C-25

イベント類似度にもとづく仮説の順位付け

Hypothesis Generation and Ranking Based on Event Similarities

宮西　大樹 †　　関　和広 ‡　　上原 邦昭 †
Taiki Miyanishi　Kazuhiro Seki　Kuniaki Uehara

1　はじめに

近年，発表される生物医学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し統合することは困難になっている [1, 3]．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 [7, 8]．

文献を基にした仮説生成の初期の研究として，Swanson[7] は，魚油とレイノー病に関係する文献を手作業で調べ上げ，抽出した知識を組み合わせることによって両者の関係を予測した．この関係は，後年，臨床的にも証明されている．その後，Swanson が手動で行った作業を自動で行うことで，大量の文献の中から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グループによって行われている [2, 6, 9] ．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．本研究では，文献から抽出した関係を基に仮説を自動的に生成し，関係間の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，低頻度の概念により導かれた妥当な仮説が提示されにくくなることに対処する．

2　仮説生成

2.1　概念ノットワークの作成

本研究では，最初に生物医学文献から固有表現として生物医学要素（薬品，病気名など）を抽出し，次に生物要素間の関係を語の共起により推定する．そして，共通する生物医学要素を基に推定した関係を結合し，ネットワークを構築する．以下で述べるように，このネットワーク上から，直接的にはつながりはないが間接的にはつながりを持つノードのペアを仮説として抽出する．**図 1** に文献から抽出した生物要素の 2 項関係を組み合わせて構築した概念ネットワークの例を示す．

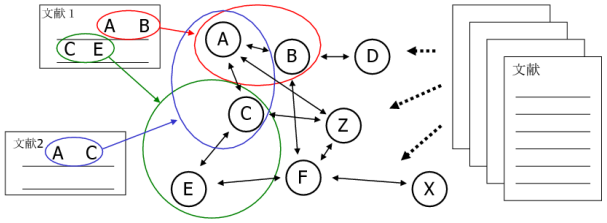

図 1: 文献の情報をもとに構築した概念ネットワーク

2.2　仮説の生成

仮説の生成は，任意の固有表現を始点として，概念ネットワーク上のノード間のエッジ（固有表現間の関係）をたどり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．

2.3　仮説の妥当性

生成された仮説は，実験と検証によるプロセスを経て，正当と認められることで知識となる．我々の目標は，自動的に導出した仮説から妥当な仮説を見つけ出すことである．本稿では，生物要素の 2 項関係をイベントと呼び，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する．我々は，妥当な仮説は類似するイベントから生まれると仮定する．

この仮定の根拠は，類似していないイベントから導出される仮説は妥当性が低いという考えに基づく．なぜなら，そのような仮説には意味的な飛躍があり，その正当性が薄いと考えられるからである．これより，イベント間の類似度が高いほど関連の強い要素間の関係，つまり，妥当性の高い仮説を見つけ易くなると考えられる．妥当性をもとめるために頻度を用いることも考えられるが，頻度を用いると低頻度のイベントを軽視しがちになる．

低頻度のイベントに対応するため，頻度以外の特徴を用いてイベントを表す．イベントを特徴付けるため，階

† 神戸大学大学院工学研究科、 Graduate School of Engineering, Kobe University
‡ 神戸大学自然科学系先端融合研究環、 Organization of Advanced Science and Technology, Kobe University

層的に体系化された MeSH（Medical Subject Headings）を用いる [4]．ただし，MeSH は文献を特徴付ける索引語であり，イベントを特徴付けるものではない．よって，文献の内容を最もよく表す部分からイベントを抽出することでイベントの特徴を文献の特徴で近似する．生物医学文献において文献の内容を最も端的に表す部分は文献のタイトルであると考えられるため，そこからイベントを抽出する．この特徴となる語を基に，語間の類似度を定義し，イベント間の類似度へと拡張する．

2.4 語間の類似度

本提案手法では，イベントに対応する MeSH 語間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．MeSH 語間の類似度測定には，Nuno らが提案したシソーラスの構造を用いた手法 [5] を使用する．

Nuno らは，シソーラス中の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いといった特性を生かし，語に対してシソーラス上の位置に応じた情報量を定義した．この特性を利用することで，上位の概念同士が類似しているよりも，下位の概念が類似しているときの方が類似度は大きくなるといった類似度計算が可能となる．

語の重み付けに頻度を用いれば，データベースに新たな文献が蓄積するたびに語の重みを計算しなければならないが，シソーラスを用いれば，重みを一度求めれば再計算する必要はない．

以下に Nuno らが提案したシソーラス中の語の情報量を表す**式 (1)** と語間の類似度を表す**式 (2)** を示す．

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

$$\mathrm{sim}(m_1, m_2) = \max_{m \in S(m_1,m_2)} \mathrm{ic}(m) \tag{2}$$

ここで，$\mathrm{ic}(m)$ は MeSH 語 m のシソーラス上での情報量であり，$\mathrm{hypo}(m)$ はシソーラス上での MeSH 語 m 以下の下位語の数である．また，$\mathrm{N_s}$ はシソーラス上にある語の数である．**式 (1)** の $\mathrm{sim}(m_1, m_2)$ は語 m_1, m_2 間の類似度を表している．**式 (2)** 中の $c \in S(m_1, m_2)$ は，m_1 と m_2 の共通の祖先であり，かつ両者から最も近い語を表す．

2.5 イベント間の類似度

次に，**式 (2)** で定義した語間の類似度をイベント間の類似度に拡張する．これを基に，類似度による仮説の妥当性を得る．

2.5.1 語の類似度の平均

次の類似度の定義は，イベントに対応する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．このイベントの類似度がイベントによって導出される仮説の妥当性（reasonability）となる．

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

ここで，$\mathbf{M}_i, \mathbf{M}_j$ 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．この MeSH 語の集合はイベントの抽出元である文献に付与された MeSH 語をイベントの抽出元である文献すべてについて，重複を許さずに集めたものである．

2.5.2 最も近い語同士の類似度の平均

次の定義は，２つの MeSH 語の集合を片方側から見て，似ている語同士の類似度だけを算出し，これを双方向に対して行い加算して類似度を求めたものである．この方法をとれば，語間の最小となる類似度だけを足していくので，イベント間の類似する特性だけを考慮することができる．

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

2.5.3 TF・IDF+コサイン類似度

シソーラスを用いた提案手法の比較対象として，テキスト間の類似度を測るためによく使われる TF・IDF とコサイン類似度を用いた手法について定義する．あるイベントには抽出元となる文献が複数対応することがあり，MeSH 語が重複することがある．その MeSH 語の数を語の TF 値とする．また，ある MeSH 語が付与された文献の数をその MeSH 語の DF 値とする．そして，この TF 値と DF 値を用いて単語の TF・IDF 値をもとめる．イベントに対応する各 MeSH 語を TF・IDF 値で重み付けし，単語とその重みを並べたプロファイルを以下に示す．

$$\mathrm{Profile}(e_i) = \{w_{i1}, m_{i1}, w_{i2}, m_{i2}, \cdots, w_{in}, m_{in}\}$$

ここで，$w_{ij} = v_{ij} / \max_k(v_{ik})$ とし，$v_{ij} = n_{ij} \times \log(N/n_j)$ とする．N はデータベース中にある総文書数，n_j は MeSH 語 m_j の文書数であり，これを DF 値とする．n_{ij} はイベント e_i の抽出元となった文献の中で，MeSH 語 m_j を含む文書の数であり，これを TF 値と見なす．

この定義から e_i, e_j に対応するプロファイルを作成し，両者のイベントの類似度を以下のコサイン類似度の**式 (5)**

で求める．この定義は Srinivasan の手法 [6] にならった．

$$\mathrm{R}_{\mathrm{tfidf}}(e_i, e_j) = \frac{\mathrm{Profile}(e_i) \cdot \mathrm{Profile}(e_j)}{|\mathrm{Profile}(e_i)||\mathrm{Profile}(e_j)|} \quad (5)$$

2.5.4 イベントの頻度

上記に示した 3 つの類似度を使った比較として，類似度を用いずイベントの頻度を基に妥当性を定量化する方法を次式で定義する．

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

イベント e_i, e_j の頻度をそれぞれ $freq(e_i), freq(e_j)$ とし，それらの相乗平均をイベント e_i, e_j によって導出される仮説の妥当性とする．

3 仮説生成の評価

3.1 評価手法

生成された仮説の評価のため，Swanson が 1986 年に示した「レイノー病に魚油の摂取が有効である」という仮説を利用する．Swanson は，レイノー病患者に高い血液粘性，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，魚油とレイノー病の関係を予測した．

上述の関係を上位に順位付けすることができるかどうかで仮説生成の良さを評価する．そのために，生物医学文献データベース MEDLINE に収録される 1985 年までの文献を用いて概念ネットワークを構築する．そして，概念ネットワークに対して，魚油を所与として，レイノー病との関係を表す仮説を生成する．これらの仮説を本稿で定義した妥当性を基に順位付けし，頻度を用いた妥当性に対して，類似度を基にした妥当性が正しい仮説を上位に順位付けできるかを検証する．

3.2 実験と考察

3.2.1 実験条件

仮説生成の前段階として，まず概念ネットワークを構築する．そのための実験条件を以下に示す．固有表現抽出に用いた MetaMap は，語と固有表現のマッピングに曖昧性が生じたとき，単一の語に対して複数の固有表現を出力することがある．その場合は，先に出力された表現を語に対応する固有表現とする．さらに，関係抽出の過程で UMLS の意味クラスによる絞り込みを行う際，特定の固有表現の再現率を大きくするため，Bloodviscosity [Laboratory or Test Result] を Blood Viscosity [PhysiologicFunction] に置換する．次に，関係抽出を行う際，特定の意味クラスに属する固有表現だけを用い，その固有表現を含む関係だけを概念ネットワークの作成に用いる．制限に使った UMLS の意味クラスは Biologic Function，Cell Function，Disease or Syndrome，Lipid，Molecular Function，Organ or Tissue Function，Organism Function，Pathologic Function，Physiologic Function の計 9 個である．

表 1: 魚油を所与として生成された仮説

ID	A-term	B-term	C-term
1	Fish Oil	Blood Viscosity	Primary Raynaud's
2	Fish Oil	Blood Viscosity	Raynaud Disease
3	Fish Oil	Blood Viscosity	Paroxysmal digital cyanosis
4	Fish Oil	Peripheral vascular disease	Paroxysmal digital cyanosis
5	Fish Oil	Atheromatosis	Raynaud Disease
6	Fish Oil	Suppression	Paroxysmal digital cyanosis
7	Fish Oil	Peripheral vascular disease	Raynaud Disease
8	Fish Oil	Development	Paroxysmal digital cyanosis

3.2.2 実験結果と考察

上記の条件で作成した概念ネットワークを用いて，魚油を始点として仮説生成を行ったところ，合計で 13677 個の仮説が得られた．その内，レイノー病を終点とする仮説が 8 つ確認された．**表 1** に，生成した魚油とレイノー病の関係を A-term，B-term，C-term を用いて掲載する．ID の順番は，類似度に基づいた妥当性による順位の平均を昇順に並べたものである．

次に前章で定義した複数の仮説の妥当性に基づいて仮説の順位付けを行い，定義の違いによって仮説の順位がどのように変化するかを実験結果により示す．**図 2** は，各妥当性による順位と類似度に基づく妥当性による順位の平均を図示したものである．

図 2 から，魚油とレイノー病の仮説に関しては，類似度による順位付けの仕方はどれも似たような順位付けとなっていることが分かる．一方，類似度に基づく順位付けと頻度による順位付けでは順位がかなり違っている．

次に，順位付けした個々の仮説のについて考察する．ID 番号 1，2，3 に対応する仮説は，Blood Viscosity（血液粘性）を B-term としたものである．これは Swanson が人手で調べ上げた関係と一致するので，これらの仮説は妥当な仮説であるといえる．**図 2** より，類似度による順位付けは頻度による順位付けより，これらの関係を比較的上位に順位付けできている．頻度による順位付けが有効に働かなかった理由としては，文献から抽出できた魚油と血液粘性の関係の数（１～４個）が極端に少なかったためと考えられる．類似度を用いた手法では，これら低頻度の概念に基づく仮説を上位に順位付けすることができている．これより，類似度を用いれば低頻度の概念にも対処できることが分かる．

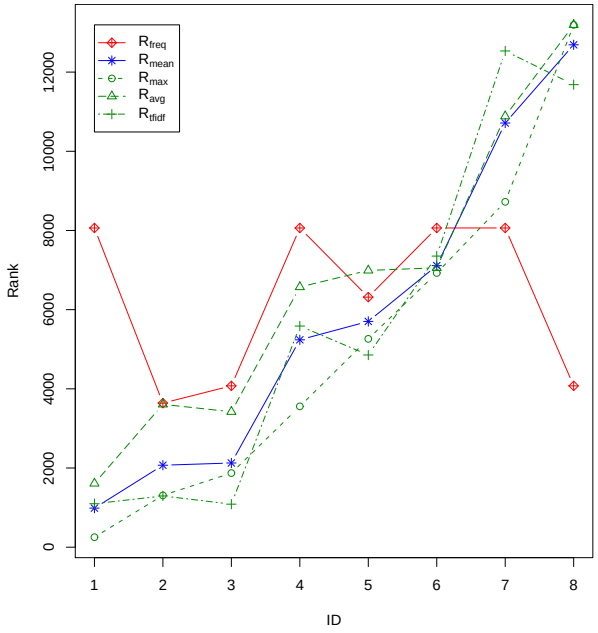

図 2: 妥当性に基づく仮説の順位

同様に，類似度の違いによる順位付けにおいても差異が見られる．文献に付与されたコンセプト間の類似度の内，すべてのコンセプト間の類似度の平均をとった R_{avg} と最も類似したコンセプトだけを比べた R_{max} とでは，後者の方が妥当な仮説を上位に順位付けできている．これは，類似度を用いて妥当な仮説を導く際には，類似した要素だけに注目することが効果的である可能性を示している．

さらに，ID1 の仮説について，TF・IDF+コサイン類似度による R_{tfidf} とシソーラスによる R_{max} および R_{avg} とでは，順位の違いが生じている．これは，コサイン類似度は同じ語彙間の値の積しか類似度に加算しない一方で，シソーラスを用いた類似度の算出方法は似た語彙同士の類似度も考慮することができることによる．これにより，妥当な仮説を上位に順位付けることができた．

ID 番号が 4 以下の仮説は下位に順位付けされている．このうち ID4，5，7 の仮説は妥当な仮説であるにも関わらず下位に順位付けされてしまっている．この原因としては，使用した文献に付与された MeSH 語が他のものと比べて少なかったり，アブストラクトが存在しなかったりし，類似度が有効に測れなかったためだと考えられる．また，文献のタイトルから抽出したイベントが MeSH 語および文献の内容と離れていたことも原因の一つであると考えられる．

ID6 と 8 の仮説は B-term があまりに一般的な語彙であるため，これを仮説の説明するものだと考えると，仮説としてあまり意味を持たない．一般的な語彙を含む仮説は，頻度を用いた順位付けだと不当に上位に順位付けされることがある．しかし，今回のように TF・IDF やシソーラスの情報量を用いて類似度を測ることで，頻出する生物要素を説明として持つ仮説を下位に順位付けることができている．

4 まとめ

本論文では，仮説の導出に必要となるイベント間の類似度を用いて妥当な仮説，特に低頻度の関係をもちいた仮説の上位順位付けを目指した．文献に付与された意味情報である MeSH 語を用いて，文献から生成される新たな知識（仮説）に関係間の類似度を基にした妥当性を定義し，この仮説の妥当性によって順位付けを行った．その結果，頻度を用いた仮説の順位付けに比べて，頻出する生物要素を含む仮説を下位に順位付けすることができ，また，低頻度の関係から導かれた妥当な仮説を上位に順位付けることができた．これにより，低頻度の概念に対応する手法を示すことができた．

参考文献

[1] Sophia Ananiadou, Douglas B. Kell, and Jun ichi Tsujii. Text mining and its potential applications in systems biology. *Trends in Biotechnology*, Vol. 24, No. 12, pp. 571–579, 2006.

[2] Michael D. Gordon and Robert K. Lindsay. 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, 1996.

[3] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. Literature mining for the biologist: from information retriev al to biological discovery. *Nature reviews genetics*, Vol. 7, pp. 119–129, 2006.

[4] NLM. Fact sheet medical subject headings, 2008.

[5] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of European Conference on Artificial Intelligence 2004*, pp. 1089–1090, 2004.

[6] Padmini Srinivasan. Text mining: generating hypotheses from medline. *J. Am. Soc. Inf. Sci. Technol.*, Vol. 55, No. 5, pp. 396–413, 2004.

[7] Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, Vol. 30, No. 1, pp. 7–18, 1986.

[8] Don R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, Vol. 31, No. 4, pp. 526–557, 1988.

[9] Marc Weeber, Henry Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos. 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.