検索意図を考慮したアクセスログ解析からのWebページの推薦に関する研究

寺口敏生† 田中成典‡ 中村健二‡ 山本雄平† 塚田義典‡
関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

1．はじめに

近年，インターネット上の情報量の増加[1]に伴い，大量の情報から必要な情報を検索するサービスや研究に注目が集まっている．このようなサービスとしては，検索語を用いて情報を検索するGoogleやYahoo! JAPANなどの検索エンジンがある．しかし，ユーザの検索意図[2][3]は多様であり，同じ検索語が入力された場合でも目的とする情報が同じであるとは限らない．そこで，検索エンジンの検索結果をクラスタリングすることで，ユーザの目的とする情報の発見を支援する研究[4][5]がある．しかし，これらの既存研究では，出力結果はWebページの記述内容に依存するため，概念的に適切な検索語を入力したとしても，Webページの内容に検索語が含まれていない場合に，適切なWebページを出力できないという問題がある．そこで，本研究では，複数のユーザの検索語，アクセス履歴や閲覧時間といったアクセスログを基に，ユーザの検索行動毎の検索意図に合致した目的ページを推定する.そして，検索語をタグとして目的ページに付与することで，新たに情報を検索する際に検索意図を考慮した適切なWebページを検索結果として推薦する手法を提案する．

2．研究の概要

本研究では，アクセスログの解析結果に基づきWebページを意味付けすることで，適切なWebページを推薦する手法について提案する．本システム（図1），1）アクセスログ解析機能，2）Webページ推薦機能により構成される．情報検索時における入力データは検索語とし，出力データは推薦するWebページとする．

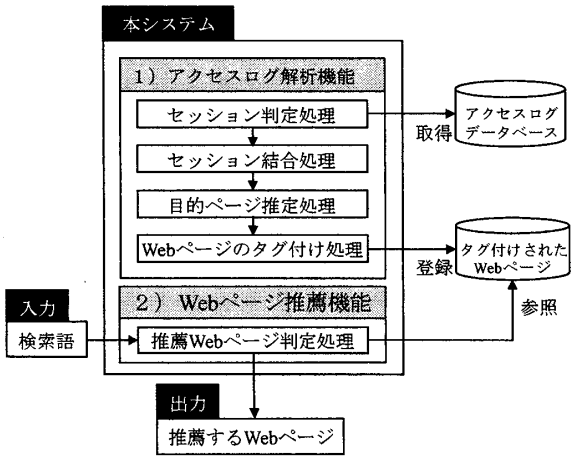

図1　システムの概要

2．1　アクセスログ解析機能

本機能では，セッション判定処理，セッション結合処理，目的ページ推定処理とWebページのタグ付け処理の4つの処理を行う．セッション判定処理では，複数のユーザから収集したアクセスログをユーザ毎に解析し，アクセス間隔と検索語の変化の情報を用いてアクセスログを検索行動の単位に分割する．このように，本研究では，検索行動の単位に分割されたアクセスログを1セッションとする．セッション結合処理では，各セッションの前後のセッションに含まれるWebページの内容を解析し，その類似度をVSM（Vector Space Model）により算出する．そして，類似度が閾値以上のセッションを結合する．目的ページ推定処理では，各ユーザにおける対象Webページの閲覧時間と平均閲覧時間の差，および，対象Webページにおける各ユーザの閲覧時間と全ユーザの平均閲覧時間の差が一定範囲内であれば，特徴的な閲覧行動であると判断し，対象Webページが目的ページであると推定する．Webページのタグ付け処理では，推定した目的ページに対して検索語をタグとして関連付ける．

2．2　Webページ推薦機能

本機能では，タグ付けされたWebページ群か

Research for Recommending Web Page with User's Search Intentions from Access Logs
†Toshio Teraguchi , Yuhei Yamamoto
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho, Takatsuki-shi, Osaka 569-1095, Japan
‡Shigenori Tanaka, Kenji Nakamura, Yoshinori Tsukada
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho, Takatsuki-shi, Osaka 569-1095, Japan

らユーザが入力した検索語に一致するWebページを検索し，一致したタグの出現頻度が高い順にランキングして推薦する．

3．システムの実証実験と考察

実証実験では，目的ページの推定精度の検証とWebページに付与したタグの妥当性を評価し，本提案手法の有用性を検証した．

3．1　実証実験

本実験の実験データとして，インターネットと同様に多様な検索行動が見られるEC（Electric Commerce）サイトにおけるアクセスログ10,810件を用いた．実験データ内の目的ページは，セッション内の購入に至った商品のWebページとそれに類似するWebページの合計270ページとした．これらのページを目的ページとすることで，推定精度を客観的に評価できると考えたためである．

本研究では，次の2種類の実証実験を行う．目的ページの推定精度の検証実験では，適合率，再現率とこれらの調和平均であるF値を評価値として推定精度を検証した．Webページに付与したタグの妥当性の評価実験では，定義した正解データと本システムが出力した正解データに基づいて付与したタグの一致状況について評価した．

3．2　結果と考察

目的ページの推定精度の実験結果（表1）では，再現率は比較的高い値であったが，適合率とF値は非常に低い値であった．そこで，誤判定したWebページを分析すると，閲覧時間が長いWebページや複数回表示されたページのように，目的ページと同様の特徴が見られた．このため，Web全体で見ると，閲覧時間に関して，目的ページの閲覧時に見られる特徴が表れるWebページは，目的ページである可能性が高いため，適合率とF値は改善されると考えられる．

Webページに付与したタグの妥当性についての実験結果（表2）では，完全一致と部分一致を併せた155件にて，正解データとほぼ同様のタグをWebページに付与できた．しかし，458件のWebページに対しては，正解データとは異なるタグを付与した．この原因として，実際にWebページに付与された主要なタグの一覧（表3，表4）を分析した結果，正解データ以外のWebページに対してタグを付与した場合においても，適切なタグが付与されるケースが見られた．しかし，複数のWebページに同様のタグが付けられており，タグ付けの方法を改善する必要があることが分かった．

表1　目的ページの推定精度

	適合率	再現率	F値
提案手法	0.22	0.70	0.33

表2　付与したタグの妥当性

正解データとの関係	本システムの出力結果
完全一致	72件
部分一致	83件
不一致	458件

表3　正解データに基づくタグの付与例

Webページの内容	タグ
水を流す圧力を利用して吸引を行う装置	金属製アスピレーター，アスピレーター，金属アスピレーター
ガラス製のビーカー	3-317-344，アルコールランプ，ガラスビーカー，キムワイプ，ピペット台，レンズクリーニング
複数の液体の混合物を分離する装置	チビタン XX42CF0RT チビタン R，遠心分離機
pHを測定する試験紙	PH，試験紙

表4　システム出力結果に基づくタグの付与例

Webページの内容	タグ
機具を固定するスタンドの部品	ゴム栓，シリコン，スタンド，パラフィルム，ビーカー，フラスコ，ルツボ，ロート，丸型クランプ，試験管，試験管立，洗浄びん，ガラス管
ガラス製の平皿	シャーレ，HCL，カウンター，サンプリングチューブ，チップ，塩酸，天秤，培地
酸の濃度を測定する試験紙	RQ，SPAD，ワンダーブレンダー，UVgl，VM-90A，ステンレスビーカー用蓋，ダウグリース
容器に蓋をするフィルムを切断する機器	ゴム栓，シリコン，スタンド，パラフィルム，ビーカー，フラスコ，ルツボ，ロート，丸型クランプ，試験管，試験管立，洗浄びん

4．おわりに

本研究では，検索語だけでは抽出できない目的ページに対し，ユーザのアクセスログを用いてタグを付与する手法を提案し，その有用性を検証した．今後は，目的ページの抽出手法を検討し，提案手法の精度向上を試みる．

参考文献

[1] 総務省：平成21年度版情報通信白書，ぎょうせい，2009.7.

[2] Broder, A. : A Taxonomy of Web Search, SIGIR Forum, ACM, Vol.36, No.2, pp.3-10, 2002.9.

[3] Marchionini, G. : Exploratory Search: from Finding to Understanding, Communications of the ACM, ACM, Vol.49, No.4, pp.41-46, 2006.4.

[4] 城市広大，三好力：ベクトル空間法とファジィ推論を用いたWEB検索結果自動分類システム，知能と情報，日本知能情報ファジィ学会，Vol.18，No.2，pp.184-195，2006.4.

[5] 大野成義，渡辺匡，片山薫，太田学，石川博：Max Flowアルゴリズムを用いたWebページのクラスタリング方法とその評価，情報処理学会論文誌 データベース，情報処理学会，Vol.47，No.SIG_4，pp.65-75，2006.3.