Vol.2010-NL-196 No.4
Vol.2010-SLP-81 No.4
2010/5/27

Web レビューの表形式化システム SECRET の開発

今井和雄[†]　吉村友希[††]　原田実[††]

近年、Web の様々なレビューサイト上のレビューの重要性が高まってきており、それらを有効活用するためにレビューサイトから自動でレビューを取得し表形式化して出力するシステム SECRET を開発することにした。SECRET はサフィックス木を用いてページ中の繰り返しを探すことで HTML のどの部分がレビューにあたるかを判定する。さらに意味解析システム SAGE を用いることで HTML に記述された様々な語句からレビューの属性として適切なものを自動で判別し、テキストマイニングツール STM に入力可能な CSV ファイルとして表形式化して出力する。9 種のサイトからレビューを取得する評価実験においては、レビューの属性のうち 72%の属性を取得し、それらに 99%の精度で適切な属性名を割り当てることができた。

The development of the system SECRET which tabulates Web reviews

KAZUO IMAI[†]　YUKI YOSHIMURA[††]
MINORU HARADA[†††]

In recent years the importance of the reviews on the various review sites in the Web rises. We developed system SECRET which acquires the reviews from the review site, tabulates them and outputs them in the automatic operation to use them effectively. SECRET judges whether the part of the HTML hits a review by looking for repetition in a page with a suffix tree. Furthermore, it is automatic and distinguishes an appropriate thing as an attribute of the reviews from the various words and phrases which are described in HTML by using semantic analysis system SAGE and it formalizes a list as the CSV file which can be input into text mining tool STM. In the evaluation experiment to acquire reviews from nine kinds of sites. It acquired attributes of 72% among the attributes of the reviews and was able to assign an appropriate attribute name to them with precision of 99%.

† 青山学院大学大学院　理工学研究科　理工学専攻　知能情報コース
†† 青山学院大学　理工学部　情報テクノロジー学科

1. 序論

近年、コンピュータの普及に伴い、インターネットを利用する機会もますます増えてきており、Web ページから様々な情報を得ようとすることも多い。現在のウェブページは様々な形式によって幅広い情報を提示しており、ユーザはその膨大な情報の中から必要とする情報を手動で見つけることが求められる。特に商品の購入やサービスの利用にあたって、商品情報そのものの他に、商品の購入者やサービスの利用者の実際の意見の有用性が高まっており、このような情報をレビューサイトを通じて得ることが一般的になっている。

しかし、レビューサイトを利用しても、商品やサービスによっては大量のレビューを持つものが多数存在し、有益な情報をピンポイントで得たり、レビューの傾向を判断したりするのは困難である。また企業にとっては、レビューのテキスト情報とレビューの様々な属性(投稿者の名前、性別、年齢、投稿日時、評価)とを関連付けて、レビュアーの意見の傾向を分析し商品やサービスの改善を行う必要があり、これらを支援するツールとしてテキストマイニングツールが普及しつつある。しかし、Web 上のレビュー情報をこれらのツールに入力できる形式に変換する作業はいまだ手作業に頼っているのが現状である。

そこで我々は、機械処理によって Web 上のレビューサイトからレビューを自動で収集し、表形式化することが必要と考えた。原田研究室では、2006 年度に意味解析システム SAGE[1]を用いたテキストマイニングツール STM[4]が開発されている。STM では SAGE による意味解析機能を用いて、アンケート文を意味グラフに展開し、2 つの意味グラフの対応する節同士の概念的な類似度や節間の深層格の類似度をベースに、類似部分グラフの大きさで 2 文の類似度を計測する。これにより、表現が異なっていても同内容の意見を持つ文を集約し分類することができる。このような背景から、レビュー本文のようなテキストデータやその他の属性などから有益な情報を得るために、Web サイトからレビューを自動収集し、STM が入力可能な形式として出力することを本研究の目的とした。

本研究で開発した SECRET(*SEmantic CRawling Enriched Tabularisation*)は、サフィックス木を利用してレビューページ中の繰り返しを発見することで HTML からレビュー内容を取得する。また、SAGE による意味解析を用いることで表形式化する際に適切な属性名を決定することができる。この方法について以下で詳しく述べる。

2. SECRET のシステム概要

2.1 システムの概要

SECRET は図 1 に示すように Web サイトの URI から 3 つの過程を経て表形式化されたレビュー情報を出力する。以下に入力と処理の概要について述べる。

Vol.2010-NL-196 No.4
Vol.2010-SLP-81 No.4
2010/5/27

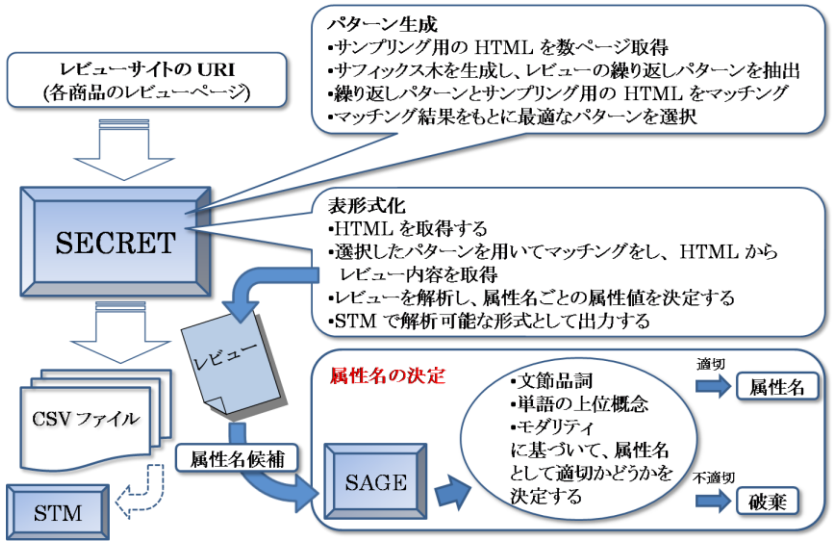

図 1 SECRET のシステム構成

- 入力

  SECRET ではレビューページの URI を入力とする。各商品のレビューページの URI を入力することによって、その商品のレビューページの HTML を取得することができる。

  - 例：http://review.kakaku.com/review/K0000058977/
  - 例：http://www.psmk2.net/title.php?title=335

- パターン生成

  レビューサイトはサイトごとにその構造が異なるため、HTML からレビュー内容を取得するためには、レビューサイトごとに HTML の構造を解析する必要がある。レビューサイトにおいてユーザが投稿したレビューは、HTML の構造が同じパターンがレビューの数だけ繰り返されていることが多い。Web ページから繰り返しのパターンを発見できれば、そこからレビューを抽出することが可能となる。 繰り返しのパターンを発見するためにはいくつかのサンプルが必要となるため、入力された URI に基づいてサンプリング用の HTML を数ページ取得する。取得した HTML を簡略化し、それをもとにサフィックス木(接尾辞木)を生成する。サフィックス木から抽出したパターンとサンプリング用の HTML をマッチングし、得られた候補の中から最適なパターンを人手で決定する。

- 表形式化

  HTML を取得し、最適な繰り返しパターンを用いてマッチングを行い、HTML からレビュー内容を取得する。これらを CSV ファイルとして表形式化するため、レビューを構成する属性名とそれぞれにあたる属性値を決定する。SAGE 解析により決定した属性名が属性値をもたなかった場合、その属性名は破棄する。

- 属性名の決定

  属性名の決定にあたって、HTML のレビュー内から属性名の候補を選出する。レビューの特性により、レビューのパターン内において常に同じ単語が繰り返されているものを属性名の候補とする。逆に、レビューによって異なるものは属性値の候補となる。属性名の候補を SAGE 解析にかけ、4 章で述べる条件を満たしたものを属性名とする。

2.2 STM に対応させた CSV ファイルの出力

SECRET では、図 2 に示すように STM に対応させた CSV ファイルとして出力する。この表の各行の形式は以下のようである。

	A	B	C	D	E	F	G	H	I
1	SampleID	ユーザ名	年齢	投稿日時	レビュー本文	購入した場所	持ちがよい	引き締め	ハリ・弾力
2	AN	AI	AN	AD	QF	AS	QC	QC	QC
3	-	-	-	-	-	-	-	-	-
4	1	サミー1006	32	2010/1/20 0:00	マルチに使える商品だと思います。	02 バラエティショップ	0	0	0
5	2	じゅんぴん	30	2010/1/20 0:00	いろんなリップを試しましたがこれ以	01 スーパー・ドラッグ	0	0	0
6	3	*しおり@*	19	2010/1/20 0:00	コストコで大きいサイズを買って以来	02 バラエティショップ	0	0	0
7	4	cosmepuff	28	2010/1/20 0:00	ベトベトしてしまうので、日中使用に	01 スーパー・ドラッグ	0	0	0
8	5	。+心菜+。	25	2010/1/19 0:00	かなり前にでっかいサイズをDSで買	01 スーパー・ドラッグ	0	0	0
9	6	ココア36	37	2010/1/19 0:00	なんと言っても、この量でこの値段は	01 スーパー・ドラッグ	0	0	0
10	7	mirity	30	2010/1/19 0:00	使い切ったことがない。。一番小さい	01 スーパー・ドラッグ	1	0	0
11	8	ぐみ	24	2010/1/18 0:00	なかなかなくならないのでコスパは	01 スーパー・ドラッグ	0	0	0
12	9	ちっぱー☆	22	2010/1/18 0:00	どんなリップクリーム使っても駄目だ	01 スーパー・ドラッグ	0	0	0
13	10	★puu★	29	2010/1/18 0:00	顔に塗って寝ると、必ず吹き出物が	01 スーパー・ドラッグ	0	0	0
14	11	はんなちゃん♪	22	2010/1/18 0:00	多用してます・①毎日のクリーム②全	01 スーパー・ドラッグ	0	1	0
15	12	*マイリー*	26	2010/1/18 0:00	リップクリームとして使っていました。	01 スーパー・ドラッグ	0	0	0
16	13	pine-apple	30	2010/1/17 0:00	家族全員でこの季節には手放せない	01 スーパー・ドラッグ	0	0	0
17	14	ちゃと美	29	2010/1/17 0:00	私の 中で定番過ぎて、口コミする事	01 スーパー・ドラッグ	0	0	0
18	15	baby	15	2010/1/17 0:00	コレ大好きです<3安いし、減りが少	02 バラエティショップ	0	0	0
19	16	canaryxkachun	18	2010/1/17 0:00	なんとなく購入しましたが、石油臭く	02 バラエティショップ	0	0	0
20	17	kiss-yu	24	2010/1/16 0:00	小さい方のを買いました♪ﾘｯﾌﾟｸﾘｰﾑ	03 通信販売・ネット	0	0	0
21	18	ｺﾏﾄ	27	2010/1/16 0:00	乾燥肌兼敏感肌の友人がいつも使っ	01 スーパー・ドラッグ	0	0	0
22	19	みいちあむ	17	2010/1/16 0:00	なんとなく有名だからという理由で買	01 スーパー・ドラッグ	0	0	0
23	20	くき子	13	2010/1/15 0:00	口コミ1位だったので、買ってみまし	01 スーパー・ドラッグ	1	0	1
24	21	りさ(ﾟ∀ﾟ*●))	20	2010/1/15 0:00	元男のりさが通るわよ～ん(ﾟOﾟ)*ﾟ★	01 スーパー・ドラッグ	0	0	0
25	22	azarashisennsei	41	2010/1/15 0:00	初めて見た時は、ジェリー？脂身み	01 スーパー・ドラッグ	0	0	1
26	23	どらみ0918	32	2010/1/15 0:00	冬場には欠かせない商品で、家族で	01 スーパー・ドラッグ	0	0	0
27	24	SUMMER	26	2010/1/14 0:00	ヴァセリンの大ファンです(*ﾟ∀ﾟ*)ｵ	02 バラエティショップ	0	0	0
28	25	a.r.s	24	2010/1/14 0:00	久しぶりに買ってみました・リップと体	01 スーパー・ドラッグ	0	0	0
29	26	m8r	17	2010/1/14 0:00	使い続けてもう1年以上。毎日使って	01 スーパー・ドラッグ	0	0	0
30	27	りゅー*ゅみ	26	2010/1/13 0:00	唇の荒れが酷くて買ってみました。	01 スーパー・ドラッグ	0	0	0
31	28	ナバロ	24	2010/1/13 0:00	突然手の甲が荒れに荒れたのでこち	01 スーパー・ドラッグ	0	0	0
32	29	みっふぃ…	17	2010/1/13 0:00	私は主にシャドウベースと失敗して	01 スーパー・ドラッグ	0	0	0
33	30	●ぐるぐるどかん	36	2010/1/13 0:00	唇、手荒れの保湿効果バッチリ！顔	01 スーパー・ドラッグ	0	0	0
34	31	quelquefois	24	2010/1/12 0:00	ハンドクリームに凝っていた時期もあ	01 スーパー・ドラッグ	1	0	0
35	32	▽.nori	15	2010/1/12 0:00	ドンキでめちゃ安く買いました！皆さんか	02 バラエティショップ	0	0	0
36	33	PHY	44	2010/1/11 0:00	粉吹くほど乾燥しすぎてビリビリしてい	01 スーパー・ドラッグ	1	0	0

図 2 STM に対応させた CSV ファイルの出力例

- 1 行目：属性名
- 2 行目：データ種別

4 行目以降に入力されるデータの形式を表す。

① AN 数値属性

回答者の年齢などの数値データ。

Vol.2010-NL-196 No.4
Vol.2010-SLP-81 No.4
2010/5/27

② AI 投稿者属性
意味解析を行う必要のない文字列。
③ QF 自由記述質問
自由記述形式の回答データ。
④ AD 日時属性
投稿日時などの日時データ。形式は下記の例に限定する。
形式：YYYY/MM/DD HH:MM
例) 2009/12/13 23:45
⑤ AS 選択属性
選択式の回答データ。
形式：選択肢番号.選択肢内容
例) 03.混合肌
⑥ QC チェック質問
チェックボックスを用いた質問に対する回答データ
形式：0 または 1

- 3 行目：回答条件指定

選択形式の質問で、ある選択肢を選んだ場合のみ回答する質問などを指定する。ただし、構造化された質問はレビューサイトにはないので SECRET では常にハイフンが出力される。

- 4 行目以降：回答データ

回答者の属性や質問に対する回答の実際のデータが入力されている。

3. レビューの繰り返しパターン発見

3.1 サンプリング用 HTML の取得

レビューサイトそれぞれにおいて、レビューのパターンを解析するためにサンプリング用の HTML を取得する。

3.1.1 文字コード

Web ページは様々な文字コードで書かれており、レビュー内容を、文字化けを起こさずに取得するには文字コードを正しく判定することが不可欠である。幸いにもほとんどのレビューサイトの HTML では<META>タグ中で文字コードを指定している。そこで SECRET では、まず文字コードを指定せずに Web から HTML を取得し、その HTML から<META>タグ中に書かれている文字コードを取得する。そして取得した文字コードを用いて再度 Web から HTML を取得することで文字化けの問題を解決している。

まれに<META>タグ中に文字コードが指定されていないサイトがあるが、その場合は Web サイトにて最も利用頻度が高いと思われるシフト JIS を用いる。

3.1.2 HTML の簡略化

Web から取得した HTML をサフィックス木作成やマッチングのために簡略化する。その際、変換するタグは HTML の構造を表すものに限定し、その他のタグは無視する。タグのオプションやパラメータはすべて削除する。さらに、Web ブラウザ上に実際に表示される文字列は<TEXT>に、HTML 中のコメントは<COMMENT>にそれぞれ置き換え、内容は別に保管しておく。

- 取得した HTML

```
<table class="tblEstimate mTop5" cellpadding="0" cellspacing="0"
width="232"><tbody><tr><th>デザイン</th><td><img
src="http://img.kakaku.com/images/itemview/item/icon_star_c_5.gif" alt="" height="11"
width="55">5</td></tr></table>
```

- 簡略化後の HTML

```
<TABLE><TR><TH><TEXT></TH><TD><TEXT></TD></TR></TABLE>
```

この際に用いるタグは HTML の構造を指定するタグに限定しているが、<FONT>や<B>などその他のタグも用いるように設定することが可能になっている。

3.1.3 次ページへのリンクの取得

レビューを多数持つ商品のレビューページは、一定のレビュー数ずつ複数のページに分けられている。そのため、各商品のレビューページのトップページの HTML 内から次ページへのリンクを自動的に取得することが必要である。

まず各商品のレビューページのトップページの HTML 内から<a>要素(アンカー要素)を用いて記述されたリンクをすべて検索する。その中から<a>要素内に次ページへのリンクであることを連想させる文字列(次、next 等)を含むということを条件として、さらに絞り込みを行う。

例：Web ブラウザ上に表示される部分にマッチする部分がある場合

```
<a href="/product/review/product_id/4691/offset/10" class="commonlnk">次のページ≫
</a>
```

例：<img>タグの中にマッチする部分がある場合

```
<a
href="./?c=ld_top_sb_2&amp;flag=1&amp;k=search&amp;q=a&amp;search_btn=1&amp;sta
rt=10"><img src="/img/cmn/pager_arrow_after.gif" alt="Next" width="12" height="23"
/></a>
```

次に、絞り込まれた<a>要素からリンクの部分のみを抽出し、絶対パスに直して次ページのリンクとする。同一ページに次や next を含むリンクが複数ある場合も考えられが、この場合、最初に入力として与えた URI と最も一致する部分が長いリンクを次ページへのリンクとする。

Vol.2010-NL-196 No.4
Vol.2010-SLP-81 No.4
2010/5/27

3.2 サフィックス木の生成による繰り返しパターンの抽出

3.2.1 サフィックス木

サフィックス木とは、以下の図 3 のように、与えられた文字列の接尾部を木構造で表すデータ構造であり、多くの文字列操作の高度な実装に利用されている。また、最長共通部分文字列問題の線形な解法の一つでもある。

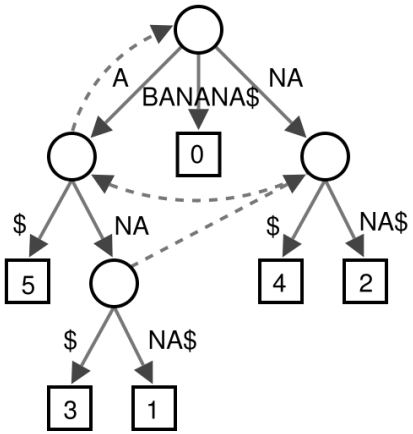

図 3 サフィックス木の例(http://ja.wikipedia.org)

3.2.2 繰り返しパターンの発見

サフィックス木を用いることで繰り返しパターンの発見をする。すなわち、簡略化した HTML を先頭から順に 1 つずつ削りサフィックス木に挿入していく。最終的に節が持っている葉の数（その節よりも下にあるすべての葉）は、その節までの辺がもつタグが HTML 中で繰り返された回数を表す。以下でタグの入力例<a><b><c><c><b><c>を用いて繰り返しパターンの発見方法を図 4 を用いて簡単に説明する。

例：入力<a><b><c><c><b><c>

① タグの並びを先頭から 1 つずつ削り、順に挿入していく。

<a><b><c><c><b><c>
<b><c><c><b><c>
<c><c><b><c>
<c><b><c>
<b><c>
<c>

挿入が完了すると以下の図 4 のようなサフィックス木が作られる。

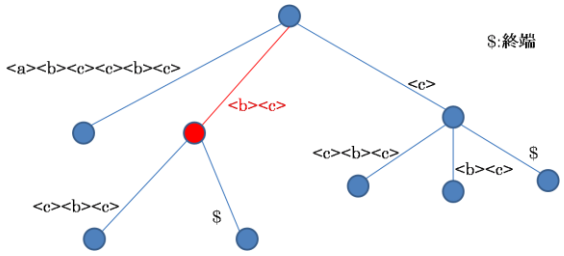

図 4 サフィックス木における繰り返しパターンの発見

② 完成したサフィックス木において葉を複数個持つ節に注目する(例：赤で示されている節)
③ 葉の数をカウントする→2 個
④ 注目した節までの辺がもつタグ→<b><c>

この場合、入力したタグ中に<b><c>が 2 回繰り返されていることがわかる。実際のプログラムにおいてはサフィックス木の構造を単純にする目的で、タグをすべて 0 と 1 からなるビット列に置き換えた後にサフィックス木を作成している。そのため、作成されるサフィックス木の節はすべて左右の計 2 本の辺を持つようになる。サフィックス木からパターンを抽出する際にビット列をタグに変換し直している。

3.2.3 繰り返しパターンの集約

サンプリングに用いたページの数だけサフィックス木を作成するので、同じパターンが複数とれる場合がある。同じパターンを複数個持つ必要性はないので、パターンが同じものは 1 つに集約する。

サフィックス木からすべての繰り返しを抽出するとその数が膨大になってしまい、この後の処理が膨らんでしまうので、この時点で緩い条件で絞り込みを行う。具体的には、繰り返されているタグの数が 5 個以上、それに含まれている<TEXT>が 3 個以上などの条件を満たさないパターンはこの時点で削除する。

3.3 繰り返しパターンと HTML のマッチング

3.3.1 マッチング

サフィックス木より発見した繰り返しパターンとサンプリング用の HTML をマッチングし、その結果をもとに人手で最適な繰り返しパターンを決定する。マッチングを行うことで、そのパターンがページ中のどの部分に何回出現するかがわかる。これらの情報を基に不適切と思われるパターンを除外しパターンを絞り込む。絞り込みに使用する条件は、パターンの出現回数、パターン中のタグの数、<TEXT>が含まれる数、パターンの出現位置から計算される密度である。これらの条件は SECRET のフォームから設定を変更することができる。

Vol.2010-NL-196 No.4
Vol.2010-SLP-81 No.4
2010/5/27

3.3.2 最適な繰り返しパターンの決定

多くのサイトでは、上記の絞り込みを行った時点ではパターンが 1 つに定まらないので、複数のパターンから最適なパターンを人手で選択する必要がある。

最適な繰り返しパターンの選択をサポートするために、取得した各パターンの以下の要素を評価し、各パターンに得点を与える。

- レビュー本文が含まれるか
  パターン中の<TEXT>にレビュー本文と判断できるものがあるか。
- ユーザ名が含まれるか
  パターン中の<TEXT>にユーザ名と判断できるものがあるか。
- 繰り返しパターンの HTML が構造化されているか
  適切なパターンであればレビュー1 件に相当する部分の HTML は構造化されている。<TABLE>や<TR>などのタグに絞って構造化の程度を判定し、その構造化の度合いに応じて加点をする。
- 繰り返しパターンの出現位置の密度が一定以上あるか
  パターンの出現位置から計算される密度が一定以上あるか。
- 繰り返しパターンがページ中に一定以上出現するか
  パターンのページ中の出現回数の平均が一定以上あるか

上記の 5 点に対して評価を行い、ユーザが最適なパターンを選択しやすいように得点の降順に並び変え表示させる。

4. レビューの表形式化

4.1 HTML の取得

レビューサイトから HTML を取得し、最適な繰り返しパターンとのマッチングを行うために、取得した HTML をサンプリング用の HTML と同様に簡略化する。

4.2 最適な繰り返しパターンを用いたレビュー内容の取得

最適な繰り返しパターンと簡略化された HTML のマッチングを行い、一致した箇所にある<TEXT>と<COMMENT>からレビュー内容を取得する。

4.3 属性名と属性値の決定

表形式化にあたって、レビューを解析しレビューを構成している属性名と属性値の決定を行う。まず HTML から取得したレビュー内容より属性名の候補を選出する。レビューの特性として、レビューごとの同じ箇所で常に同じ単語が繰り返されているものが属性名である可能性が高く、逆にレビューごとに異なる単語は属性値となる可能性が高い。これを基準に属性名の候補を選出する。

4.3.1 意味解析に基づく属性名の抽出

属性名の候補の中には、レビューから得たい情報には適さないものも含まれる。そこで意味解析システム SAGE を利用し、属性名候補を意味解析し、その主辞が以下の 3 点の条件を満たすか判断する。

① 名詞節あるいは断定節である
② 「断定、現実、現象描写」のモダリティをもつ
③ 特定の概念を上位概念にもつ

これらの条件を満たしたものを属性名とし、満たさなかったものは破棄する。

まず属性名候補を取得したレビュー内容から抽出する。すべてのレビューにおいて同じ言葉が繰り返されている<TEXT>を選択する。ただしその<TEXT>が記号のみ、もしくは 15 文字以上ある場合は、属性名としては不適切だと判断して破棄する。選択した<TEXT>を意味解析にかけ、上述の条件を満たした場合、その<TEXT>を属性名として用いる。ただし、投稿日時やレビュー本文など、属性値からその属性の属性名が決定できる場合は、そちらを優先する。なお、下記の 9 個の概念 ID を用いている。

- ID:3aa938 場所
- ID:30f7c8 物事に対する評価
- ID:3f9871 性状・性向
- ID:3aa943 行動に関わるもの
- ID:444dd9 ものを対象とする行為
- ID:4444c4 静物
- ID:4444e1 自然現象によってできる物
- ID:444800 行為における役割で捉えた人間
- ID:4448db 書いた情報

4.3.2 属性名と属性値の対応の決定

属性値として用いられうるものは属性名候補とされなかった<TEXT>である。ほとんどの場合、属性名と判断された<TEXT>には、その後に属性値と判断された<TEXT>が存在する。一方、<TEXT>を属性値と決定した際にその属性名が決まる（日時やユーザ名など）ため、属性名と属性値の対応の決定は容易である。

しかし、一部のレビューサイトでは、レビュー部分に<TABLE>タグを用いて表形式でデータを掲載している。その場合、属性名の<TEXT>が連続したり、属性値の<TEXT>が連続したりすることがあり、この部分の順番を入れ替える処理が必要となる。この処理は、<TABLE>タグ中の<TD><TH>タグを用いて入れ替えるべき位置を判断し、属性名の<TEXT>の後に適切な属性値の<TEXT>がくるように入れ替えを行う。

4.3.3 STM に対応した CSV 形式での解析結果出力

本研究で扱う属性は STM に対応させるために AI,AD,AN,AS,QF,QC のいずれかのデータ種別に分類する必要がある。このデータ種別の決定には属性値を用いる。また、

Vol.2010-NL-196 No.4
Vol.2010-SLP-81 No.4
2010/5/27

そのデータ種別によっては属性値の各項目も書き換えることが必要となる場合もある。

① AN 数値属性
属性値が常に数値である場合。

② AI 投稿者属性
属性値がユーザ名と判断された場合。

③ QF 自由記述質問
属性値がレビュー本文と判断された場合。

④ AD 日時属性
属性値に常に日時が含まれており、投稿日時と判断された場合。

⑤ AS 選択属性
属性値が複数の値からの選択属性だと判断された場合。

⑥ QC チェック質問
属性値が 2 値で「ある、なし」で表現できる場合。

データ種別が AS となった場合、各属性値の先頭に「01.」「02.」といった選択肢番号を挿入する。また、QC の場合、そのレビューにおいてその属性値がある（質問に対してチェックがついている）場合 1 を、ない場合 0 を属性値として書き換える。日時データである AD はすべて YYYY/MM/DD HH:MM の形式に統一する。

5. SECRET の利用方法

SECRET の利用方法について解説する。図 5 は SECRET のメインフォームである。レビューページの URI を入力し、パターン取得ボタンを押すと図 5 の①のようにレビューパターンの候補が表示される。Score は 3.3.2 で述べた方法で評価した得点であり、この得点の降順で表示される。出現回数はサンプリングで使用した各ページ中にそのパターンが出現する回数を表す。

①でパターンを選択すると、②にそのパターンで取得できるレビュー内容のサンプルが表示される。さらに、ブラウザ表示ボタンを押すと、図 6 のようにそのパターンがレビューページ中のどの部分にあたるのかが視覚的に表示される。図 6 の③がそのパターンの開始位置で、④が終了位置である。図 6 ではレビュー1 件がパターンの開始位置と終了位置にちょうど挟まれているので、選択したパターンが最適な繰り返しパターンだと判断できる。

最適なパターンを選択した状態でレビュー取得ボタンを押すと、レビューサイトからレビューを取得し、STM 用の CSV 形式で出力する。

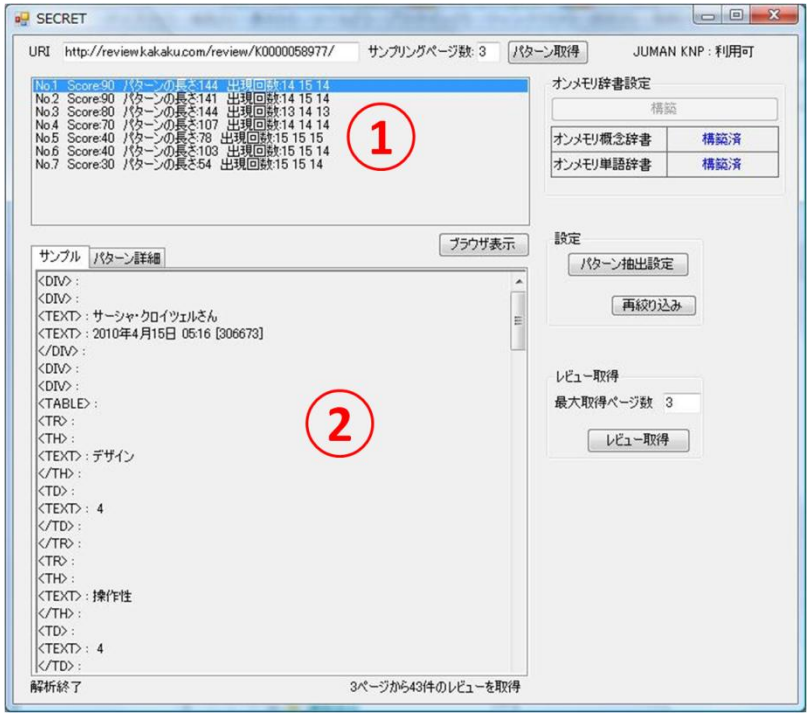

図 5 SECRET のメインフォーム

Vol.2010-NL-196 No.4
Vol.2010-SLP-81 No.4
2010/5/27

図 6 ブラウザ表示の例

6. 評価実験

SECRET の精度を評価するために、Web 上の複数のレビューサイトを対象に実験を行った。実験に用いたサイトは以下の 9 サイトである。

- 価格.com http://kakaku.com/
- Amazon.co.jp http://www.amazon.co.jp/
- @cosme http://www.cosme.net/
- 楽天 http://review.rakuten.co.jp/
- 漫画レビュー.com http://www.manngareview.com/
- PlayStation mk2 http://www.psmk2.net/
- @nifty 温泉 http://onsen.nifty.com/
- ReComic http://recomic.jp/
- Yahoo!ショッピング http://shopping.yahoo.co.jp/

上記の 9 つのサイトから投稿レビュー数が一定数以上ある商品をそれぞれ 10 件選び、SECRET によりレビューを抽出した。90 件の商品から取得した総レビュー数は 11556 件となった。以下がその結果である。

表 1 実験結果

サイト名	レビュー取得率	属性取得率	属性名	データ種別
価格.com(30)	82%	93%	92%	92%
Amazon.co.jp(100)	76%	60%	100%	100%
@cosme(1000)	81%	55%	100%	100%
楽天(60)	49%	72%	100%	100%
漫画レビュー.com(50)	92%	80%	100%	100%
PlayStation mk2(40)	100%	80%	100%	100%
@nifty 温泉(60)	100%	67%	100%	100%
ReCOMIC(40)	83%	43%	100%	100%
Yahoo!ショッピング(40)	96%	96%	100%	100%
平均	84%	72%	99%	99%

サイト名の右の括弧内は各商品から取得を試みたレビュー数である。括弧内の数字以上のレビューが投稿されている商品のみを対象とし実験を行った。属性取得率は、レビューから取得すべきと考えられる属性に対する実際に取得できた属性の件数の割合である。属性名の項目は、その取得した属性において適切な属性名があてられていると判断した割合である。同様にデータ種別の項目は、取得した属性において適切なデータ種別があてられていると判断した割合である。

7. 結論

7.1 評価実験からの考察

本研究ではサフィックス木を用いて Web ページ中の繰り返しパターンを発見することで、多くのサイトから高い取得率でレビューを集めることができた。また、SAGE 解析によって属性名を決定し、STM に対応した CSV ファイル出力をすることができた。

レビューの HTML 上の繰り返しパターンが常に一定なサイト（PlayStation mk2 や @nifty 温泉）においては、商品ごとに 1 つのレビューパターンですべてのレビューが取得でき、高い精度が出せている。しかし、レビューによって属性が増減してレビューのパターンが一定でない楽天サイトにおいては、低いレビュー取得率となってしまっている。これは、サフィックス木の性質上、繰り返しが一定でないものは同一のパ

ターンだと認識できないためである。

データ種別の判定を誤っているものは、レビューに現れるその属性の種類が不足していたためである。例えば、AS として取得した属性があったとしても、取得したレビューにその属性値が 2 値しかなかった場合、その属性が AS だと判定することは非常に困難になり、QC と判定してしまう。

7.2 今後の課題

現在の SECRET では複数の繰り返しパターンがあった場合は対応しきれずに、取得できるレビューが大きく減ってしまうか、そもそも繰り返しパターンが取得できないという問題が起きている。

属性名として使用できる言葉が HTML 中になかった場合、現在は属性値から容易に属性名が決定できる「年齢、ユーザ名、投稿日時、レビュー本文、評価」のみしか属性として採用できないが、それ以外でも属性値からその属性名を推測できるようにすれば属性名の精度を維持したまま属性取得率を上げられると考える。

1 つの<TEXT>に複数の属性がある場合、例えば「プレイ時間：15 時間以上 30 時間未満(クリア済)shiromaru さん [2009/04/18 掲載]」はプレイ時間、クリア状況、ユーザ名、投稿日時の 4 属性がひとつの<TEXT>にまとまっている。このような場合、テキストを適切に分割し、複数の独立した属性として扱う処理が必要となる。

8. 参考文献

1) 原田実，田淵和幸，大野博之，"日本語意味解析システム SAGE の高速化・高精度化とコーパスによる精度評価"，情報処理学会論文誌，Vol.43，No.9，pp.2894-2902，(2002.9)
2) 原田実,水野 高宏: “EDR を用いた日本語意味解析システム SAGE "，人工知能学会論文誌, Vol.16, No.1, pp.85-93 (2001.1).
3) 梅澤俊之，西尾華織，松田源立，原田実: “意味解析システム SAGE の精度向上とモダリティの付与と辞書更新支援系の開発”，言語処理学会第 14 回年次大会発表論文集, E3-1 ,pp. 548-551 (2008.3).
4) 西脇 剛,保立哲志,原田実: “意味解析に基づくテキストマイニングシステム STM”,情報処理学会第 69 回全国大会論文集,2C-03,第 2 分冊 pp. 89-90. (2007.3).
5) Chia-Hui Chang , Chun-Nan Hsu , Shao-Cheng Lui, Automatic information extraction from semi-structured Web pages by pattern discovery, Decision Support Systems, v.35 n.1, p.129-147, 01 April 2003
6) Chia-Hui Chang, Shao-Cheng Lui, IEPAD: information extraction based on pattern discovery, Proceedings of the 10th International Conference on World Wide Web, Hong-Kong, Lecture Notes in Artificial Intelligence vol. 2336, Springer, 2001, pp.223– 231.
7) Algorithms with Python / トライ，パトリシア
http://www.geocities.jp/m_hiroi/light/pyalgo09.html
8) 南野朋之 齋藤豪 奥村学: 繰り返し構造を用いた Web ページの構造化に関する研究，情報処理学会研究報告．自然言語処理研究会報告 IPSJ SIG Notes 2003(23) pp.185-192 20030306