社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (12)
2008－ＮＬ－183 (12)
2008／1／22

blogを対象とした薬剤服用情報抽出手法

高橋 美佳　荒木 健治

北海道大学大学院 情報科学研究科

消費者が企業に対し安全・安心を求める傾向が強くなっており，製造者自身が進んで情報を収集する必要性が高まりつつある．レビュー投稿型サイトではなく消費者が記述しているblogから情報を収集することにより，情報を網羅することが必要となってきている．情報を網羅するために対象とする商品だけではなくその周辺の商品や分類名までを抽出基準決定のための材料とすることにより，データの不足を補い少ない情報からより多くの情報を収集することを試みる．本稿では，薬剤服用情報を収集して提示することをタスクとするシステムを構築し評価を行う．

Extraction of Information on Medication Usage from Japanese blogs

Mika TAKAHASHI　Kenji ARAKI

Graduate School of Information Science and Technology, Hokkaido University

Users often think that manufacturers should pay more attention to reliability and safety of products. From the viewpoint of the supplier, it is necessary to cover information by extracting opinions not only from review sites but also blogs. We try to collect more data from limited samples, using not only one particular product but also other products of the same type and their category name. In this paper, we propose a system that extracts information on medication usage from blogs and then evaluate it automatically.

1. はじめに

近年，消費者が企業に対し安全・安心を求める傾向が強くなっており，製造者自身が進んで情報を収集する必要性が高まりつつある．消費者は企業へとより安全で安心のできる商品開発や，問題が発生した場合の迅速な対応をもとめている．また，企業は消費者からの情報を待つ受身の態勢をとるのではなく，積極的に情報を集めることにより消費者へ安心を提供するケースが増加している．

本研究では，製造者がblogから商品に関する情報収集を容易にするシステム構築を目的としている．この際に必要となるのが，検索エンジンによる情報を含んだサイトのランク付けではなく，デマや勘違いを含む情報の網羅性であると考えられる．ユーザーは商品に関する情報を主に口コミ登録サイトや掲示板やblogで発信している．サイトの扱う情報の種類や商品の普及率によっては，口コミ登録サイトからの情報収集よりもblogからの情報収集が有効である場合がある．また，レビューを投稿することが目的である登録サイトは同一の

フォーマットの中で表示されているので一覧性が高く，情報の整理がしやすい．しかし情報の網羅性には欠けており，企業側としてもblogの記述を網羅することは有用であると考えられる．そこで本研究ではページランクなどで順位付けされた検索エンジンを用いず，ping情報を元に構築されているblog検索エンジンを用いて情報を抽出し，「企業の」情報収集を支援を目的としたシステムの提案を行う．

未知の少数の情報を探すという特徴から，抽出のための正解データがほとんど存在しない．そこで，関連する概念へと拡張して抽出スコア算定のための正解データを収集する．今回の実験では薬剤服用情報をタスクとしてシステムを構築した．そこで正解データとして用意するデータとして，特定の薬ではなくさまざまな薬を服用したあとの感想へと拡張し文章を用意する．

本稿では，作成したシステムの概要と対象データの拡張方法，及び評価実験と今後の予定について述べる．

2. 関連研究

blogからの評判情報の抽出に関する研究としては，奥村ら[2]の研究がある．この研究では評価情報が多い商品に対しての評価抽出及び判定を対象としている．しかしながら，記述の多い商品を対象としているこの手法は，薬剤に関する評価情報のように記述の少ない情報に対して的確な判定が行えないという問題点がある．

峠ら[4]は対象を特定の掲示板に絞り込み，そこからドメインの特徴語自動抽出を行っている．掲示板やレビュー投稿サイトはそれぞれのローカルルールがあり，特徴が類似する傾向にあるので有効であるが，blogでは記述の形式も異なっている．

医療分野におけるWeb利用に着目している研究として長沼ら[5]の研究がある．HTMLタグを手がかりにして分割したpassage[6]を文の類似度使って分類しユーザへ提示する研究である．passageには不要な情報も提示され，より一層の情報の絞込みが必要である．

Kimら[1]は医療情報を投稿しているblogへとアンケートを行い，患者はblogからさまざまな情報を得ようとしていることを示した．本研究ではblogから薬剤に関する情報を収集するタスクを設定しているが，薬剤に関する経験談は薬剤の種類が多く，薬剤あたりの服用体験談が一般の商品と比較して圧倒的に少数となっている．奥村らは評価情報の記述が少ない場合については考慮しておらず，検索対象を薬剤とした場合に適切な結果が得られない．そこで，本システムではまず，検索対象に特化したデータベースを使用することにより検索範囲を広げ，検索結果の件数を増やす．さらに，検索結果から情報を抽出する際に，抽出ルールを拡張することにより，より多くの体験談の取得を可能とする．

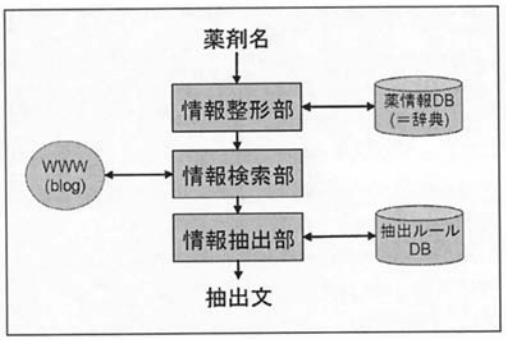

図 1: システム概要

3. システム

システムは大きく分けて図1に示すような情報整形部，情報検索部，情報抽出部の3つの処理部からなる．システムへの入力は薬剤名であり，blogから抽出した薬剤を服用した際の経験談が述べられた文が出力となる．

まず，情報整形部において入力に関連する情報をデータベースから取得して複数回検索を行うことで，検索結果の件数を増やすことを可能とする．さらに，情報抽出部において，上位概念を基に拡張して決定した抽出スコアを用いることで，より多くの情報を抽出することを可能とする．これにより，提案システムでは検索対象を薬剤という対象に絞った際にも，効率的な情報の取得が可能となる．

3.1 情報整形部

この処理部では薬剤データベースから情報を検索する際の商品名の差異を吸収するための処理を行う．クエリとして入力される薬の名称は製品名であり，データベースには成分名で登録されている．ある製品Aがクエリとして入力された場合に，同じ成分からなる製品Bもクエリとすることで，

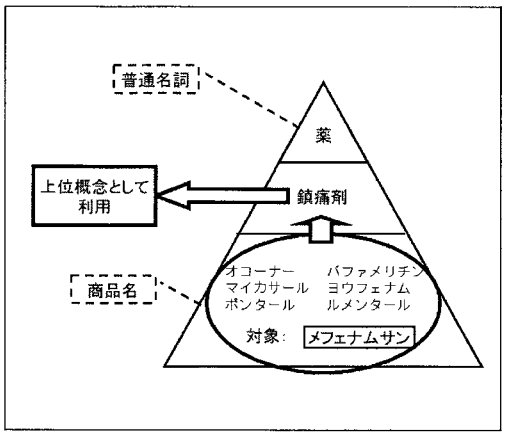

図 2: 概念拡張のイメージ

より多くの情報を得ることを可能とする．情報の拡張方法には 3.1.1 で定義する方法をとる．

3.1.1 対象範囲の拡張

抽出基準作成のためのデータを収集するために，基準クエリ集合の上位概念を利用する．上位概念について記述されている文章と，固有名詞で記述される文章は同様の傾向が見られる．そこで本稿では薬剤を対象としているので，上位概念として「頭痛薬」や「鎮痛剤」などの薬の分類名を用いる．上位概念の定義は薬剤データベースの分類を用いた．

次に，参考として同じ傾向のある対象の性質の登録を行う．今回対象としているクエリは，インターネット上の記述が少ないものを対象としている．同一成分や，同一部品の商品へと対象を広げることにより，より広範囲の情報提供を行う．

本研究では実験システムのタスクを薬剤情報収集としているので同一成分の薬をデータベースに登録している．すでに公開されているデータベースの一部を人手で登録しているが，どんな製品でも，何らかの部品や成分からなり必ず既存のデータベースが存在する．これらのデータを利用することにより，より広範囲の情報収集が可能となり，OEM による製品などのタスクでも適用できると予想される．

3.2 情報検索部

情報整形部により得られた薬剤の成分名・製品名をクエリとして blog 検索エンジンによる検索を行う．blog 検索エンジンを用いる理由としては，一般の検索エンジンを用いると，企業サイトに設置されている薬のデータベースまでもが検索結果として含まれてしまうためである．

本研究ではテクノラティブログ検索のキーワード検索 API[7] を利用した．テクノラティブログ検索のキーワード検索 API では Google のようにページランクをつけて結果を順位付けするのではなく，ping を収集して新着順に表示している．公開している blog の大半は ping を送信する性質を利用し，情報の網羅率を高めるために利用した．

ping に含まれる情報としては blog タイトルと記事 URL があるが，API で検索を行うことで，加えて投稿日時，記事タイトル，記事の検索クエリ周辺の文章を取得することができる．インターネット上の商品情報は「発売前に予想される情報」と「発売後に実際に使用した上での情報」の 2 つに分類することができる．今回は明確な発売日を考慮するタスクではないため，投稿日時は利用しないが，発売日を考慮する必要がある際にも容易に発売後の情報を用いることができる．一方，RSS では更新された日時が表示され，HTML 上ではユーザー自身が自由に設定した日時が表示される．記事全体が最初に更新された時間を取得するには，ping からの時間情報が有効である．さらに，blog 検索エンジンは一般的な検索エンジンと違い blog に特化しているので，複数記事が含まれる blog のトップページやカテゴリごとにまとめて表示したページは検索結果として出現しないため，確実に対象記事の記述がある blog の記事を効率的に取得できる．

3.3 情報抽出部

情報検索部で取得した検索結果の blog 記事から，スコアに基づき体験談の抽出を行う．blog 記事の本文には薬を服用した体験談とそれ以外の薬とは関係のない部分が含まれているため，これらの部分を分離する必要がある．このとき使用する文のスコアを，図 2 を基に拡張することで，より多くの情報の取得を可能とする．

まず，取得した blog から本文部分を抽出する．HTML タグを取り除き，API による検索結果にある投稿記事タイトルを利用して本文部分を取得する．一般的な blog には，コメントとトラックバックの機能が実装されている．つまり記事本文のページ構造上の特徴として，投稿記事タイトルと「コメント」や「トラックバック」というキーワードの間に記述されていることとなる．そこで，本稿で

```
<!-- generator="Technorati API version 1.0 /search" -->
-<tapi version="1.0">
  -<document>
    -<result>
      <query>フラペリック</query>
      <querycount>35</querycount>
      <rankingstart/>
    </result>
    -<item>
      -<weblog>
        <name>幻燈鏡月(偲</name>
        <url>http://lunamoony.exblog.jp</url>
        <rssurl>http://Lunamoony.exblog.jp/index.xml</rssurl>
        <atomurl>http://Lunamoony.exblog.jp/atom.xml</atomurl>
        <inboundblogs>3</inboundblogs>
        <inboundlinks>72</inboundlinks>
        <lastupdate>2007-10-14 17:57:06 GMT</lastupdate>
      </weblog>
      <title>フラペリック錠</title>
      -<excerpt>
        熱も下がり、体調もだいぶ良くなってきたんです、が。今、あたしの身体
        る音が全て、いつもより低く聴こえる」... で貰った咳止め薬「<strong clas
        ネット検索して分かったことなんですが、<strong class="keyword">フラ
        て聴こえる」とい ... ないつもと違うお声!! ってか「<strong class="keyword
      </excerpt>
      <created>2007-10-14 17:48:05 GMT</created>
      <permalink>http://Lunamoony.exblog.jp/6308438/</permalink>
    </item>
    -<item>
      -<weblog>
```


図 3: テクノラティブログ検索による検索結果

は投稿記事タイトルと「コメント」「トラックバック」間のページ内でもっとも長い文章群を記事本文と定義する．

近年の blog の文章の特徴として，

- 句読点がなく，改行で句読点の代用をしている
- 段落の概念がない (文の間に必ず 1 行挟む)

といった傾向が見られる．これらのことから，行をまたいだ「一文」を定義するには形態素解析ツールに頼ることとなり非常に難しい．そこで本研究では，「改行まで」あるいは「句点」など確実に文の終了を表す記号までを 1 文とした．

文単位に分割したのち，形態素解析を行い，3.3.1 で述べるスコアを求める．本システムは Java で構築されているため，MeCab[9] を Java へ移植した Sen[8] を利用した．形態素解析に用いた辞書は Sen の標準辞書である IPADIC に，実験条件としている薬剤の固有名詞を登録したカスタム辞書を作成して構築した．

3.3.1 文章スコア

事象 A を抽出対象文であると定義し，事象 B を単語 w が文に含まれる事象と定義する．事象 B が与えられたときの事象 A の条件付確率を単語スコアとすると式 (1) のようになる．

$$P(A|B) = \frac{\text{単語 } w \text{ が抽出対象文に含まれる確率}}{\text{全文章中で単語 } w \text{ が含まれる確率}} \quad (1)$$

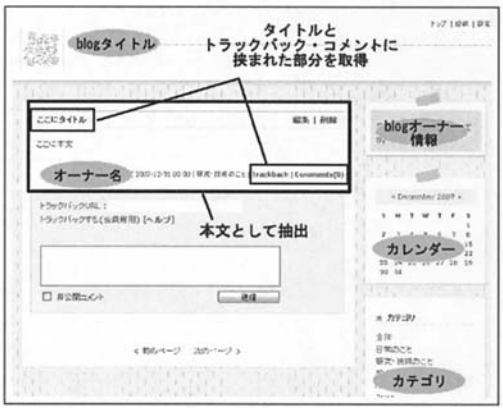

図 4: blog 記事本文

図 5: システム画面

その文章が抽出文であるかどうかの確率は本来，対象とする文章に含まれる単語のスコアの積であらわされるべきであるが，式 (2) の式より抽出対象文に含まれていない単語のスコアが 0 であるため，未知語が抽出すべき文のスコアを 0 にしてしまう．そこで，本手法では文章のスコアを含まれる単語スコアの平均とする．

$$Score = \frac{\Sigma P(word)}{\text{スコア計算の対象となった単語数}} \quad (2)$$

式 (2) によるスコアをもとに抽出文を決定する．

4. 実験

3 章で示したシステムを用いて薬品に関する記述をどの程度抽出できるかを評価する．

表 1: HTML からの本文抽出精度

分類	記事数	割合
抽出成功	29	68%
抽出失敗	21	42%

表 2: 抽出失敗の分類

原因	記事数
異なるタイトルを ping 送信する blog	8
コメント・トラックバックを拒否している	4
タイトルに除去対象の記号 (<) を含む	4
ping 送信後にタイトルを変更した	1
メニュー部分の誤抽出	3
エラーページを誤取得	1

まず，実験で利用する正解データについて説明する．本研究では blog 検索 API の検索結果として得られた blog のエントリを抽出対象としている．本実験では検索結果上位 500 件の HTML を取得し，情報抽出部と同様の処理で HTML を除去，本文抽出を行う．実験で利用する検索クエリとしては，同一成分で別の名称を持つ薬剤 7 種と，上位語 1 種で正解データを作成した．52,384 文の本文が得られ，72 文を正解データとした．

本実験では，**3.3** で述べた方法で HTML から blog 記事の本文を抽出している．あるクエリの検索結果 50 件の抽出精度を調査したところ，表 1 に示すような結果が得られた．本文抽出に失敗した blog の原因としては，表 2 の通りである．

異なるタイトルを ping 送信する blog では，タイトルと blog 記事の分類タグを同時に ping 送信するが，実際の blog では逆順で表示しているなど，本文領域を断定できなかった失敗例が表の二重線よりも上の 4 原因である．これらの原因に当てはまる blog 記事からは，本文およびメニュー部分を含む文を対象として取得している．つまり，先ほど述べた 52,384 文については本文と除去しきれなかったメニュー部分の総数となっている．一方，メニュー部分を誤検出した例や，エラーページを取得した例では本来の本文を除去してしまっているので，52,384 文には含まれず，その中から決定した 72 文の正解データは誤って除去された本文には含まれない．

それぞれの薬剤のヒット件数と正解文数は表 3 の通りである．これらの薬剤のうち，マイカサールとオコーナーについては対象からはずした．マイカサールについては検索結果が 0 件であったので

表 3: 薬剤のヒット件数と正解文数

薬剤名	ヒット件数	正解文数
オコーナー	1932	0
バファメリチン	2	3
マイカサール	0	0
ヨウフェナム	1	15
ポンタール	215	3
メフェナムサン	14	22
ルメンタール	19	3

本文は抽出できない．オコーナーについては，ヒット件数上位 500 件がすべて人名のオコーナーについてであった，また，他の薬剤 20 種類についてもヒット件数と検索結果の blog のエントリの記述の関連性を確認したところ，2 種類はヒット件数が 1,000 件を超えていたが，他の意味を持つ言葉であるか，その一部であった．18 種類については 1,000 件を超えず，検索のクエリとした薬剤についての記述が確認できた．そこで，本実験では，すべての検索結果は薬剤「オコーナー」についてではないと判断し，対象から除外した．

4.1 結果と考察

本実験における適合率と再現率，F 値については次の式で定義する．

$$\text{適合率 } (Recall) = \frac{\text{システムで出力した正解文}}{\text{システムで出力した文}} \tag{3}$$

$$\text{再現率 } (Precision) = \frac{\text{システムで出力した正解文}}{\text{正解文}} \tag{4}$$

$$F \text{ 値} = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{5}$$

本手法との比較対照として，奥村ら [2] の手法を用いてサービスを提供している SHOOTI（シューティ）[3] で同様の検索を行った．SHOOTI は日本中の blog・レビューサイトからクチコミ情報だけを検索して提示するシステムで一般に公開されている．このシステムは本手法同様，ping サーバーから blog を収集しているため，他手法よりも抽出対象が同じ blog となる傾向となっている．

出力文としては，記事を削除されていない blog から本文抽出精度 100 % と仮定して出力を集計した．文への分割は本手法の基準を適用し，SHOOTI の検索結果のサマリーを文単位に分割して集計した．サマリーとしての出力の文字数は可変で，文

表 4: 出力結果

手法	正解文	出力文	出力中の正解文	Recall	Precision	F 値
提案手法	22 文	30 文	12 文	0.400	0.545	0.486
SHOOTI	22 文	70 文	11 文	0.157	0.500	0.239

の途中から開始したり，文の途中で終了したりしているが，その文も集計に加えた．

SHOOTI では本実験で正解データと定めた文章の一部が検索結果のサマリーで表示された場合に正解とした．本手法では「メフェナムサン」というクエリに対し，12 文の正解データが得ることができ，同様のクエリで SHOOTI は 11 文の正解データが得られた．SHOOTI では 1 文単位ではなく，その周辺も同時に表示しユーザへと提示するので，本稿で定義した 1 文へと分割すると出力文が 70 文と大量になってしまっている．本手法では情報を網羅しながらも，的確な出力を行っているということができる．

本実験の出力結果で最も高いスコアとなった出力文章が「歯痛歯痛歯痛」となっているが，これは構成している単語数が少ない文章に，複数個の「抽出文に含まれやすい単語」が含まれていた場合，確実にスコアが高くなるためである．しかし，このような短い文章では十分に服用に対する感想を読み取ることができないので，文章の長さによる文章スコアへの重み付けを検討する必要がある．

5. まとめと今後の予定

本稿では，薬剤服用経験談を Web 上，特に blog から抽出してユーザへ提示するシステムについて述べた．薬剤服用情報解析システムを構築したのち，経験談抽出についての評価を行った．

今後は，検索クエリと抽出結果の因果関係を判断するために，クエリ集合で抽出した文章同士の類似度を求め，共通する現象を提示するシステムの構築や，拡張概念の自動取得の検討を行いたい．

参考文献

[1] Sujin Kim, Deborah S. Chung, “ Characteristics of cancer blog users “, Journal of the Medical Library Associationl, 95(4), 2007.

[2] 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕, “ blog ページの自動収集と監視に基づくテキストマイニング“, 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01, 2004.

[3] 株式会社ブログウォッチャー SHOOTI(シューティ）, http://shooti.jp/

[4] 峠泰成, 大橋一輝, 山本和英, “繰り返し学習を用いた話題に順応する意見文抽出 “, 情報処理学会, 情報学基礎研究会, pp43-50, 2004.

[5] 長沼 潔, 速水 悟, “ 医療分野における Web 文書からの話題抽出方法 “, 人工知能学会全国大会，2005.

[6] 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

[7] テクノラティブログ検索 キーワード検索 API, http://www.technorati.jp/developers/api/

[8] 形態素解析器 Sen（Java 形態素解析ツール）, http://ultimania.org/sen/

[9] 工藤拓，MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.jp/

O(≧▽≦)Oバファ○ン達では効かなかったぁたしの腹痛も一気に落ち着くぜぃ（V^?°）ちょっと高いけど…今日はPSPでみんごる中。

別の薬をもらいにある薬局へいってみたら、そのまま「メフェナムサン」という薬が！

これはホントによく効きます

痛みに弱い私が、もうお守りのように常に持ってる薬がコレ

消炎・鎮痛・解熱に優れた効果を発揮メフェナムサンカプセル

まだ頭が微妙に痛いよ☆もう薬飲んじゃおうかな☆私の常備薬☆よく効くのでおススメです☆

図 6: 出力例