Web閲覧履歴の共有による検索効率改善のためのグループ形成手法の提案

山口 雄大† 新美 礼彦† 小西 修‡
公立はこだて未来大学 システム情報科学部 情報アーキテクチャ学科†
公立はこだて未来大学 システム情報科学部 複雑系科学科‡

1 はじめに

Web 上の情報量は増加の一途をたどっており、その膨大な情報源から、効率的な情報収集を実現するために、様々なサービスの開発や研究が行われている。その一つに、グループの検索活動を支援する研究がある。興味や関心が似ているグループ内では、Web 検索の目的、閲覧 Web ページの内容に重複があり、それらを利用することで検索要求を効率良く満たせる可能性が示されている[1]。しかし、それらの研究では、同じ検索目的を持っている、または興味や関心の似ているユーザグループを明示的に定義しているため、適用範囲が限られている。

本研究では、多数のユーザ間で Web 閲覧履歴を共有することによって Web 検索の効率を改善するシステムを提案し、本稿では、そのシステム内における、多数ユーザの履歴情報の整理について検証する。

2 提案手法

本研究では、個人の検索活動の支援に多ユーザの Web 閲覧履歴を利用することを考える。膨大なユーザの、検索目的や閲覧 Web ページの内容の重複を整理することで、ユーザグループを特定せずとも、個人の Web 検索の効率を改善できると考えられる。そこで、本研究が提案するシステムは、多数のユーザの Web 閲覧履歴を検索目的別に一括管理し、それらの履歴情報を逆引き検索できるシステムである(図1参照)。

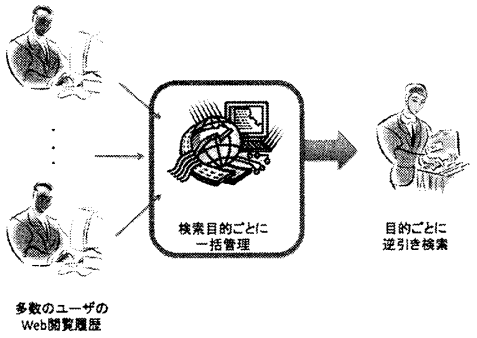

図1. 提案システムの概要と本稿での検証箇所

A Method of Organizing the Group to Improve Search Efficiency with Sharing History of Web Browsing
† Takehiro Yamaguchi・Future University Hakodate
† Ayahiko Niimi・Future University Hakodate
‡ Osamu Konishi・Future University Hakodate

つまり、ユーザをグループ化するのではなく、履歴データをグループ化することを提案する。Web 閲覧履歴を検索目的ごとに整理することで、適切な検索キーワードの選出作業を軽減する「逆引き検索」を可能とするだけでなく、履歴を共有するユーザ数が増加しても特徴量が薄くなることを軽減できると考えられる。

本稿では、一つの検索目的に沿った履歴集合を「検索タスク集合」、検索目的の類似する検索タスク集合を「検索タスクグループ」と定義し、多数のユーザの Web 閲覧履歴集合に含まれる検索タスク集合から、検索タスクグループの自動生成について検証する。検証手順としては、各ユーザの Web 閲覧履歴集合から、検索タスク集合に分割したものを多ユーザ分集め、それぞれの検索タスク集合の特徴ベクトルを比較し類似度を算出する。その類似度を基にクラスター分析を行い、出来たクラスターを検索タスクグループとする。そして、同じ検索タスクグループ内に、同じ検索目的である検索タスク集合がどの程度含まれるかを検証する。

2.1 Web閲覧履歴集合の細分化

既存の Web ブラウザが保持している履歴データは、時系列に整理されたデータである。さらに、ユーザによって検索目的を達成するための検索活動の過程は多種多様であり、検索目的ごとの履歴データを一意に区切ることは困難である。そこで、本研究では、Google などの Web 検索エンジンに検索目的を示したキーワードを入力したページから、そのキーワードと相関のないキーワードが入力された Web 検索エンジンのページまでを一つの検索目的に沿った検索タスク集合と定義し、本稿ではその一つ一つの検索タスク集合を明示的に定義した。

2.2 検索タスク集合の特徴ベクトル

各検索タスク集合を特徴づけるキーワードを抽出し、検索タスク集合それぞれの特徴ベクトルとして定義する。検索タスク集合内の各履歴ページの本文に対して、形態素解析を行い、すべての名詞を抽出する。抽出した名詞から不要語フィルタを用いて抽出したキーワードを属性に、そのキーワードの出現頻度(tf)に重み(idf)をつけたスコアを要素とした特徴ベクトルを形

成する。ある検索タスク集合 H においてキーワード w が出現する頻度を tf(w,H)、全ユーザの Web 閲覧履歴集合内の履歴ページ総数を N、そのキーワードの検索タスク集合内における出現履歴ページ数を n、そのキーワードの重みを idf(w)とすると、そのキーワードのスコア S(w)、そのスコアによって形成される検索タスク集合ごとの特徴ベクトル v は、それぞれ以下の式で定義する。

$$S(w_i,H) = tf(w_i,H)\cdot idf(w_i) = tf(w_i,H)\cdot\left(\frac{N}{n}\right)$$

$$v(H) = \left[S(w_1,H),\cdots,S(w_i,H)\right]^t$$

検索タスク集合内のキーワードの出現頻度に全ユーザの Web 閲覧履歴集合におけるキーワードの重みを用いることで、各検索タスク集合の特徴を際立たせることができる。

2.3 検索タスク集合間の類似度

各検索タスク集合の特徴ベクトルの類似度計算にはベクトルの長さが類似度に影響がない、コサイン類似度を用いる[2]。２つの特徴ベクトル v_1, v_2 の類似度 $sim(v_1, v_2)$ を以下の式によって定義する。

$$sim(v_1,v_2) = \cos(v_1,v_2) = \frac{\sum_w (v_1(w)\cdot v_2(w))}{\sqrt{\sum_w v_1(w)^2}\cdot\sqrt{\sum_w v_2(w)^2}}$$

特徴ベクトル v_1, v_2 は、それぞれある１つの検索タスク集合の特徴ベクトルを表しており、その属性であるキーワードはすべての検索タスク集合から抽出したキーワードで形成されている。しかし、その履歴集合に出現しないキーワードの要素であるスコアは 0 となるため、ベクトルの長さが類似度に影響することはない。

2.4 検索タスクグループの形成

各検索タスク集合間の類似度を基にクラスター分析を行い、検索タスクグループを形成する。各検索タスク集合を１つのクラスターとし、類似度の最も高い検索タスク集合同士を併合する。併合後の類似度には、併合されたクラスター内のクラスターと、他のクラスターとの類似度の最大値を用いる。そして、予め定めた検索タスクグループ数になるまで再帰的に併合する。

3 評価実験および実験結果

提案手法を評価するため、実験を行った。61 人の被験者に 4 つの検索課題を与え、その際の Web 閲覧履歴を収集し、明示的に定義した各検索タスク集合を、同じ検索課題である検索タスク集合ごとに、同一クラスター内にクラスタリングできるかを検証した。本検証では、検索タスクグループ数を予め定めずに、クラスター間の類似度の最大値が 0.5 以下になった時点でクラスターの併合を止めた。その時の検索タスクグループ数は 22 であり、そのうち最大のもの４つを取り上げ、結果とした。実験結果について、再現率、適合率を用いて評価した。検索課題 i における検索タスク集合を同一クラスター内にクラスタリングできた検索タスク集合数を C_i、そのクラスター内にクラスタリングされた他の検索課題における検索タスク集合 O_i、そのクラスター内にクラスタリングできなかった検索課題 i における検索タスク集合数を E_i とし、再現率 R_i、適合率 P_i を以下の式で定義する。

$$R_i = \frac{C_i}{C_i + E_i} \quad P_i = \frac{C_i}{C_i + O_i}$$

再現率は平均 85%であり、適合率は全て 1.0 であった(表１参照)。正しくクラスタリングできなかった検索タスク集合は、正確に履歴データを収集できなかったものや、検索課題を解答するには見当違いな履歴ページが含まれている検索タスク集合、複数の検索課題の回答をサポートするような履歴ページが存在する検索タスク集合がほとんどであったため、平均 85%のクラスタリング精度でも良い結果であったと考えられる。しかし、正しくクラスタリングできた検索タスク集合のほとんどが、少数の Web ページで検索目的を達成している検索タスク集合であったため、多数の Web ページを横断的に閲覧するような検索課題に対しても有効な手法かを検討する必要がある。

表 1. 実験結果

	課題１	課題２	課題３	課題４
再現率	0.85	0.90	0.73	0.91
適合率	1.0	1.0	1.0	1.0

4 おわりに

本稿では、本研究が提案する検索効率を改善するシステムの概要、そして、そのシステム内における Web 閲覧履歴の整理に対する検証結果を示した。今後の展開として、明示的に定義した検索タスク集合を自動的に判別できるようにすることなどが挙げられる。

参考文献

[1] 武田達弥, 五十嵐健夫: グループでウェブの探索を効率化する検索共有インタフェース; 情報処理学会研究報告.HCI, ヒューマンコンピュータインタラクション研究会報告, Vol.2008, No.11, pp.93-98

[2] 大島裕明, 小山聡, 田中克己: 文書群をクエリとした“似て非なる”文書の検索; 日本データベース学会 Letters, Vol.5, No.1, pp.121-124