ISS-P-41

一般学生を対象とした医学情報検索支援の検討

森　武士†　　出口　幸子††
† 近畿大学大学院システム工学研究科　　†† 近畿大学工学部電子情報工学科

1. はじめに

近年，テキストマイニングを応用して，医学系の文献から診断や治療に有益な情報を抽出する研究が行われている[1]．一方，ネットワーク上には多くの医学系の情報があり，その真偽が問題となっている．本研究では，医学を専門としない人が，医学系の1次情報を検索するための支援システムを検討した．そのために，ネットワーク上の英語の医学系文献サイトPubMedとニュースサイトyahoo.comにおいて情報を収集した．

2.　医学系文献サイトにおける情報収集

医学系文献サイトにおいては，身近な物質から始めて病気の原因を探ることを目指し，次に検索するべきキーワードを見つけるための支援を行う．

2.1 身近な物質に関わる病気の探索

身近な物質はVitamin Dとし，PubMedで"Vitamin D disease"と"Vitamin D prevention"を検索し，各200件分の論文概要を取得した(新着順に取得．約300KB)．2つの検索結果をマッチングし，さらに病名DB（CTD）とマッチングし，共通単語を得た．この中から特定の病気としてAlzheimer's Diseaseを決め，また他の病気としてdementia, atherosclerosis, diabetesを選んだ．

2.2 複数の病気に共通する要因の探索

前述の検索で得られた4つの病名とVitamin Dを組合せ，再度PubMedで検索し，各200件分の論文概要を取得した．4つの検索結果を互いにマッチングすることにより，共通語を探した．また，医学用語DB（MeSH）ともマッチングを行い，不要語を除去した．この結果から共通要因と考えられるinflammation, immunity等を見つけた．

2.3 特定の病気の原因の探索

先に見つけた共通要因と考えられる単語と，特定の病気を組み合わせることで，特定の病気の原因となりうる単語の抽出を試みた．PubMedで"alzheimer's inflammation"の検索を行い，200件の概要部分を得た．このデータを医学用語DBとマッチングし，品詞情報により絞り込み，出現順に並べたデータにした．分析は，(a)概要200件の上位100単語の集計，(b)KeyGraph[2]，(c)ネットワーク分析，(d)クラスター分析，および(e)提案手法について行った．KeyGraphは頻度の上位80単語およびそれらと共起(1文において同時に出現)する上位20単語を使用した．ネットワーク分析では，2-gramの上位100組を使用した．クラスター分析では，頻度の上位100単語について，各論文における出現回数を記録した200次元の特徴ベクトルを作成し，主成分分析で100次元にして使用した．

提案手法は，複数の病気の共通要因と考えられる単語(ここではinflammation)が出現した場所に注目し，inflammationを含む文＋後ろ2文＋前1文(計4文)における単語の集計を行う．その後，頻度の上位80単語およびそれらと共起(4文において同時に出現)する上位の組を使用し，単語の関連付けを行った．

今回の提案手法では他の分析方法と比較して優位な結果は得られなかったが，重要と考えられる単語を示すことができた．

3. ニュースサイトにおける情報収集

ニュースサイトについては，情報の分析で新しい知見が得られた場合，医学系文献サイトで1次情報を検索して分析し，真偽を判定することを目指している．

yahoo.comのニュースサイトから単語を検索し，検索結果にある記事のリンク先を巡回して情報を収集する．毎日，定時に自動実行し，フォルダを自動生成し，検索結果のファイルを保存している．

病気の原因の探索例として，"Alzheimer's factor"を検索し，情報を収集した．分析方法は2.3節の(a)(b)(c)と同じである．今回，分析対象としたデータが2週間分のニュース記事とそのリンク先の情報(約250KB)であったため，新しい知見は得られなかった．また，PubMedにおける"Alzheimer's factor"の検索結果と比較したところ，原因と考えられる単語はニュースの方が少ないことが確認された．

4. おわりに

本研究では，医学系の1次情報の検索を支援するため，情報収集の枠組みを検討し，種々のツールを作成し，また分析の手法を提案して既存の分析方法と比較した．今後の課題としては，提案手法の評価・改良，作成したツールの学習支援システムへの組込み，ニュース記事の蓄積と検索結果の評価等があげられる．

謝辞　プログラムとデータの作成にご協力頂いた近畿大学工学部の藤原氏・森脇氏・梶原氏に感謝致します．

参考文献

[1]　Nagarajan, M. et al.: Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature, Proc. of the 21th ACM SIGKDD, pp.2019-2028, 2015.

[2]　大澤，ベンソン，谷内田: KeyGraph:語の共起グラフの分割・統合によるキーワード抽出，信学会論文誌，J82-D-I(2)，pp.391-400，1999.