applications in systems biology, *Trends in Biotechnology*, Vol.24, No.12, pp.571–579 (2006).
2) Cohen, A.M. and Hersh, W.R.: A survey of current work in biomedical text mining, *Briefings in Bioinformatics*, Vol.6, No.1, pp.57–71 (2005).
3) Jensen, L.J., Saric, J. and Bork, P.: Literature mining for the biologist: from information retrieval to biological discovery, *Nature reviews genetics*, Vol.7, pp.119–129 (2006).
4) Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspectives in Biology and Medicine*, Vol.30, No.1, pp.7–18 (1986).
5) Swanson, D.R.: Migraine and magnesium: eleven neglected connections, *Perspectives in Biology and Medicine*, Vol.31, No.4, pp.526–557 (1988).
6) DiGiacomo, R.A., Kremer, J. and Shah, D.: Fish-oil Dietary Supplementation in Patients with Raynaud's Phenomenon: A Double-Blind, Controlled, Prospective Study, *The American Journal of Medicine*, Vol.86, No.2, pp.158–164 (1989).
7) Gordon, M.D. and Lindsay, R.K.: Toward discovery support systems: a replication, re-examination, and extension of Swanson 's work on literature-based discovery of a connection between Raynaud's and fish oil, *Journal of the American Society for Information Science*, Vol.47, No.2, pp.116–128 (online), (1996).
8) Hristovski, D., Džeroski, S., Peterlin, B. and Rožić-Hristovski, A.: Supporting Discovery in Medicine by Association Rule Mining of Bibliographic Databases, *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, Springer-Verlag, pp.446–451 (2000).
9) Koike, A. and Takagi, T.: Knowledge discovery based on an implicit and explicit conceptual network, *Journal of the American Society for Information Science and Technology*, Vol.58, No.1, pp.51–65 (2007).
10) Kostoff, R.N., Briggs, M.B., Solka, J.L. and Rushenberg, R.L.: Literature-related discovery (LRD): Methodology, *Technological Forecasting and Social Change*, Vol.75, No.2, pp.186–202 (2008).
11) Lindsay, R.K. and Gordon, M.D.: Literature-based discovery by lexical statistics, *Journal of the American Society for Information Science*, Vol.50, No.7, pp.574–587 (online), (1999).
12) Srinivasan, P.: Text mining: generating hypotheses from MEDLINE, *Journal of the American Society for Information Science and Technology*, Vol.55, No.5, pp. 396–413 (online), (2004).
13) Swanson, D.R. and Smalheiser, N.R.: An interactive system for finding complementary literatures: a stimulus to scientific discovery, *Artificial Intelligence*, Vol.91, No.2, pp.183–203 (online), (1997).
14) Wanda, P. and Meliha, Y.-Y.: LitLinker: capturing connections across the biomedical literatur, *Proceedings of the 2nd international conference on Knowledge capture*, pp.105–112 (2003).
15) Weeber, M., Klein, H., deJong-vanden Berg, L. T.W. and Vos, R.: Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries, *Journal of the American Society for Information Science and Technology*, Vol.52, No.7, pp.548–557 (2001).
16) Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T., Aronson, A. R. and Molema, G.: Generating Hypotheses by Discovering Implicit Associations in the Literature: A Case Report of a Search for New Potential Therapeutic Uses for Thalidomide, *Journal of the American Medical Informatics Association*, Vol. 10, No. 3, pp. 252–259 (online), available from ⟨http://www.jamia.org/cgi/content/abstract/10/3/252⟩ (2003).
17) Wren, J.: Extending the mutual information measure to rank inferred literature relationships, *BMC Bioinformatics*, Vol.5, No.1 (2004).
18) Aronson, A.R.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program, *Proceedings of American Medical Informatics 2001 Annual Symposium*, pp.17–21 (2001).
19) SparckJones, K.: Statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–20 (1972).
20) Hersh, W.R., Bhupatiraju, R.T., Ross, L., Roberts, P., Cohen, A.M. and Kraemer, D.F.: Enhancing access to the Bibliome: the TREC 2004 Genomics Track, *Journal of Biomedical Discovery and Collaboration* (2006).
21) NLM: Fact Sheet Medical Subject Headings (2008).
22) Jiang, J.J. and Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *In International Conference Research on Computational Linguistics*, pp.9008+ (1997).
23) Lin, D.: An Information-Theoretic Definition of Similarity, *In Proceedings of the 15th International Conference on Machine Learning*, pp.296–304 (1998).
24) Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy, *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp.448–453 (1995).
25) Seco, N., Veale, T. and Hayes, J.: An Intrinsic Information Content Metric for Semantic Similarity in WordNet, *Proceedings of European Conference on Artificial Intelligence 2004*, pp.1089–1090 (2004).