1D1-3

時系列テキストマイニングによる類似用語の語彙体系内距離の比較

A Comarison of Similarity of Technical Terms in Temporal Patterns on MeSH

阿部 秀尚[*1]　　津本 周作[*1]
Hidenao ABE　　Shusaku TSUMOTO

[*1]島根大学医学部医学科医療情報学講座
Dept. of Medical Informatics, Shimane University, School of Medicine

In this paper, we present an analysis of a relationship between temporal trends of medical terms in medical research document and their similarities on a structured vocabulary. In order to obtain the temporal trends, we used our temporal pattern extraction method that combines an automatic term extraction, an importance index of the terms, and clustering for the values in each period. By using a set of medical research documents that were published every year, we extracted temporal patterns of the automatically extracted terms. Then, we assigned trends of the patterns based on the linear regression technique. Finally, the similarity measure on the medical taxonomy by defining a distance on the tree structure is compared between the two trends of the temporal patterns.

1. はじめに

近年，情報システムに蓄積される多くのデータの増加により，電子的に保存される文書も顕著に増加している．これらの文書は，検索技術と併せ，定性情報の共有や普及に貢献することが期待される．一方，学術論文を対象とした Web 上の文献データベースは，情報科学のみならず多くの分野で公開され，生命科学・医学分野でもその研究発展に大きな役割を果たしている．しかしながら，このような文献から有用な知識を見出すことは，人手による読解によるところが多く，これを支援する手法が求められている．

用語やその用法は，用語自体の出現頻度や構成要素の単語と用語の出現頻度の差異として計量化され[*1]，有用な語句の抽出を行うテキストマイニングで用いられる．また，用語に関する計量化指標の変化では，タグクラウドをはじめとする用語の出現頻度の変化が注目され，傾向抽出手法 [Kontostathis 03] として開発されてきた．しかしながら，従来の傾向抽出手法では，辞書に現れる単語について特定の計量化指標の傾向を予め与えたモデルとして得ることが前提とされ，辞書に無い語句や複数の計量化指標の考慮，複数の傾向の同時抽出や傾向の比較が困難であった．

このため，我々は，辞書によらない用語自動抽出手法から得られた用語に対し，時間毎の重要度指標を算出し，各用語の時系列データの類似性から時系列パターンを得る手法を開発してきた [阿部 10]．本手法によって，時間経過に伴って変化する用法の類似した用語群を得られることが確認されたが，各時系列パターンが示す傾向と含まれる語句の意味上の類似性は定性的に評価したのみであった [Abe 10]．このため，各時系列パターンの傾向と各パターンに含まれる用語の意味上の類似性を定量的に評価することが必要と考えた．

本稿では，計量化指標の変化に基づく時系列パターン抽出手法によって得られた類似語群の語彙体系上での類似度について，各時系列パターンの直線的傾向を基に比較を行う．これにより，意味上の類似度と時系列における振る舞いの傾向の関係を明らかにする．

連絡先: 阿部秀尚，島根大学医学部医学科医療情報学講座，〒693-8501 島根県出雲市塩冶町 89-1，電話番号 (0853)20-2174，FAX 番号 (0853)20-2170，abe@med.shimane-u.ac.jp

*1 これらを総称して計量化指標と呼ぶ．

2. 医学用語シソーラス MeSH

MeSH(Medical Subject Headings)[MeS] は，米国医学図書館 NLM によって公開されている生命科学および医学分野の用語を網羅する用語集である．MeSH は，定義語 (Descriptor)，用語 (Entry Term)，定義記述などから成り，定義語は階層構造上の単一あるいは複数の節点や葉の番号 (TreeNumber) が割り当てられている．2011 年現在，MeSH に掲載された定義語は 25,588，用語は 464,282 語である．

MeSH の階層構造の最上位は，カテゴリと呼ばれ，16 のカテゴリから構成されている．特に，解剖 (A)，生物 (B)，疾患 (C)，化学物質・薬物 (D) など医学に関連が強いカテゴリから，情報科学 (L) に関する用語まで広い範囲を網羅することが特徴である．MeSH の階層構造では，複数の定義語が 1 つの節点や葉，あるいは 1 つの定義語が複数の節点や葉の番号をもち，類義語辞書としての性格が強いため，厳密な概念階層とはなっていない．MeSH の階層構造の一部を図 1 に示す．

3. 用語計量化指標の時系列パターンに基づく類似用語群の抽出手法

辞書に掲載されていない新規の概念や用語について，用語の用法に関係する計量化指標から，様々な傾向を同時に時系列パターンとして抽出する手法は，以下の 4 フェーズから成る．

1. 辞書に依らない文書群からの語句の抽出
2. 語句の重要度指標
3. 時系列パターンの生成
4. 時系列パターンの傾向割り当てによる用法の傾向分析

本手法では，まず，全時点の文書群あるいは一部の文書群を対象として，用語を抽出する．次に，各用語について，時点毎

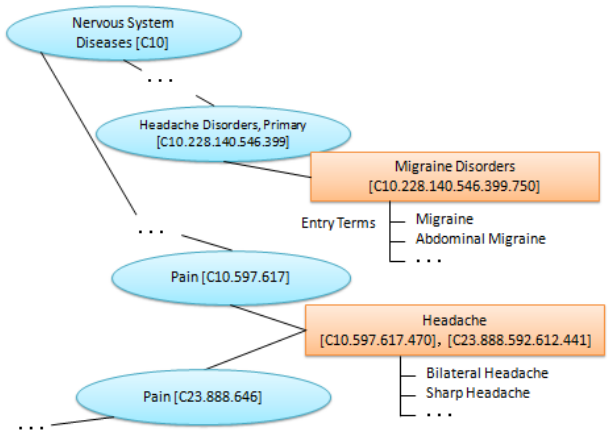

図 1: MeSH における概念語と用語の関係と階層構造の例.

の文書群における計量化指標の値を算出し，各用語を行，各時点の計量化指標の値を列とするデータセットを作成する．用語 $term_i$ のある一定期間の文書 D_{period} 中の出現頻度としては，用語自体の出現頻度 $TF(term_i, D_{period})$ や用語を含む文書数 $DF(term_i, D_{period})$ がある．本稿においては，次の指標を計量化指標として扱う．

テキストマイニングにおいて単語（あるいはフレーズ）の重要度として広く用いられる tf-idf[Sparck Jones 88] は，各用語 $term$ のある一定期間での文書 D_{period} に対する tf-idf 値 $TFIDF(term_i, D_{period})$ として以下のように計算される．

$$TFIDF\ (\ term_i, D_{period}) = TF\ (\ term_i, D_{period}) \times log_e \frac{|D|}{DF(term_i, D_{period})}$$

ここで，$TF(term)$ は，サイズ $|D_{period}|$ の文書における $term_i$ の出現頻度を表し，$DF(term)$ は $term$ を含む文書数を表している．

生成されたデータセットに対し，各用語の時系列について類似度を算出し，クラスタリングによって時系列パターン c_k を生成する．生成された時系列パターンについて，時間方向の傾向を解釈する．そこで，各時系列パターン c_k の傾き $Deg(c_k)$ は，代表元の値 y_i について，時間経過 $x_1, ..., x_n$ に対して，以下のように算出する．このとき，n は時点数を表す．

$$Deg(c_k) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

この傾きとそれぞれの平均 $\bar{y}, \bar{x}$ を用いて $Int(c_k)$ は，以下のように算出される．

$$Int(c_k) = \bar{y} - Deg(c_k)\bar{x}$$

4. 時間毎に出版された医学論文での類似用語群抽出と評価

3. 節で述べた計量化指標の時系列パターンに基づく類似用語群抽出手法を用いて，医学論文中の用語の時系列パターンを得る．時間毎に出版される文書群として，時間経過に伴って薬物による治療法が変化していることが指摘される片頭痛に関し [Abe 10]，PubMed への検索クエリを用いて 1980 年から 2009 年までに出版された論文のタイトルを得る．ここで用いる検索クエリは，”migraine/drug therapy [MH:NOEXP] AND YYYY [DP] AND clinical trial [PT] AND english [LA]” であり，YYYY の部分はそれぞれの検索対象の年を示す 4 桁の数字とした．

本検索シナリオから得られた論文のタイトル数を図 2 に示す．この期間で検索クエリによって得られた論文タイトルの総数は 3,678 である．

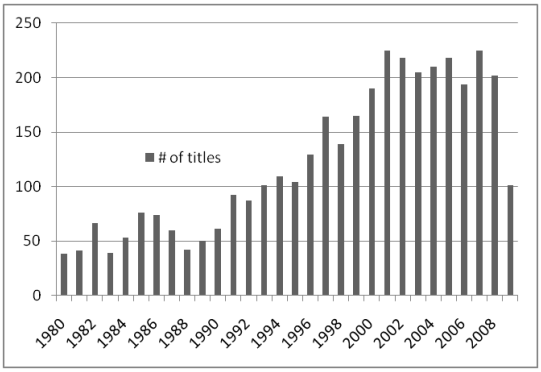

図 2: 片頭痛の薬物治療に関して 1980 年から 2009 年までの 30 年間で MEDLINE に掲載された論文のタイトル数.

これら，全ての文書群に対し，FLR スコアに基づく自動抽出手法 [Nakagawa 00][*2]によって用語を自動抽出した結果，1,469 語が得られた．この 1,469 語の用語について，30 年間の各年に出版された文書群で tf-idf に基づく指標値を計算し，データセットを得た．

以上のデータセットについて，ユークリッド距離に基づく k-means によるクラスタリングを適用し，用語の計量化指標の時系列パターンを得る．2 つの用語の時点毎の計量化指標による時系列データ $\mathbf{x}$ と $\mathbf{y}$ について，以下のように定義されるユークリッド距離尺度を用いた．

$$Sim(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

このとき，生成パターン数 k は，全用語数の 1%である $k = 14$ とした．実装は，Weka[Witten 00] によるものを用い，500 回のクラスタ割り当ての更新中に空のクラスタの発生を許容するものとした．この結果得られた時系列パターンおよび代表用語を表 1 に示す．

次に，MeSH 上の類似度を計測するため，以下のように同一クラスタ c_k に属する用語 $term_{i1}$ と $term_{i2}$ の間で類似度を計算する．例えば，図 1 に示した migraine (Migraine Disorders [C10.228.140.546.399.750]) と sharp headache (Headache [C10.597.617.470], [C23.888.592.612.441]) の間の距離 $Dist(migraine, headache)$ は 8 または 9 となる．MeSH 上の位置 $TNum(term_i)$ は，用語 $term_i$ に MeSH 用語が含まれるときに複数候補が得られるため，$Dist(term_{i1}, term_{i2}))(i_1 \neq i_2)$ は 2 用語間での最小値を用いることとする．ただし，別カテゴリでの距離の計算は行わず，

*2 公開され TermExtract モジュール（http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html にて配布）の実装のうち，ストップワード除去による方法を適用した．

表 1: 片頭痛の検索シナリオで得られた用語の時系列パターンと直線的傾向.

k	代表語	$Deg(c_k)$	$Int(c_k)$
1	migraine patients	0.059	3.874
2	clinical efficacy	0.063	-0.142
3	placebo-controlled study	0.069	-0.299
4	cluster headache	0.050	0.203
5	5-ht1b/1d agonists	0.044	-0.139
6	migraine	1.816	3.589
7	double-blind study	0.058	0.206
8	migraine therapy	0.339	-1.870
9	patients	0.748	1.363
10	management of migraine	0.025	0.816
11	tension-type headache	0.058	0.196
12	oral sumatriptan	0.255	0.955
13	migraine prophylaxis	0.025	0.523
14	acute treatment of migraine	0.789	-3.5661

同じカテゴリの階層構造内のみで距離を計算する．次に，クラスタ c_k 内の平均類似度 $MSim(c_k)$ を次のように定義する．

$$sumS(c_k) = \sum_{i1 \neq i2} \frac{1}{1 + Dist(term_{i1}, term_{i2})}$$

$$MSim(c_k) = \frac{sumS(c_k)}{|mat(c_k)|}$$

このとき，$|mat(c_k)|$ はクラスタ内の用語 $term_{i1}$ と $term_{i2}$ について，それぞれの TreeNumber リスト $TNum(term_{i1})$ と $TNum(term_{i2})$ が取得され，距離 $Dist(term_{i1}, term_{i2})(i_1 \neq i_2)$ が計算できた用語の組み合わせ数を表す．

以上による MeSH の階層構造上における平均類似度を表 2 に示す.

表 2: 片頭痛の検索シナリオで得られた用語の時系列パターンの傾向と MeSH 階層構造上での平均類似度

k	傾向の解釈	$MSim(c_k)$
1	定常・頻用	0.133
2	新興	0.138
3	新興	0.133
4	定常・頻用	0.131
5	新興	0.135
6	定常・頻用	0.160
7	定常・頻用	0.135
8	新興	0.120
9	定常・頻用	0.174
10	定常・頻用	0.144
11	定常・頻用	0.141
12	定常．頻用	0.131
13	定常・頻用	0.133
14	新興	0.121

表 2 に示した結果について，時系列パターンの直線的傾向について切片が負のものを**新興**，それ以外を**頻用・定常**として，MeSH 階層構造上の平均類似度を比較する．平均類似度について各傾向を群として比較すると，**新興**では中央値が 0.132，**頻用・定常**では中央値は 0.138 となり，頻用・定常群のほうが若干近接した用語を用いている結果となった．以上から，新興の傾向をもつ用語群は，比較的語彙体系上で遠い位置に定義される語句を含みながら，それらの使用頻度が増加する傾向であることが明らかとなった．

図 3 に c_{14} にパターン番号 $k = 14$ に含まれる用語と時系列パターンの代表元の値の変化を示す．このパターンは，22 語中 10 語がトリプタンに関する薬剤名から成るが，代表語が示すように患者の急性治療に関する用語も含まれている．このため，より正確に MeSH 階層構造での距離を測定し，密度の差異に関する指標などを用いて，時系列パターン内での用語の属するカテゴリの比較を可能とすることが必要と考えられる．

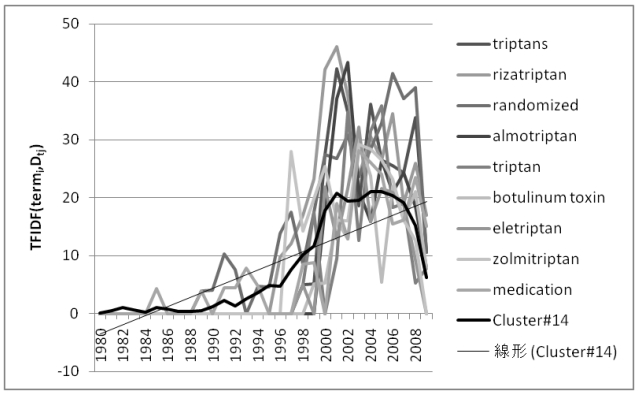

図 3: 片頭痛の薬物治療に関する論文タイトルにおいて,「新興」となったパターンと含まれる用語（上位 10 用語）.

5. おわりに

本稿では，用語の計量化指標の変化に基づく時系列パターン抽出手法によって得られた類似語群の語彙体系上での類似度を各時系列パターンの直線的傾向を表す指標および密度指標と比較について述べた．この結果，類似用語から成る時系列パターンの時間経過に対する傾向と意味上の類似度指標の間では，注目を集めている用語を含むパターン内での値が低くなった．これは，新興の傾向を示すパターンに含まれる用語がより広範囲の概念を含め，新たな研究の方向性の模索することと合致する．しかしながら，より正確に MeSH 用語と自動抽出による用語の間で概念定義上での距離を得るため，自動抽出された用語と MeSH 上の用語との意味上の曖昧性を事前に解消する必要があると考える．

今後は，MeSH に定義された用語あるいは，定義語の解説文からの用語を用い，時間毎に出版された論文での用語の用いられ方のパターンを取得し，これらの時系列パターンと現在の語彙体系上でのより正確な近接度を明らかとすることが課題である．また，新規に追加される用語について，これと関連する文書での用語用法の変化パターンと新規用語の追加位置を数値予測手法により予測する方法の開発を行っていく．

参考文献

[Abe 10] Abe, H. and Tsumoto, S.: Finding Temporal Patterns of Medical Terms in MEDLINE Documents, in *2010*

Intelligent Data Analysis in Biomedicine and Pharmacology (IDAMAP2010), pp. 73–77 (2010)

[Kontostathis 03] Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S., and Phelps, D. J.: A Survey of Emerging Trend Detection in Textual Data Mining, *A Comprehensive Survey of Text Mining* (2003)

[MeS] MeSH (Medical Subject Headings), `http://www.ncbi.nlm.nih.gov/mesh`

[Nakagawa 00] Nakagawa, H.: "Automatic Term Recognition based on Statistics of Compound Nouns", *Terminology*, Vol. 6, No. 2, pp. 195–210 (2000)

[Sparck Jones 88] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval, *Document retrieval systems*, pp. 132–142 (1988)

[Witten 00] Witten, I. H. and Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann (2000)

[阿部 10] 阿部 秀尚, 津本 周作：専門用語の用法に関する計量化指標の時系列パタン分析, 2010 年度人工知能学会全国大会 (第 24 回) (2010)