Web からのイベント情報の抽出および情報の再編集

田村和香子　(指導教員：小林 一郎)

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 [1]

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，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．

3　処理の概要

処理の概要を図 2 に示す．

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，GoogleAPI を用いて検索を行う．

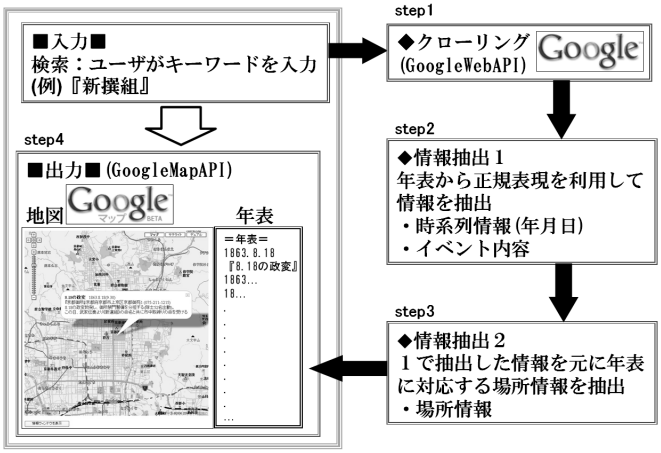

図 2: システムの概要

step2 . 年表情報の抽出

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してくる．このとき，年表の多くが table タグを用いて書かれていることを利用する．

step3 . 年表情報に対応した場所情報の抽出

step2 において抽出したイベント情報から場所情報を抽出する．

step4 . 抽出したデータの書き出し

step2，**step3** において抽出したデータを GoogleMapAPI へ読み込ませるために，JSON(JavaScriptObjectNotation) 形式でテキストファイルに書き出す．

step5 . 取得情報の再編集および地図上への表示

step4 において作成したテキストファイルからデータを読み込み GoogleMap 上へ表示する．

3.1　年表を含む 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[2] を使用している．今回は，歴史的事柄として「新撰組の活動」を対象とし,「新撰組＋年表」のクエリを用いて，検索したページ上位 10 件を情報再編集のための対象 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．このことから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 年表情報に対応した場所情報の抽出

取得されたイベント内容を表すテキストから，地名を表す名詞を抽出する．地名を表す単語の抽出には日本語構文解析器 CaboCha/南瓜 [3] を用いる．イベントの内容を表すテキストを解析し，名詞を示す品詞情報の内，地域を含むもの (例：名詞-固有名詞-地域-一般) だけを正規表現を用いて抽出し，その名詞をイベントが発生した場所情報を示すランドマークとして用いる．

3.4 抽出したデータの書き出し

年表から抽出したイベント内容，時系列情報，場所情報を JSON 形式でテキストファイルに書き出す．JSON 形式で表された実際のイベント内容を以下に示す．

JSON 形式のデータ

```
{ data : [
{ year:”1863 年”,month:”2 月”,days:”23 日”,add:”壬生”,
doc:”幕府浪士組 230 人余り、壬生村に到着。”},
] };
```

3.5 取得情報の再編集および地図上への表示

複数の Web ページから取得した年表情報を用いて，より情報量の多い詳細な年表を再編集する。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，各イベントから抽出された場所情報をそのまま GoogleMapAPI[4] に入力し，イベント内容を地図上に表示する（図 3）．

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，その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．今後の課題としては，取得した Web ページからのより正確な情報の抽出，およびそれらの再編集機能の充実を通してシステムの汎用性を高めるつもりである．

参考文献

[1] http://www.eva.hi-ho.ne.jp/iwaemon/sinsen-nenpyou.htm

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

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

[5] 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）

[6] 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），http://www-hagi.ist.osaka-u.ac.jp/ fujimoto/WebSearch/Address/index.html