検索シーンにおけるユーザの行動情報を考慮した閲覧履歴の推薦

藤沢 和哉 † 坪川 宏 †
†東京工科大学 コンピュータサイエンス学部 コンピュータサイエンス学科

1 はじめに

近年のIT技術の発展の伴い，ユーザは膨大な数のWebページから必要な情報を取得するため，検索エンジンでの情報探索を効率的に行う技術が要求されてきている．しかし，検索技術のないユーザは情報の取捨選択に時間を費やしてしまうという現状である．また，多くの検索エンジンではどのユーザに対しても同一検索キーワードの検索結果が同じになっているため，各ユーザに合わせた検索支援が行われていない．

これらの問題を解決するため，各ユーザの閲覧履歴からユーザの目的に沿った推薦の手法が提案されている．しかし，既存の手法ではユーザが検索により目的とする情報に至までに各Webページ中で起こした行動を考慮に入れていないため，ユーザが真に注目していたWebページが曖昧であり，ユーザが注目していないWebページ でもアクセス履歴があれば関係のある情報として扱われてしまうという問題がある．

2 目的

本稿では各ユーザの検索における目的や検索領域を検索シーンと定義し，閲覧履歴からユーザの目的を推測し，ユーザの検索シーンに合わせた推薦を行うことを目的とする．従来のようにページタイトルやURLなどのWebページ情報だけを閲覧履歴と考えるのではなく．滞在時間やブックマーク要求，キーボード操作などのWebページ内で起こす行動情報も閲覧履歴として考慮に入れる．これらの行動履歴を取得することにより，閲覧した各Webページに重み付けが可能となり，ユーザが閲覧したWebページがどの程度注目されていたかということを推測できるため，ユーザの検索シーンを明確に把握できると考えた．本システムによりユーザの目的をベクトルで表現し，個々のユーザの目的に合わせた特徴ベクトルを生成して推薦を行う手法を提案する．

3 提案手法

3.1 システムの概要

図1に本手法の概要を示す．本手法の流れとして目的別にグループ化された閲覧履歴から基準となる基準特徴ベクトルを抽出し，グループ中の各閲覧履歴の特徴語と行動情報から特徴ベクトルの要素の値を変化させることによりユーザの目的を反映した拡張特徴ベクトルを生成する．そして，生成した各グループの拡張特徴ベクトルの類似度を計算することで推薦する閲覧履歴を決定する．

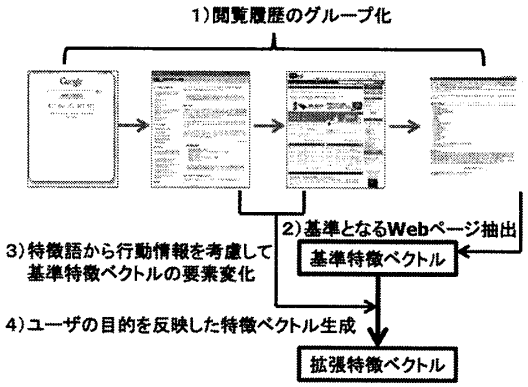

図1: 提案手法の概要

3.2 閲覧履歴の取得

取得する情報は検索情報として検索エンジンに入力された検索キーワードとアクセスWebページのリンク元URL，Webページ情報としてページタイトルとURL，行動情報として印刷要求有無，ブックマーク要求有無，滞在時間，マウス移動量，検索文字列，コピー文字列，クリックアンカーテキストをWebページが切り替わる度に閲覧履歴を表す1つの単位として取得する．

3.3 閲覧履歴のグループ化

ユーザが検索行動を行う際，1つのWebページだけから目的とする情報を得ようとするのではなく，複数のWebページを閲覧することにより目的を達成することも考えられる．そのため，1つの目的に沿った検索行動を閲覧グループとしてグループ化することでユーザの閲覧履歴を管理する．

そこで，ユーザの目的を表すと考えらる検索キーワードに着目し，時系列順に取得した閲覧履歴の検索キーワードの前後関係を比較することにより閲覧グループを判定する．

Recommendation of browsing history that considers user's action information in retrieval scene
† Kazuya Fujisawa
† Hiroshi Tsubokawa
Tokyo University of Technology (†)

3.4 特徴語の抽出

閲覧グループの各 Web ページに対して特徴語を抽出するため，TFIDF 法を用いる．Web ページからテキスト文書を取得後，テキスト文書に対して形態素解析を行い名詞のみを抽出する．以下の式を用いて TFIDF 値を求め，TFIDF 値が上位の名詞を特徴語として抽出する．

$$W_{ij} = tf_{ij} \times log(\frac{N}{df_{ij}})$$

W_{ij}：Web ページ i における単語 j の TFIDF 値
tf_{ij}：Web ページ i における単語 j の出現回数
N：閲覧グループの総 Web ページ数
df_j：キーワード j を含む Web ページ数

3.5 注目度の測定方法

検索シーンにおける閲覧グループの各 Web ページに対する重要性はユーザによって様々であるため，注目度という尺度を独自に定義し，ユーザの行動情報から各 Web ページがどの程度注目されたかということを評価する．

表 1 の対応表から，以下の式に示したように各行動の出現頻度と重み係数の積算により Web ページの注目度を算出する．k は行動の種類の数を表す．

$$Att_i = \sum_k (Ac_k \times Fr_k)$$

Att_i：Web ページ i の注目度
Ac_k：行動種類 k の重み係数
Fr_k：行動種類 k の出現頻度

表 1: 行動と重み係数の対応

行動種類	重み係数
滞在時間 (秒)	1
ブックマーク要求	300
印刷要求	300
文字列検索	60
文字列コピー	30

3.6 特徴ベクトルの生成

閲覧グループにベクトル空間モデルを適用し，特徴ベクトルを生成することでユーザの目的を表現する．特徴ベクトルは Web ページのテキスト文書に形態素解析を行い，全名詞数を特徴ベクトルの要素数，各名詞の出現回数をその要素の重みとすることで抽出する．まず，閲覧グループの中から注目度が最も高い Web ページを基準特徴ベクトルとして表す．次に，注目度に閾値を設定し，閲覧グループの注目度が閾値以上の Web ページはユーザが注目した適合 Web ページだと仮定し，その Web ページの特徴語や行動情報から抽出した単語は基準特徴ベクトルの該当する要素の重みを大きくする．注目度が閾値未満の Web ページはユーザが注目しなかった不適合 Web ページだと仮定し，同様の方法で該当する要素の重みを低くする．この新たなベクトルを拡張特徴ベクトル q'，基準特徴ベクトルを q とすると以下の式のように表すことができる．

$$q' = q + \sum_{d_m \in D_R} d_m - \sum_{d_n \in D_S} d_n$$

d_m：適合 Web ページ i の特徴ベクトル
d_n：不適合 Web ページ j の特徴ベクトル

3.7 閲覧グループの類似度

閲覧グループはユーザの検索シーンを表す一連のプロセスのため，この閲覧グループを推薦に用いる 1 つの単位とする．各閲覧グループの類似度をコサイン類似度で計算し，類似度の降順に閲覧グループをユーザに推薦する．以下の式は特徴ベクトル X と Y の類似度を表す．

$$sim(X, Y) = cos(X, Y) = \frac{\vec{X} \cdot \vec{Y}}{|X| \cdot |Y|}$$

4 検証結果

検証方法として PHP という単語を主キーとしたときの検索を，検索エンジンが提示する副キーワードの組み合わせてにより複数パターン行い閲覧履歴を収集した．検索支援を要求するユーザは PHP の概要を表すような Web ページを基準とし，データベース関連についての Web ページを適合 Web ページ，プログラム関連についての Web ページを不適合 Web ページとなるよう閲覧を行った．基準特徴ベクトルから類似度を計算した場合をパターン A とし，行動情報を考慮して拡張特徴ベクトルを生成した場合をパターン B とした時，A では下位に存在した適合 Web ページの特徴語を副キーワードや特徴語に持つ履歴履歴が B では上昇し，A からの類似度の増加値が平均 0.15 となった．また，A では上位に存在した不適合ページの特徴語を持つ閲覧履歴が B では下降し，A からの類似度の減少値が平均 0.12 となったことから，閲覧履歴の特徴語と行動情報からユーザの目的を表現することが可能だと考えられる．

5 まとめ

今後は閲覧履歴の可視化のための表示インターフェースの実装を行い，提案手法の有用性を評価したいと考えている．

参考文献

[1] 中島 伸介，黒田 慎介，田中 克己:閲覧履歴を反映したコンテクスト依存型 Web ブックマーク，情報処理学会論文誌:データベース，2002
[2] 森田 哲之，倉 恒子，日高 哲雄，大浦 啓一郎，田中 明通，加藤 泰久，奥 雅博:体験獲得情報を想起させる行動検索手法，情報処理学会論文誌，2007