大規模ウェブアーカイブを用いた社会分析の試み

鍛治伸裕　　豊田正史　　喜連川優

東京大学 生産技術研究所
{kaji, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

1　はじめに

近年，ウェブは我々の生活に密着したものとなり，あらゆることがウェブに書き込まれるようになってきている．例えば政治に対する意見から趣味の話題にいたるまで，様々な内容の書き込みが見られる．ブログなどのツールの普及も手伝って，今後この傾向はますます強まるものと思われる．

こうした背景から，我々は大規模なウェブアーカイブの解析を通して，現実社会における人々の意見や行動を分析することを試みている．ウェブアーカイブの解析に必要となる自然言語処理技術は，評価情報分析のような比較的応用よりの技術から，形態素解析や多義性解消といった基礎的な技術まで多岐に渡る．本論文では，ウェブアーカイブを用いた社会分析における我々の取り組みの一部を紹介する．

2　ウェブを通した社会分析

我々の研究室では 1999 年から継続的に日本語ウェブページの収集を行っている．2007 年 1 月現在までに累計で 10 億件以上のウェブページを蓄積したアーカイブが構築されている．ウェブページは取得した時間情報を保持して蓄えられているので，過去に存在したウェブページを閲覧することや，ウェブページの時間的変遷を観察することも可能である．このウェブアーカイブが我々の分析対象である．

この大規模なウェブアーカイブ全体を一度に解析することは，コストが高いため現実的ではない．そこで，現在のところは，ある検索語 (人名や商品名など) について記述しているウェブページだけをアーカイブから取り出してきて，その検索語に対する関心度や意見内容を解析するという問題設定で議論をしている．分析に使う検索語は，マーケティングの専門家らと議論をしながら，人々の関心や意見が反映されやすそうな語を随時選定している [4].

具体的な分析内容としては，まず検索語に関する話題の広がりを分析するために，ウェブページ数の時間的推移やリンク関係の解析を行っている．また，意見内容を分析するために，検索語が出現するテキストの解析も行っている．自然言語処理技術としては，まず評判情報分析技術 (Sentiment Analysis) が課題となる．これは後者の分析を行うためには必須である．さらに，両分析に共通する技術として，検索語の多義性解消や形態素解析の高度化などの基礎解析技術にも取り組んでいる．

大規模ウェブアーカイブを対象した自然言語処理の面白さは，ウェブアーカイブが分析対象であると同時に，知識獲得の対象でもあるという点である．近年では億単位の大規模なウェブページを使える環境が整備されつつあり，コーパス (ウェブアーカイブ) からの知識獲得が現実的なものになってきたと考えている．そのため，ウェブアーカイブの解析に必要な知識 (または機械学習に必要な訓練事例) を，ウェブアーカイブから獲得するという手法を模索している．

3　評判情報分析のフレームワーク

評判情報分析のための基盤的なリソースをウェブアーカイブから自動構築し，それを用いて評判情報抽出を行うというフレームワークを提案している．これまでに，我々は評価文コーパスと評価表現辞書を自動構築する手法を提案した [1, 7]．現在は，その評価表現辞書を用いて各種実験を進めている．

3.1　評価文コーパスの自動構築

リソース整備の第一歩として，ウェブページから評価文コーパスを自動抽出する手法を考案し，大規模な評価文コーパスを構築した．基本的アイデアは，テキ

ストとレイアウトの構造的手がかりを利用するというものである．

レイアウト構造は箇条書き形式と表形式の 2 種類を利用した．例えば，図 1 のような箇条書きは「良い点」「悪い点」という見出しを持っているため，箇条書きに評価文が記述されていることを判定できる．本論文では「良い点」「悪い点」のような，評価文の存在を示唆する表現を手がかり表現と呼ぶ．

良い点
- 変に加工しない素直な音を出す．
- 曲の検索が簡単にできる．
- お気に入りのプレイリストを作って楽しめる．

悪い点
- リモコンに液晶表示がない．
- ボディに傷や指紋がつきやすい．
- ライトを点灯し続けると直ぐに電池がなくなる．

図 1: 箇条書き形式で記述された評価文

表形式も箇条書き形式の場合とほぼ同様である (図 2)．表の 1 列目に手がかり表現 (気に入った点，イヤな点) が存在していて，これが見出しの働きをしている．そして 2 列目には評価文が記述されている．

燃費 (市街地)	7.0km/litter
燃費 (高速)	9.0km/litter
満足度	95%
気に入った点	４ドアなのにカッコよすぎる．
イヤな点	シートがぼろくライトが暗い，色がはげてきてる．

図 2: 表形式で記述された評価文

次に，定型的なテキスト構造に着目した．

(1) これの 良いところは 計算が速い ことだ．
(2) 慣れるまで時間がかかる ところが，悪い点だ．

それぞれの例文には「計算が速い」「慣れるまで時間がかかる」という評価文が含まれている[1]．いずれも「良いところは〜こと」「〜ところが悪い点」といった定型的なテキスト構造を使って記述されているので，単純な語彙統語パターンを用いて自動抽出できる．

上記の手法を，ウェブアーカイブ中の約 10 億件のウェブページに対して適用したところ，約 50 万文の評価文を抽出することができた[2]．抽出された評価文の例を表 1 に示す．

表 1: 評価文の例

評価極性	評価文
好評	順応性が素晴らしくある．
	使い方がわかりやすい．
	何と言っても，料金が良心的だ．
不評	費用が高い．
	いい加減な意見，ふざけた意見などが出てくる．
	エンジンが非力で少々うるさい．

3.2 評価表現辞書の自動構築

次に，この評価文コーパスから評価表現辞書を自動構築した．形容詞/形容詞句の出現頻度が好評文 (または不評文) にどの程度偏っているかを Pointwise Mutual Information を用いて定式化することによって，評価表現の自動獲得を行った．その結果，表 2 に示すような評価表現を辞書に登録することができた [7]．

表 2: 評価表現の具体例

好評表現	不評表現
謙虚だ	ダサい
支障が無い	厄介だ
エキサイティングだ	消耗が早い
漏れが少ない	魅力が無い
能力が高い	しょぼい

評価文コーパスから構築した評価表現辞書を用いて，再帰的に評価文コーパスを解析すると，興味深いデータが得られる．例えば，下記の例文はいずれも不評文として評価文コーパスに登録されている．

(3) a. まだまだ 魅力的な商品 が少ない
    b. 美味しいお店 が少ない

自動構築した辞書には「魅力的だ」「美味しい」が好評表現として登録されている．そのため「魅力的な商品」「美味しいお店」は，いずれも《望ましい物》であると解釈することができる．このことを利用すると「《望ましい物》が少ない」ことは不評であるといった知識を獲得することができる．これは句の汎化を行っていることに相当する．

3.3 評判情報抽出への応用

評価表現辞書を使えば，簡単ではあるが評判情報抽出を行うことができる．検索語の評判は「《検索語》の

[1] 厳密には文ではなく節と呼ぶべきだが，レイアウト構造を用いて抽出される評価文との整合性を考えて文と呼ぶ．
[2] http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp/

《属性》は《評価表現》」という三つ組で記述されることが多い．これに着目して，検索語の評判を抽出するため，そのような抽出パターンを人手で作成した．

抽出パターンを用いて，実際にウェブアーカイブから抽出された評判情報の例を以下に示す．

(4) a. Aのサラダは，どれも 食べやすい です．
b. はっきり言ってNの番組制作 能力は高い．
(5) a. Gの店員は 愛想がない ！
b. K内閣の答弁が 無茶苦茶 です．

(3) が好評情報で (4) が不評情報である．評価表現部分には下線部を引いている．なお，検索語に相当する部分はアルファベットに変換している．

もちろん，このような方法にはまだ改善の余地がある．例えば，実際のウェブテキストでは省略が頻繁に行われるため，上記のような抽出パターンを使うだけでは高い再現率を得ることが難しい．そのため，省略解析の適用などが今後の課題であると考えている．

4 基礎解析技術

基礎的な解析技術の高度化にも取り組んでいる．ここでは検索語の多義性解消と，形態素解析のための未知語獲得に関する取り組みを紹介する．いずれも機械学習を用いているが，訓練事例を人手で作成するのではなく，ウェブアーカイブから自動構築するというアプローチをとっている．

検索語の多義性解消 検索語が多義である場合，その多義性解消は重要な課題となる．例えば，シャンプーの「ラッシュ」に関する評判を分析しようとしたとき，単純に「ラッシュ」を検索語とするのでは「通勤ラッシュ」などが検索結果に混入してしまう．

これまでの多義性解消処理は，機械学習にもとづく手法が主流となっている．しかし，検索語ごとに人手で訓練事例を作成するのは現実的ではない．そこで，ウェブアーカイブから擬似的な訓練事例を自動構築することを試みた．例えば「ラッシュ」の場合であれば「シャンプー」や「髪」のような関連語と共起しているテキストを擬似的な正例として活用することが考えられる．精度はまだ十分に調査していないが，予備実験の結果おおむね良好な結果が得られている．擬似的な訓練事例を用いる方法以外にも，ウィキペディアのような大規模な語彙資源を活用する方法も検討中である．

未知語の獲得 ウェブには，話し言葉的なくだけた文体のテキストが多く見られる．従来の形態素解析器は，新聞記事のような書き言葉を前提として開発されているため，くだけたテキストを高い精度で解析できない．

評判情報分析を行うことを考えると，評価表現の多くは形容詞であることから，形容詞の解析精度の向上が重要となる．特に我々は「ンマい」「ウザい」といったカタカナ形容詞に着目した．この種の用言は，現在の一般的な形態素解析辞書に登録されていないため，正しく解析することができない．

そこで，カタカナ用言の自動獲得に取り組んでいる．ウェブアーカイブからカタカナ文字列とそれに後続する文字列を抽出し，その後続文字列を素性として機械学習を適用する手法を試している [5]．この場合も，訓練事例はウェブアーカイブから自動構築している．

5 リンク解析との融合

ここまでは，ウェブを大規模な文書集合として見てきたが，一方でウェブはハイパーリンクで結合された巨大な文書のネットワークとしての側面も持っている．ウェブ上では，互いに関連のある情報がリンクで密に結合される傾向があるため，密なページ間のリンク構造を抽出することでウェブ上のトピックを抽出し，その時系列的な変化を観測することが可能となる．我々はこれまでに，リンク解析を用いたウェブ上の全トピック抽出，その時系列的な変化の追跡，および追跡を容易にする可視化手法について研究を行ってきた [2, 3]．

リンク解析を用いることで，ウェブにおける話題の概要を把握することが可能になるが，より詳細な話題の分析には自然言語処理との融合が不可欠となる．例えば，リンク解析においてはリンクで結合されたページ同士が関連を持つことを仮定するが，テキストの類似度を用いることで解析上ノイズとなるリンクを排除し，重要なリンクの重みを増すことで，より高精度な話題の抽出が可能となる．また，リンク解析ではリンクで結合されていない文書を取り扱うことが出来ないが，リンク解析で抽出された話題に，テキストの類似度を用いてリンクのない文書を補完すれば解析対象を拡張することが可能になる．リンク解析と自然原書処理の融合については，既にいくつかの予備実験を行っているが，今後詳細な解析を行う予定である．

また評判情報分析など，詳細な自然言語処理の結果をどのように分析者に提示するかも重要な研究課題となる．ウェブにおける話題の時間的変化は，それに関

連するページとリンクからなる動的なグラフと考えることができる．我々は動的グラフを見やすく，対話的に操作可能な状態で提示する可視化技術の開発にも取り組んでいる [6]．図 3 は「生協の白石さん」について記述しているウェブページを検索し，ページ間のリンク関係の時系列変化を可視化したものである．時間を追うにしたがって，リンクが増加する様子が分かる．また，流行の火付け役となったウェブページも簡単に見てとることが出来る (一番下のスナップショットの右側中央)．今後，このような話題の地図上に評判情報をマッピングして可視化するなど，自然言語処理との融合を行い，マーケティング研究者や社会学者などのエンドユーザにとっても理解しやすい分析結果の提示手法を検討していく予定である．

6　おわりに

本論文では，ウェブアーカイブを用いた社会分析における我々の取り組みを紹介した．現在，分析結果のマーケティングへの応用を検討しているところであり，これまでに一定の成果を得ている [4]．今後は，マーケティング応用からのフィードバックを得ながら，研究を進めていく予定である．

参考文献

[1] Nobuhiro Kaji and Masaru Kitsuregawa. Automatic construction of polarity-tagged corpus from html documents. In *Proceedings of COLING/ACL, Poster Sessions*, pp. 452–459, 2006.

[2] Masashi Toyoda and Masaru Kitsuregawa. Web Community Chart: a Tool for Navigating the Web and Observing its Evolution. *IEICE Transactions on Information and Systems*, Vol. E86-D, No. 6, pp. 1024–1031, June 2003.

[3] Masashi Toyoda and Masaru Kitsuregawa. A System for Visualizing and Analizing the Evolution of the Web with a Time Series of Graphs. In *Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia (Hypertext 05)*, pp. 151–160, September 2005.

[4] 馬渡一浩, 富田英裕, 新井範子, 豊田正史, 鍛治伸裕, 喜連川優. ブログからレピュテーション分析の

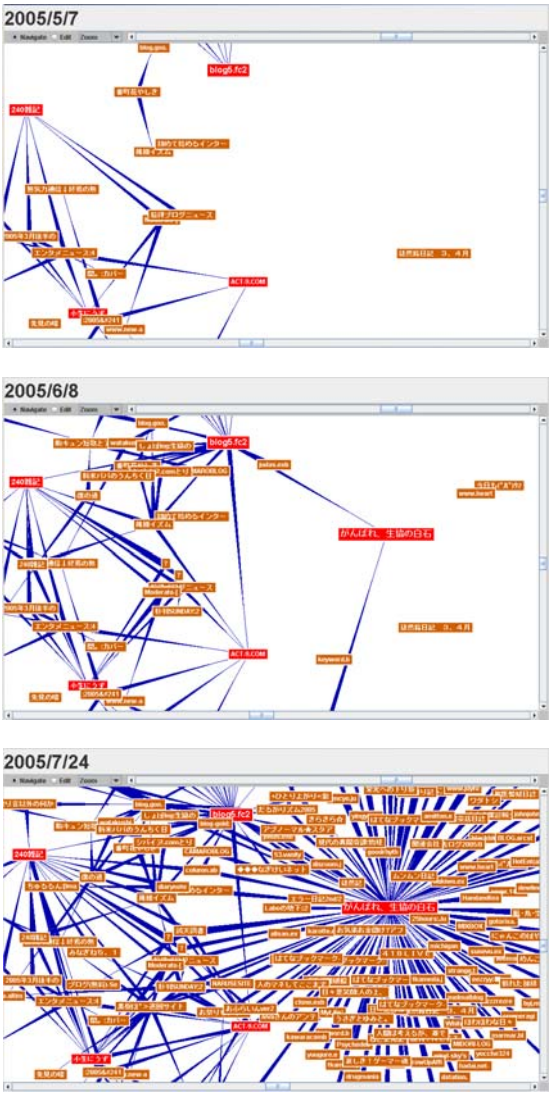

図 3: 「生協の白石さん」に関するウェブページとリンクの時系列変化

可能性を探る－web2.0 時代の新たな方法論へのトライアル－. 日本広報学会第 12 回研究発表大会予稿集, pp. 60–63, 2006.

[5] 福島健一, 鍛治伸裕, 喜連川優. 機械学習を用いたカタカナ用言の獲得. 言語処理学会第 13 回年次大会, 2007.

[6] 豊田正史. インタラクティブな動的グラフレイアウト手法を用いたウェブグラフ発展過程の可視化. In *Proceedings of WISS*, 2006.

[7] 鍛治伸裕, 喜連川優. HTML 文書からの評価表現辞書の自動構築. 言語処理学会第 13 回年次大会, 2007.