1E1-01

医療分野における Web 文書からの話題抽出方法

Extraction of Topics from Web Documents in the Medical Domain

長沼 潔[*1]
Kiyoshi Naganuma

速水 悟[*2]
Satoru Hayamizu

[*1] 岐阜大学大学院工学研究科
Graduate School of Engineering, Gifu University

[*2] 岐阜大学工学部
Faculty of Engineering, Gifu University

This paper describes Web documents filtering of specific topics in the medical domain. Recently, huge information has been accumulated on Web. Moreover, it is useful because it includes latest information. We aim to produce statistical information from these Web documents. We collected Web documents by using the search engine. We divided the document using tag code. Next, we have extracted an important word by using TF*IDF. And, we clustered by using the result between documents, and merged the documents with high level similarity. We show the experimental results using some words in medical domain.

1. はじめに

近年, インターネットの利用人口の増加により, Web 上には膨大な情報が蓄積されている. また, 常に最新の情報を提供しているという点からも, これらの情報は有用であると言える. しかし, Web 文書の 1 ページ中には, 様々な分野の記事が掲載されていることが多く, 利用者にとって不必要な情報が含まれていることがある.

また, ある特定の話題のみを抽出することにより, その話題に特化した文書のみを収集することが可能となり, 読み手にとって必要な文書のみを取得することができるので, 必要性は高い. 一方, 医療において病気に関連した情報は, 知りたい利用者が多いと考えられ, 必要性は高いと言える.

そこで本研究では, Web 文書から医療分野の特定の話題部分のみを抽出し, 収集した文書群から医療に関する統計的情報を抽出するための素材を作成することにする.

2. システムの構成

本節では, 本研究で構築するシステムの概要を示す.

まず, 本研究で提案するシステムの流れを図 1 に示す. このように本システムは, 検索, 項目分類, 候補群選択のモジュールから構成されている.

はじめに入力用語 (病名) に対して、検索エンジンを使用して Web から必要な文書を収集する. 次に, 取得した Web 文書をあらかじめ指定しておいたタグコードを分割箇所として Web 文書を意味の通る小さなブロック(passage)ごとに分割する処理を行う[山本 05]. そして, passage ごとに分割した文書群に対して, 知りたい項目ごとに文書を分類する. 次に, 分類した passage 文書に対して形態素解析をし, その passage 文書の重要語を抽出する. そして, その結果を利用して文書間のクラスタリングをし, 類似度の高い文書をまとめる. 最後に, 頻度の高い内容の文書を出力とする.

本研究では, 入力用語 (病名) と知りたい項目については, 以下のように限定して行うことにする.

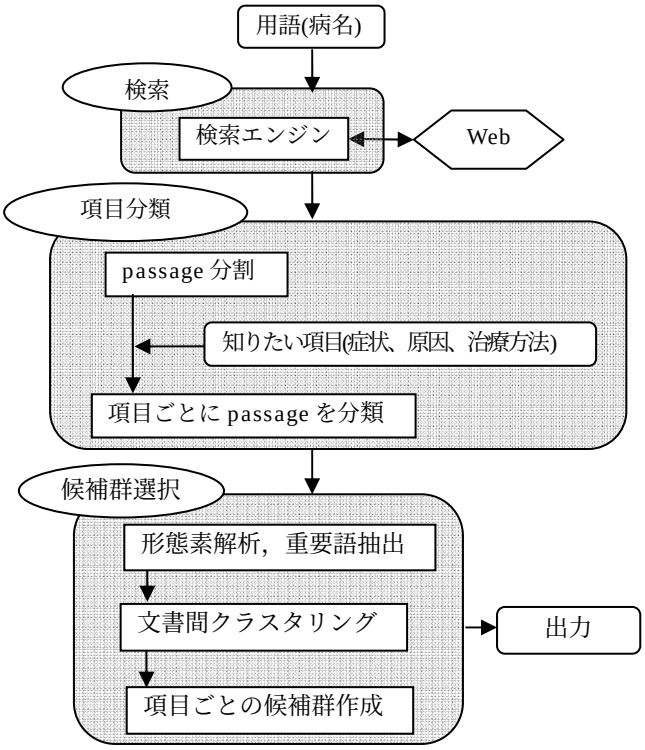

図 1:本システムの構成

- **入力用語 (病名)**:脳梗塞
- **知りたい項目**:症状, 原因, 治療方法

以下に3つのモジュールごとについての説明を記す.

2.1 Web 巡回によるテキストの収集

本研究での Web 巡回では, 検索エンジンとして Google[*1] を用いる. まず, 入力用語を検索エンジンにかけ, 上位 200 件のアドレスを取得する. そして, 入力用語 (病名) を含む Web 文書の HTML テキストを集める. このとき, 取得したリンク先には, リンク先が不明の Web 文書が存在する場合がある. このような場合, タイムアウトをする処理を行う. 取得したアドレスの HTML テキストを収集した後, 収集した HTML テキストから入力用語を含

連絡先:長沼潔, 岐阜大学大学院 工学研究科 応用情報学専攻, 〒501-1193 岐阜市柳戸 1-1,
E-mail:kiyoshi@hym.info.gifu-u.ac.jp

[*1] Google: http://www.google.co.jp

むリンク先を巡って HTML テキストを収集し, 200 件の HTML テキストが得られるまで収集するようにする.

また, 何度も同じリンク先を巡るとテキストに偏りが出るので, それを回避するために, 巡回の重複を避けるようにする.

2.2 項目分類

収集した Web 文書群の中には, 複数の話題が含まれていることがある. そのため, 話題ごとに Web 文書を分割する必要がある. そこで, Web 文書を小さなブロック(passage)に分割する処理を行う. そして, passage ごとに分割をした文書群を知りたい項目ごとに分割する.

(1) passage 分割

Web 文書を話題ごとに分割する方法の一つとして, HTML のレイアウトを利用する方法がある[藤井 02]. これは, Web 文書を高度に解析することなく passage ごとに分割することが可能にする. この手法は, Web 文書はあるタグコードにより意味的にまとまったブロックに分割されているという考えから, タグコードを分割箇所として Web 文書を分類する. 表 1 に分割箇所とするタグコードを示す.

表 1: passage の分割箇所とするタグコード

タグ	意味	タグ	意味
HTML	HTML 文書	TABLE	表作成
P	段落	DIR	リスト
TITLE	タイトル	OL	順序ありリスト
H	見出し	UL	順序なしリスト
DIV	ブロック要素	DL	定義リスト
PRE, PLAINTEXT	ソースの表示	MENU	メニューリスト
XMP, LISTING	ソースの表示	BLOCKQUOTE	引用文

(2) passage 分類

表 1 のタグコードによって passage 分割をした文書群に対して, 分類を行う. 本研究では, 知りたい項目として「症状, 原因, 治療方法」の 3 つの項目に分類することにする. しかし, 例えば症状について表記してある文書では, 「症状」という単語以外で症状の内容について表している場合が考えられる. そこで, 3 つの項目のそれぞれについて, その項目の内容の文書を導き出すための単語リストを作成することにする. その単語リストの例を表 2 に示す.

表 2: 各項目における項目識別単語の例

項目	単語リスト
症状	症状, 病状, 痛み, 前兆, 主訴, etc
原因	原因, 要因, 疾患, 遺伝, 先天性, etc
治療方法	治療, 処方, 投薬, 服用, 手術, etc

ある項目の単語リストに含まれているいずれか一つの単語と入力用語が含まれている passage 文書は, その項目についての内容が表記してあると考えて分類を行う.

2.3 候補群選択

passage 文書を項目ごとに分類した後, その文書群の中には類似する文書があると考えられる. そこで, 各文書の重要語を抽出し, 文書間の類似度を求めて文書間クラスタリングを行う. そして, 多くの類似する文書が存在する場合, その文書は信憑性が高いと判断できるので, 上位にランク付けすることができる.

(1) 形態素解析

passage 文書の重要語を抽出するにあたり, まず形態素解析を行い品詞情報を取得することにより, 重要語となる単語候補を抜き出す必要がある. ここで抜き出す品詞は, 名詞と未知語とする. また, passage 文書を形態素解析するために, ChaSen[松本 03]を使用する.

(2) 重要語抽出

形態素解析の結果から, 各 passage 文書における重要語を求める. 本研究では, 重要語抽出を行うのに TF*IDF 法を使用する. ある passage 文書 m における語句 t の重要度 $w(t,m)$は, 式(1)のようにして求める.

$$w(t,m) = \frac{1}{M}\left(tf(t,m)\cdot\left(\log\frac{N}{df(t)}+1\right)\right) \quad (1)$$

ここで, $tf(t,m)$は語句 t の文書 m 内の出現頻度(回数), M は文書 m における総単語数, N は総文数, $df(t)$は語句 t を含む文数を示している. このようにして求めた TF*IDF 値の上位 10 語をその passage 文書における重要語とする.

(3) 文書間クラスタリング

各 passage 文書の類似した文書を求めるために, 各二文書間の類似度を求めて文書間クラスタリングを行うことにする[余 03]. 本研究では, 二つの文書 m_j, m_k の類似度 $sim(m_j, m_k)$ は式(2)で求める.

$$sim(m_j, m_k) = \frac{\sum_{i=1}^{n} w(t_i, m_j)w(t_i, m_k)}{\sqrt{\sum_{i=1}^{n}(w(t_i, m_j))^2 \sum_{i=1}^{n}(w(t_i, m_k))^2}} \quad (2)$$

ここで, $w(t_i, m_j)$と $w(t_i, m_k)$は、それぞれ語句 t_i の文書における重要度である. このようにして類似度を求め, 類似度が高い二つの文書を統合し, 類似している文書の頻度を調べる. そして, 類似頻度が高い文書は信憑性が高いという考えから, その文書群は上位にくるようにランク付けを行う. このようにして, 各項目に対して候補の文書を求める.

3. 評価実験

3.1 実験対象と方法

本論文で行った各項目の候補群の選択についての性能を調べるために評価実験を行った. 実験の対象は, 入力用語を「脳梗塞」, 知りたい項目を「症状, 原因, 治療方法」の三項目とする. そして, その各項目で求めた文書群において, 上位 50 件における精度について調べた. ここで言う精度とは, 上位 50 件についてその項目のことを述べている文書がいくつあるかということである. また, この正解かどうかの判定は人手で行った.

3.2 実験結果と考察

実験の結果を表 3 に示す. そして, 各項目の抽出した文書の例を表 4 に示す. また、失敗例を表 5 に示す.

表 3：実験結果

	正解数	精度
症状	31	0.62
原因	35	0.70
治療方法	27	0.54

表 4:各項目の抽出した文書の例

症状	当初から脳硬塞のため記憶が部分的に途切れる状態が出ていらっしゃるそうです。ご主人のNさんが倒れた1998年8月7日の朝‥吐き気とめまいがあったそうです。血圧は200を越えていらして、救急車で病院へ運ばれました。病院で容態が急変されました。糖尿病から来る脳硬塞と診断されました。
原因	■脳卒中（脳梗塞、脳出血）脳の血管が破れたり詰まったりして起こる病気の総称を「脳卒中」と言います。このうち動脈硬化が原因のものが「脳硬塞」です。脳硬塞には「脳血栓（血管そのものが動脈硬化等で閉塞する。大半はこちら）」と「脳栓塞（他の部位から、異物（栓子）が飛んできて血管が閉塞する）」があります。
治療方法	最近MRIなどの医療機器の発展により、安全な脳の検査が可能となりました。このため以前は外来では困難だった脳の健康診断が行えるようになってきました。当院でも従来の人間ドックに加え"脳ドック"を新たに行うことになりました。

表 5：失敗例

症状	以下に脳動脈瘤の違いによる穿通枝障害（脳硬塞をおこす）について述べてみます。
	・脳硬塞等の脳血管障害の患者さんを、当院の循環器内科

表 3 の結果より，各項目の平均では 0.62 の精度が得られた．これにより，上位には必要な情報が存在していることがわかった．しかし，不必要な情報も含まれているので，更に上位の項目に対して取捨選択をする必要がある．また，表 4 の成功例を見てわかるように，脳卒中の中に脳梗塞が含まれるため，両方の病名について書かれている文書も含まれていた．表 5 の失敗例は，その文書には症状のことについては述べていないので失敗としている．ただし，表 5 の上段の例の文書は文脈から判断して，前後の文書に症状のことについて詳しく述べられている可能性があると想像できる．

4. 今後の課題

実験の結果より，今後研究を進めていくにあたり，以下の課題について解決する必要がある．

(1) 文書の中に複数の病名について表記されている

脳卒中の中に脳梗塞が含まれているというように，病気というのは階層的な構造で表現されている．そのため，文書中に複数の病名について表記されていることがある．

(2) 不必要な情報が上位に含まれることがある

クラスタリングの結果の上位には，関係のない話題の不必要な情報が含まれていることがあった．この原因としては，passage 文書の単語数が多い場合，複数の話題があるためにこのようなことが起こると考えられる．よって，更に取捨選択をして，不必要な情報を取り除く必要がある．

(3) 文書が整っていない

本研究では，Web 文書を抜き出しただけなので，文書の形式に一貫性がない．

(4) クラスタリングに時間を要する

今回の実験では，分類した文書の全て(各項目 600 文書前後)を対象にして二文書間のクラスタリングを行った．そのため，一日程度の処理時間を要してしまう．このような状態では，本システムのように入力してから Web 文書を収集しクラスタリングを行う場合，リアルタイムに処理ができないので，時間の短縮が必要となる．

5. おわりに

本論文では，Web 文書から医療における話題を抽出して医療分野の統計的情報を抽出するためのシステムの提案と構築をし，その評価実験を行った．

本システムは，検索，項目分類，候補群選択の三つのモジュールから構成されている．検索では検索エンジンを利用して Web 文書を収集し，項目分類ではタグコードを利用して，収集した文書を passage ごとに分割することにより意味的にまとまったブロックに分割し，項目ごとに分類した．そして，候補群選択では，各 passage 文書の重要語を抽出し，二文書間の類似度を求めて文書間のクラスタリングを行い，候補群を選択した．

実験では，上位に必要な情報が存在していることがわかり，有効性が確認できた．更なる改良を行うことにより実用性を高められるのではないかと考えられる．

本論文で使用した話題抽出は，他の分野にも応用ができると考えられる．また，音声認識の統計的言語モデルを作成する際にも有効ではないかと考えられる．

謝辞

速水研究室の日高幸範氏には，本システムの構築に協力して頂きました．また，その他の速水研究室の皆様には研究討論会などを通じ貴重な御助言を頂きました．以上の皆様には，感謝の意を表したいと思います．

参考文献

[山本 05] 山本正範，延澤志保，太原育夫：Web 文書の抜粋を回答とする質問応答システム，言語処理学会第 11 回年次大会論文集，pp.1076-1079, 2005

[藤井 02] 藤井敦，石川徹也：World Wide Web を用いた辞典知識情報の抽出と組織化，電子情報通信学会論文誌 D-II, Vol.J85-D-II NO.2, pp.300-307, 2002

[松本 03] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原 正幸：日本語形態素解析システム『茶筌』version 2.3.3 使用説明書，2003

[余 03] 余東明，石川孝：コミュニティウェブにおけるアクティブ情報検索のトピック抽出，第 17 回人工知能学会全国大会論文集，2003