1G3-3

Webから抽出した歴史的イベント情報の地図上への再編集

Extraction of Historical Event Information from the Web and Recompilation on a Map

田村 和香子[*1]　小林 一郎[*2]
Wakako Tamura　Ichiro Kobayashi

[*1]お茶の水女子大学理学部情報科学科
Dept. of Information Sciences, Faculty of Science, Ochanomizu University

Today, there are quite a few Web pages on the Internet, however, the amount of information obtained from each Web page varies. Therefore, in order to obtain the information we want, we have to brwose many Web pages and that makes us very laborious. In this study, we aim to extract information from Web pages and recompile the pieces of information extracted from the pages to make a more informative document. Especially, we focus on event information, i.e., when, where, and what; recompile the extracted information; and display it on a map. As a concrete example of the proposed method, we apply our method to recompile historical event documents and display the recompiled information on a map.

1. 研究背景と目的

現在，Web 上には多量の Web ページが存在しているが，個々のページから得られる情報量は異なっている．そのため，自分の知りたい情報を集めるためには多数のページを閲覧しなくてはならず，多大な労力を必要とする．このことから本研究では，Web ページからの情報抽出とその再編集を目的とする．特に，本研究では時系列情報と住所情報を有するイベント情報 (「いつ，どこで，何があった」というような情報) に着目し，抽出した情報を地図上に整理して表示することを目指す．今回は特にイベント情報を抽出しやすい歴史的事柄に関して，提案する枠組みの適用を試みる．

2. 歴史的事柄の情報再編集

歴史的事柄に関する Web ページの例を図 1 に示す．

「新選組」略年表

1863	文久3	2	2/23 幕府浪士組230人余り、壬生村に到着。土方ら八木邸泊。 その夜、清河八郎、新徳寺にて尊王攘夷の意志を表明。
		3	浪士組に東下の命令くだり、清河ら約220人江戸へ戻る。
		3	近藤勇、芹沢鴨ら17人、松平容保に嘆願書提出。会津藩お預かりとなり壬生浪士組と名乗る。
		8	8月18日の政変。壬生浪士組、御所の南門を守る。
		9	芹沢鴨・平山五郎、八木邸にて暗殺される。
		9	9/25「新選組」の隊名与えられる。
1864	元治1	6	池田屋事件。
		6	明保野亭事件。
		6	長州軍の攻撃に備えて竹田街道銭取橋付近に布陣。
		7	蛤御門の変。(禁門の変)
		7	新選組、真木和泉ら17人を天王山に追いつめ自刃させる。
1865	慶応1	2	山南敬介、脱走の罪により切腹する。
		3	屯所を壬生から西本願寺の集会所へ移転する。
		3	伊東甲子太郎ら13人御陵衛士を拝命し、新選組を脱退。

図 1: 歴史的事柄に関する Web ページの例

一般に，このようなページには，事柄が年表の形式でまとめられている場合が多い．このことから，歴史的事柄の時系列情報と事柄の説明を一括で取得することが出来る．本研究では情報の再編集の主旨を，情報を時系列順序に整理し，事柄の説明を地理情報に関連付けて行うこととする．

連絡先: 田村和香子，お茶の水女子大学理学部情報科学科小林研究室，g0320522@edu.is.ocha.ac.jp

3. システムの概要

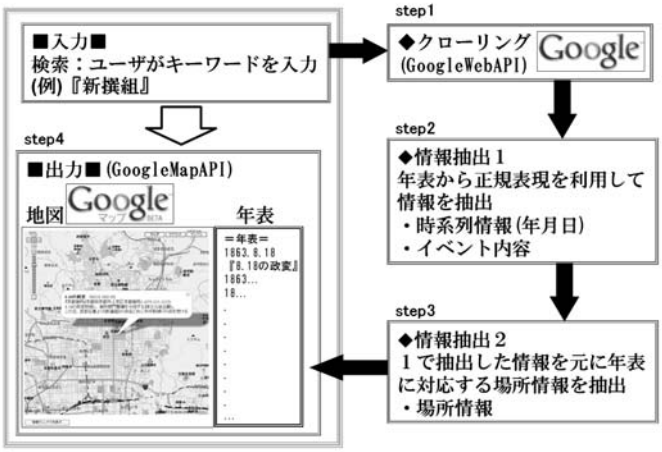

図 2: システムの概要

以下，処理の手順ごとにその内容を説明する．

step1．年表を含む Web ページの取得
イベント情報のうち,『いつ，何があった』というイベント内容と時系列情報を一括で取得するために，検索キーワードを歴史的事柄に限定し,「検索キーワード＋年表」というクエリを用いて，既存の検索エンジンで検索を行う．

step2．取得 Web ページの解析および情報抽出
step1 において取得した年表を含む Web ページの HTML タグを解析し，イベント内容と時系列情報を抽出する．年表の多くが table タグを用いて書かれていることを利用し，table タグ内から必要な情報を取得する．

step3．年表情報に対応した場所情報の抽出
step2 において抽出してきたイベント情報に対応した場所情報を取得する．場所情報の判別には CaboCha/南瓜 [3] を用いる．

step4．取得情報の GoogleMap 上への表示
step1〜**step3** で取得してきた Web ページの情報を地図上に再編集して表示する．

3.1 年表情報を含む Web ページの取得

本研究では検索および検索結果 Web ページの HTML 取得に GoogleWebAPI[1] を使用している．今回は，歴史的事柄として「新撰組」を採用し,「新撰組＋年表」のクエリを用いて，検索したページ上位 10 件を情報取得のための対象となる Web ページとした．

3.2 年表情報の抽出

取得された上位 10 件の Web ページのうち，記載されている年表の HTML 表記を分析した結果，大半が table タグを用いて書かれていることが判明した．正規表現を用いることにより，table タグ内の情報を HTML ファイルから取得し，さらに正規表現を用いてその中からイベント内容と時系列情報を抽出する．

ソースコード
＜ TITLE ＞新選組詳細年表＜/TITLE ＞
(中略)
＜ TABLE border=”1”＞
＜ TR ＞
＜ TD ＞文久３年＜/TD ＞
＜ TD align=”right”＞２月　２３日＜/TD ＞
＜ TD ＞幕府浪士組 230 人余り、壬生村に到着。＜/TD ＞
＜ TD ＞１８６３年＜/TD ＞
＜/TR ＞
(中略)
＜/TABLE ＞

3.3 取得したイベント情報に対応する場所情報の取得

取得されたイベント内容を表すテキストから場所情報を抽出する．テキスト中における場所情報の特定には，CaboCha/南瓜 [3] を用いる．CaboCha/南瓜で解析したテキストから名詞を表す品詞情報の内，地域を含むものを正規表現を用いて抽出する．CaboCha/南瓜による解析例を以下に示す．

CaboCha/南瓜の解析結果例
* 0 1D 2/2 0.37332273
幕末 バクマツ 幕末 名詞-一般 O
浪士 ロウシ 浪士 名詞-一般 O
組 グミ 組 名詞-接尾-一般 O
、 、 、 記号-読点 O
1 2D 1/2 0.00000000
壬生 ミブ 壬生 名詞-固有名詞-地域-一般 B-LOCATION
村 ムラ 村 名詞-接尾-地域 I-LOCATION
に ニ に 助詞-格助詞-一般 O
2 -1O 0/0 0.00000000
到着 トウチャク 到着 名詞-サ変接続 O
。 。 。 記号-句点 O O

これにより，上記年表のイベント内容を表すテキスト「幕府浪士組 230 人余り、壬生村に到着。」から「壬生村」が場所情報として抽出できる．しかし，これだけでは場所情報抽出の精度が不十分なので，場所情報抽出の精度を向上させる為に，格助詞を利用することを検討中である．文献 [2] による格の分類において，一般的に場所情報は,「ニ格」,「デ格」,「カラ格」,「マデ格」などの格によって担われる．このことに基づき，CaboCha/南瓜 [3] を用いて，テキストを形態素解析，係り受け解析し，上記４つの格が受ける語を抽出し，それらが場所を示す語であるかを，特定の述語が取る格の関係を編集した辞書 (文献 [4] においては「意味格抽出用辞書」と言っている) を用いて判断する予定である．

3.4 地図上への再編集

取得してきたイベント情報 (イベント内容，時系列情報，地名又は住所情報) を GoogleMapAPI[5] を用いて再編集して出力する．GoogleMap が日本語の地名からジオコーディング可能であるため，地名又は住所情報に関しては，抽出してきた文字列をそのまま利用する．

図 3: 出力例：地図上への表示

4. まとめと今後の課題

本研究では Web からのイベント情報の抽出とその再編集を目的とし，その一環として，イベント情報を取得しやすい歴史的事柄を対象として手法の提案を試みた．現在，GoogleMap 検索では，あらかじめ登録されている Web ページのみが検索結果として返ってくるのに対して，本研究では，Web 上の未登録の Web ページから情報を取得し，地図と関連付けて情報の再編集を試みている．しかし現段階では，抽出した情報の再編集という点において，まだまだ汎用性が低いという問題点があり，今後の課題としては，本研究で構築したシステムを拡張し汎用性を高めるということが上げられる．

参考文献

[1] http://www.google.com/apis/

[2] 益岡隆志，田窪行則：基礎日本語文法，くろしお出版，1992.

[3] 奈良先端科学技術大学大学院松本研究室，日本語構文解析器「CaboCha/南瓜」，
http://chasen.org/ taku/software/cabocha/

[4] 河野安友未，小林一郎：Web 上のヘッドラインニュースを情報源とした質問応答システムの構築，FIT2005 第 4 回情報科学技術フォーラム

[5] http://www.google.com/apis/maps/

[6] 星野厚，岡瑞起，加藤和彦：位置情報を用いたブログサービス“ ろぐの細道 ”の提案 社団法人　電子情報通信学会　第二種研究会資料 Ｗ１２-２００６-５１（2006.7）

[7] 藤本典幸，森本泰貴，長屋務，萩原兼一：ウェブ検索 API とトピック主導型クローリングに基づくロボット型住所関連情報検索システム　社団法人　電子情報通信学会　第二種研究会資料 Ｗ１２-２００６-６６（2006.7）