指定ドメインにおけるWebページ評価手法

2X－02

片山佳則，古川淳子，西野文人
(株)富士通研究所 ドキュメント処理研究部
E-mail: {katayama.yoshin, furukawa.junko, nishino}@jp.fujitsu.com

1．はじめに

企業や個人が、情報発信を行うために、インターネットやイントラネットのWebコンテンツ(ページ)が活用されている。しかし、そのページにあるべき情報が記述されていないために、情報伝達がスムーズに行えない事や、各種の検索エンジンによる検索結果から漏れてしまう事などが起きている。これらは、企業にとって、ビジネスの機会を逃してしまうという重大な問題につながる可能性がある。従って、Webページをコンテンツの充実度の観点で評価でき、記述内容に関して示唆が行える方法が必要である。

本稿では、これらの目的に合う評価手法として、各業界やコミュニティなどのドメインで、既に情報発信されている発信内容に基づいて、ダイナミックに評価するWebページ評価手法について述べる。次節以降では、評価手法の概要とその特徴を述べ、その評価の予備実験例、さらには、情報サービスへの活用可能性、応用技術としての発展性をまとめる。

2．Webページ評価手法の概要

2.1 従来のWebページ評価手法

既に、各種のWebページ評価方法が提案されている[1]。それらは、一般的な出版物に対する評価項目を基に、(a)Webコンテンツの現状を考慮して新たな視点を付加した、チェックリストを用意して評価するものや、(b)表示スピードなどの利用者の利便性を重視したもの、(c)登録メンバに対するアンケートによる評価など、に発展させている。これらの評価手法は、評価価値判断の変化や利用者の状況に応じて評価値が変わり、ドメイン毎に柔軟に対応することは困難である。

2.2 発信内容によるWebページ評価手法

ディレクトリの充実や、キーワードを用いた検索技術も日々発展していることからも、Webページもドメインに応じた発信内容の充実が重要である。Webページの内容の充実には、対象ドメインごとの発信情報の内容に踏み込んだ評価が必要であり、それらの結果をフィードバックすることで、それぞれのWebページの価値を高め、完成度が高められる。

Web-page evaluation method for specific domain
Yoshinori KATAYAMA, Junko FURUKAWA, Fumihito NISHINO
FUJITSU Laboratories LTD.

発信内容によるWebページ評価手法の、基本的な流れを図1に示す。主な手順は、評価の対象ドメインを定め、そのドメインでの評価セットを作成し、評価セットにより Web ページを評価することである。評価セットは、情報発信のための基本となる共通項目と、ドメイン毎の特有項目などのメニュー表示項目、およびそのドメインの情報抽出技術により得られるテキスト情報内に埋め込まれている詳細情報項目からなる。これらにより、ドメイン毎の情報整理を実現する。

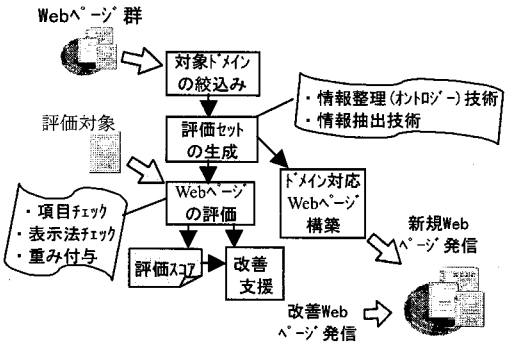

図1 Webページ評価手法の基本処理

本Webページ評価手法の特長は次の4点である。

- □対象ドメインを、カテゴリやWebページ指定などにより自由に規定できる。
- □メニュー表示等の情報表示項目から、テキスト情報に表現されている詳細情報項目まで、様々な粒度で内容情報の評価ができる。
- □評価結果のスコア算出に各種の重みを選択付与できる。
- □評価セットとなる各種の項目情報は、ドメイン毎の知識として蓄積される。

2.3 内容評価の効果

発信項目に基づくメニュー表示項目に加えて、情報抽出技術による詳細情報項目の内容情報による評価と、各内容情報の表示形式(深さ)による評価が行えることにより、主観的・表面的な判断でなく、客観的なスコアを用いて判断でき、改善点が明確に示される。また、対象ドメイン内での独自性の追及や、他ドメインにおける特徴の活用も、評価ドメインをダイナミックに変更することで容易に行える。

3．Webページ評価手法の予備実験

本評価手法では、評価セットとなる各項目の抽出が重要になる。抽出の基本は以下とする。

ジャンルにまたがる共通項目：複数のジャンルに対して、共通度の高いメニュー項目をPick-upする。

ドメイン依存の特有項目：対象ドメインに適当なWebページを複数指定し、それらのドメインに共通度の高いメニュー項目を、共通項目以外でPick-upする。

詳細情報項目：対象ドメインに適用可能な情報抽出ルールにおいて、情報抽出される項目をPick-upする。

これらの項目のPick-upは、自動化処理に向けたツール化を進めている。ここでの予備実験では、人手によりWebページをチェックして項目のPick-upを行い、以下のような評価セットを用意して評価を行った。

4種のジャンル(建築、自動車、化粧品、食品ﾒｰｶｰ)から、ランキング上位を対象に、主に7割以上の出現頻度のある１０項目を共通項目とした。

共通項目＝10項目(新着,ﾘﾘｰｽ,会社概要,採用,ｷｬﾝﾍﾟｰﾝ,業績,商品,問合せ,ｻｲﾄﾒﾆｭｰ,ﾒﾙﾏｶﾞ)

ドメイン依存の特有項目としては、食品ﾒｰｶｰを対象ドメインとして、知的情報収集サービス[2]で集められた優良ﾄｯﾌﾟﾍﾟｰｼﾞ128ページに対して、共通項目以外で5割以上の出現頻度のある5項目＋全ページを参照して、食品ﾒｰｶｰに必要と判断した2項目の計7項目をドメイン依存の特有項目とした。

特有項目＝7項目(工場見学,ﾄｯﾌﾟﾒｯｾｰｼﾞ,ｼｮｯﾌﾟ,ｲﾒｰｼﾞｷｬﾗ,ﾚｼﾋﾟ,健康,CM)

詳細情報項目としては、新聞記事情報からの企業情報抽出に用いた情報抽出ルールから、11項目を取上げた。

詳細情報項目＝11項目(企業名,資本金,社長,従業員数,住所,郵便番号,Tel,Fax,地図,決算,株価)

これらを評価セットとして、先の優良ﾄｯﾌﾟﾍﾟｰｼﾞ128ページの評価を実施した。

＜予備実験結果＞

評価対象において、HTML 間のリンク解析技術を応用したページの人気度ランキング[3]上位と下位での評価結果を比べたものを以下に示す。

○ 食品ﾒｰｶｰの優良ﾄｯﾌﾟﾍﾟｰｼﾞ人気度ランク上位50ページに対する各項目の出現率平均

共通項目出現率平均＝77.8%、特有項目出現率平均＝52.6%、詳細項目出現率平均＝72.8%

○ 同人気度ランク下位50ページに対する各項目の出現率平均

共通項目出現率平均＝57.3%、特有項目出現率平均＝21.9%、詳細項目出現率平均＝67.4%

予備実験である、今回の食品ﾒｰｶｰのWebページ評価では、詳細情報項目に関しては極端な違いは出ていないが、ジャンルにまたがる共通項目およびドメインに依存する特有項目に関しては、20ポイント以上の明確な違いが出ている。この結果から、評価セットとして採用した今回の各項目は、そのドメインにおいて必要な項目であり、人気度下位のページに対しては、発信項目や内容の示唆が行えることが解る。

＜蓄積された言語情報＞

評価セット内の各情報項目の抽出において、今回は人手で、表現の揺れや同義語などの統一化処理を行った。この処理により、多種多様な言語情報が蓄積された。特に、共通項目の整理における言語情報に加えて、ドメイン依存の特有項目の整理において、蓄積された言語情報は、各ジャンルの重要な知識となる。

＜評価における問題点＞

- **項目名称の扱い**：予備実験では、人手 Pick-upのため適当な項目名を代表としたが、同じ内容での表記の統計などによる判断が必要。
- **対象リンク範囲の指定**：今回の評価に関しては、ﾄｯﾌﾟﾍﾟｰｼﾞから同じサイトと判断できる範囲で扱ったが、今後、処理の自動化に際して検討が必要。
- **pdf やイメージデータへの対応**：今回の評価では全て対象外としたが、この形態での情報発信も行われているため今後検討が必要。

4．おわりに

ドメインに応じた評価によってWebページの完成度を高めることは、情報発信者にとっても利用者にとってもメリットが大きい。

また、Webページ評価によって蓄積される評価セットの情報は、ここで述べた評価だけでなく、ツール化によって、新規Webページを構築する際の内容項目支援に活用できる。さらには、評価セットとして、情報項目を整理する際に得られた言語情報は、これまで、専門的知識を必要としていた、ドメイン単位のオントロジーの構築に発展させられる。

謝辞　本手法に関して議論していただいた、富士通(株)ﾈｯﾄﾜｰｸｻｰﾋﾞｽ本部ｲﾝﾀｰﾈｯﾄｿﾌﾄ部、および富士通研究所ドキュメント処理研究部の方々の、ご協力に感謝いたします。

参考文献

1)J.Alexander,M.Tate:Evaluating Web Resources http://www.science.widener.edu/~withers/webeval.htm 1996.8

2)富士通㈱, ｼﾞｰ・ｻｰﾁ：「ValueContents」 http://pr.fujitsu.com/jp/news/2001/12/6-2.html 2001.12

3)津田,鵜飼,三末：Web ディレクトリのためのページメタデータの自動付与の試み、情報学ｼﾝﾎﾟｼﾞｳﾑ2002.1