D-036

Web情報信頼性分析のためのページ外観情報の抽出
Extraction of Appearance Information for Web Information Credibility Analysis

赤峯 享† 加藤 義清† 乾 健太郎† 黒橋 禎夫†‡
Susumu Akamine Yoshikiyo Kato Kentaro Inui Sadao Kurohashi

1. はじめに

筆者らは，健康食品や時事問題など情報の信頼性が問題となるトピックや命題に対して，Web 上の様々な発信者の情報を多面的に分析し，Web 情報の全体像を様々な観点から集約して提示する Web 情報分析システム WISDOM を開発している[1]．WISDOM では，信頼性評価の観点として，(a)情報発信者，(b)内容(意見や主要表現)，(c)ページ外観という３点を中心に開発を進めている．本稿では，その中のページ外観について，抽出の目的，WISDOM での実装イメージ，抽出対象と抽出手法の方針，及び，予備実験について報告する．

2. ページ外観情報の抽出の目的

人は，Web ページの内容を深く理解しなくても，外観的な特徴を把握するだけで，そのページが有用そうか，そのページが信頼できそうか，そのページの主要部分はどこか等を，ある程度，判別することが可能である．例えば，健康食品の効果についての意見を調べたい場合，写真とキーワードだけで文の少ない商品販売ページやリンク集ページはその内容を深く確認するまでもなく，不用であると見なしてスキップする．また，住所などの実世界の連絡先が記述されている Web ページは信頼度が増加し，広告が多い Web ページは信頼度が減少するなどの評価を行っている[2]．

ページ外観抽出は，上記を機械的に処理することで，以下を実現することを目的としている．

- 情報信頼性検証のための外観分析
  Web ページ中に，電話番号や住所等の実世界の連絡先が含まれるか，アフィリエイト広告が過剰に含まれるか等，Web ページの信頼性に関連する外観的な特徴を抽出し，利用者に提示する．

- 情報分析の不用ページ削除(ページタイプ分類)
  外観的な特徴を元に，検索結果，商品販売，リンク集などのページタイプ分類を行い，情報分析には雑音となる不用なページタイプを検索結果から削除する．

- 外観から判別可能な Web の形式的情報の付与
  ブログや掲示板などの外観的特徴があるページを判別したり，ページ中のヘッダ，フッタ，本文，広告部分など抽出したりことで，発信者分析や内容分析の基盤となる形式的情報を付与する．

情報分析システムの処理において，ページ外観の利用箇所を図１に示す．

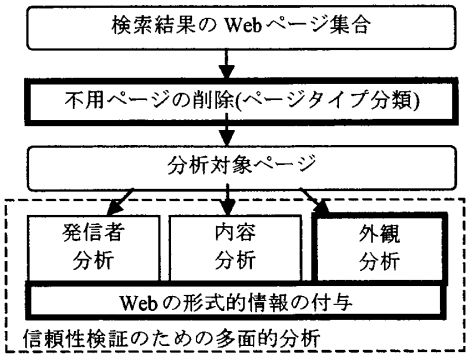

図１ Web情報分析における外観抽出の利用

3. Web情報分析システムでの実現例

Web 情報分析システム WISDOM での信頼性検証のための外観分析の実現例を図２に示す．外観情報は，「実世界の連絡先(住所・電話番号)があるか」，「広告が多いか」，「プライバシーポリシーが記載されているか」等の Web ページの信頼性の手掛かりを利用者に与える情報がアイコンで表示される．

4. 抽出対象とページタイプ分類

今回，外観抽出としては，プロトタイプシステム[3]での検討結果を元に，自動処理の実現性を加味して，「連絡先」，「広告」などを抽出対象とした．また，ページタイプ分類としては，情報分析で雑音となる「商品販売」や「リンク集」などを分類対象とした．抽出対象とページタイプ分類として選択したものを図３に示す．

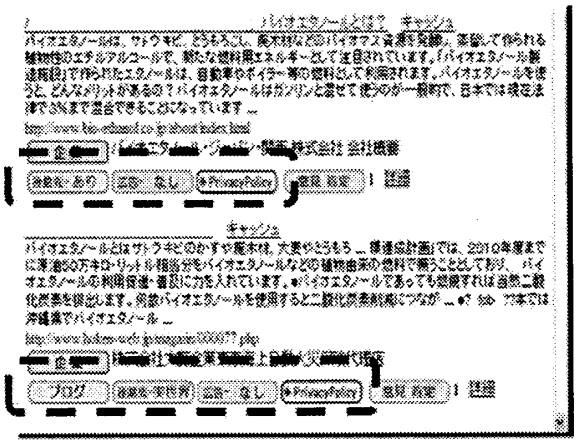

図２ Web情報分析システムでの実現例

† 情報通信研究機構, NICT
‡ 京都大学, Kyoto University

信頼性検証のための外観抽出対象	連絡先(住所，電話番号，電子メール，問合せ先)
	広告数
	プライバシーポリシーの有無
	文体（口語表現）
基盤の抽出対象	文数(名詞句・文），画像の数(サイズ大・サイズ小），リンク数(サイト内・サイト外），日付表現
ページタイプ(不要)	検索結果，商品販売，写真集 リンク集，メニュー
ページタイプ(基盤)	ブログ，掲示板

図 3　抽出対象とページタイプ分類

5. 外観情報の抽出手法の方針

外観情報の抽出は，HTML ファイルをパーズし，タグ情報，テキスト情報，リンク情報・先頭からのオフセットを取り出し，以下の方針で外観情報を抽出する．

- 出現位置や共起文字列等の複合的な利用

Web ページ中で，連絡先の住所，電話番号や，プライバシーポリシーの記述などは，ページの中央部ではなく，フッタ部分に多く出現する．また，本文中に住所の表記があらわれた場合は，必ずしも Web ページの発信者の連絡先ではない可能性が高い．従って，Web ページ中の出現位置を考慮した抽出を行う．

また，例えば，電話番号は，「連絡先」，「TEL」などの特定の文字列や住所とも共起しやすいので，共起情報も利用した複合的な条件を用いて抽出を行う．

- ブートストラップ的なパターンの獲得

個々の要素の抽出は，予め辞書として蓄えた特定のパターン（正規表現など）にマッチするかどうかで行うが，最初に全てのパターンを人手で作成することは困難である．したがって，以下の方法でブートストラップ的に，2-4 を繰り返すことで，辞書を拡大する．

1. 種となる初期パターンを辞書に登録する．
2. 辞書中のパターンを用いて実際の Web ページからパターンにマッチする箇所とそのパターンに共起するパターン（文字列）を見つける．
3. 共起パターンを用いて，Web ページから新規パターンの候補を獲得する．
4. 新規パターンの候補を人手でチェックし，辞書に登録する．

例えば，「プライバシーポリシー」については，実際の Web ページでは，「個人情報保護」，「個人情報規定」など様々な文字列で表現される．これらの表現は Web ページの後方（フッタ部分）に，「会社案内」，「会社概要」，「サイトマップ」，「お問い合わせ」などの文字列と共起して現れることが多い．この共起文字列を用いて，「プライバシーポリシー」を表す新規の文字列を獲得する．

- 外観抽出とページタイプ分類の融合

(a)「連絡先」，「サイト内／外へのリンク数」，「日付表現」などの個々の外観情報の抽出結果を総合的に判断することで，ページタイプ分類を求めるボトプアップ的な処理と，(b)ページタイプ分類を仮定することで，個々の外観情報の抽出していくトップダウン的な処理を融合した手法を用いる．

6. 予備実験

「アガリクス」，「バイオエタノール」，「ジェネリック医薬品」，「裁判員制度」などの健康・社会問題に関する 50 トピックに対して，検索エンジン TSUBAKI[4] の検索結果の上位 1000 ページを評価対象ページとして，外観情報抽出を抽出して，予備評価を行った．

出現位置の利用については，出現位置を考慮せずに，連絡先の住所を抽出した結果の 50 ページを確認したところ，連絡先でない住所が 20 ページ存在した．タグ数と文字数の割合が，80%以降(フッタ)の箇所に絞り込むことで，連絡先でない住所の 11 ページが抽出対象から除かれることが確認できた．しかしながら，連絡先自体が 1 ページになっているものについては，ページ中央部に住所があり，出現位置の利用により副作用が生じる例もあった．また，ヘッダとフッタの位置とタグ数と文字数の 80%以降という条件のズレによる誤りも発生した．今後，精度を上げるためには出現位置だけでなく，共起文字列の情報や，住所が現れるブロックの種別の情報などを複合的に用いる必要があると考えられる．

パターンの獲得については，「プライバシー」，「個人情報保護」を元にブートストラップ的な手法で，「privacy policy」，「個人情報の取扱い」，「個人情報保護管理」，「プライバシー規約」，「プライバシー保護方針」などの多様な表現を獲得できることを確認できた．

また，予備実験で抽出したページ外観の抽出結果の「連絡先」，「広告数」，「プライバシーポリシー」，「文体」とページタイプの「ブログ」については，その結果を WISDOM に実装して，Web 情報分析システム上でその効果を確認中である．

7. おわりに

Web 情報の信頼性検証のためのページ外観情報抽出について，その目的，Web 情報分析システム WISDOM での実現例，抽出方針を報告した．また，50 トピックの検索結果について，予備実験を実施し，抽出手法の有効性を確認した．今後は，統計的手法を導入し，出現位置，共起情報，ページタイプなどを総合的に判断するモデルを用いることで，抽出精度の改善を行う予定である．

参考文献

[1]赤峯享,宮森恒,加藤義清,中川哲治,乾健太郎,黒橋禎夫,木俵豊, “Web 情報の信頼性検証のための情報分析システム WISDOM”, 言語処理学会年次大会,2008

[2]B.J. Fogg et al. “What makes a Web site credible? A report on a large quantitative study”, Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems, Vol. 1, pp.61-68, New York: ACM Press , 2001.

[3]Hisashi Miyamori, Susumu Akamine, Yoshikiyo Kato, Ken Kaneiwa, Kaoru Sumi, Kentaro Inui and Sadao Kurohashi: “Evaluation Data and Prototype System WISDOM for Information Credibility Analysis”, Proceedings of the First International Symposium on Universal Communication, pp. 234-237, 2007.

[4] 検索エンジン基盤 TSUBAKI:http://tsubaki.ixnlp.nii.ac.jp/index.cgi