Web 検索効率改善のための Web 履歴の分類とグループ化

Grouping of Web browsing history to improve efficiency of Web searching

山口雄大[*1] 新美礼彦[*2] 小西修[*2]
Takehiro Yamaguchi Ayahiko Niimi Osamu Konishi

[*1]公立はこだて未来大学大学院システム情報科学研究科
Graduate School of Systems Information Sciences, Future University-Hakodate

[*2]公立はこだて未来大学システム情報科学部
Systems Information Sciences, Future University-Hakodate

Information contents on the Web have grown steadily. And, to achieve effective collecting information from the huge information source, there are various developments of service and research. We propose a system to improve efficiency of personal Web searching. The system organizes a lot of user 's web browsing history into same purpose of Web retrieval and reuses their history. In this paper, we show classification of their history focused on changing the search keywords and grouping of their history with similarity of the search keywords in the system.

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 [武田 08]．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，適用範囲が限られる．そこで，本研究ではユーザグループを特定しない，多ユーザ間の Web 履歴共有システムを提案し，そのシステムにおける，個人の Web 履歴の分類と多ユーザの Web 履歴のグループ化について検証する．

2. 提案システム

本研究では，Google などの検索エンジンを利用するユーザ数の多さ [Forbes 08] に着目したシステムを考察した．検索エンジンを利用するユーザ数が多いほど，それらのユーザ間で日々の検索活動の目的に重複が存在する可能性が高くなり，それら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，各ユーザにとって扱いやすい状態で Web 履歴を共有できると考えられる．そこで，本研究が提案するシステムを図 1 に示す．

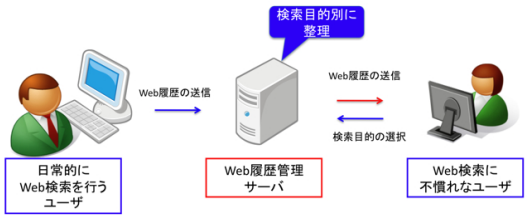

図 1: 提案システム

連絡先: 山口雄大,
公立はこだて未来大学大学院システム情報科学研究科,
北海道函館市亀田中野町 116 番地 2,
Mail: g2109046@fun.ac.jp

提案システムでは，日常的に Web 検索を行うユーザの Web 履歴を多数用意し，それらの履歴データをサーバ上で一括管理する．サーバ内では，それらの履歴データを同じ目的で検索された際の履歴データごとに整理する．そして，Web 検索に不慣れなユーザがそのサーバが提示する検索目的の一覧から，自分の検索目的と類似するものを選択することで，それらの履歴情報を逆引き検索できるシステムである．Web 履歴を検索目的ごとに整理することで，既に同じ事柄について調べたユーザの履歴データを順に追う事ができるため，Web 検索に不慣れなユーザの検索キーワード選出作業を軽減することが可能になると考えられる．

3. Web 履歴の整理手法

本研究では，個人の Web 履歴に含まれる，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，個人の Web 履歴を検索タスク集合ごとに自動分類し，多ユーザの Web 履歴に含まれる検索タスク集合から，検索タスクグループを自動生成する手法について検証する．これまでに，明示的に定義した検索タスク集合から，各履歴ページに含まれるキーワードを基にベクトル空間法と階層的クラスタリングを用いて，検索タスク集合を自動生成する手法について検証してきた [山口 09]． 本研究では，検索タスク集合を明示的に定義せず，検索キーワードの変化とその類似性に着目した個人の履歴データから検索タスクグループの自動形成までの手法について具体化する．

3.1 Web 履歴の分類

個人の Web 履歴を一つの検索目的ごとに分類する際には，入力する検索キーワードの時間的前後関係に着目する．具体的には以下のステップに従い，Web 履歴の切り分けを行う．

Step 1: 個人の Web 履歴を検索結果が出力された履歴ページごとに区切り，履歴集合を形成する．

Step 2: 形成された履歴集合の前後間で検索キーワードを比べる．

Step 3: 同一の検索キーワードが含まれる場合に，それらの履歴集合を統合する．

[鈴木 02] で示されている複数回検索しているユーザの検索パターンから，同じ目的で連続して複数回の検索が行われる場合に，その前後で同じ検索キーワードが存在する可能性が高いと考えられる．そこで上記のステップによって切り分けられた，または切り分けられ統合された履歴集合を検索タスク集合とした．図 2，3 は，それぞれのステップの具体例を示している．

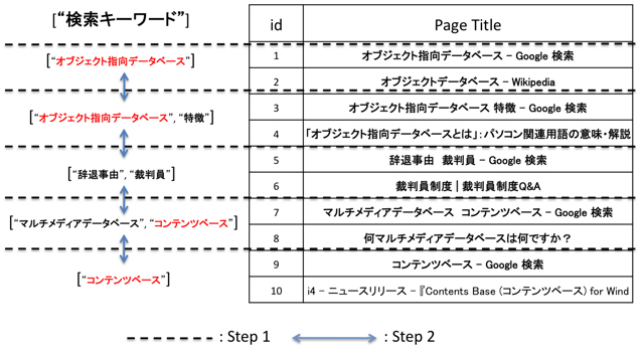

図 2: Step1，Step2 の具体例

図 2 内のテーブルはユーザの履歴データを時系列に整理したものであり，それぞれの履歴ページのタイトルの一覧を表している．Step1 によって，破線で示したように 5 つの履歴集合（id:1,2），(3,4），(5,6），(7,8），(9,10）に切り分けられる．Step2 によって，Step1 で形成された履歴集合の前後間で検索キーワードが比較され，Step3 で「オブジェクト指向データベース」,「コンテンツベース」が含まれるそれぞれの履歴集合が統合される．Step3 によって，統合された履歴集合は図 3 の赤枠で囲まれた履歴集合である．

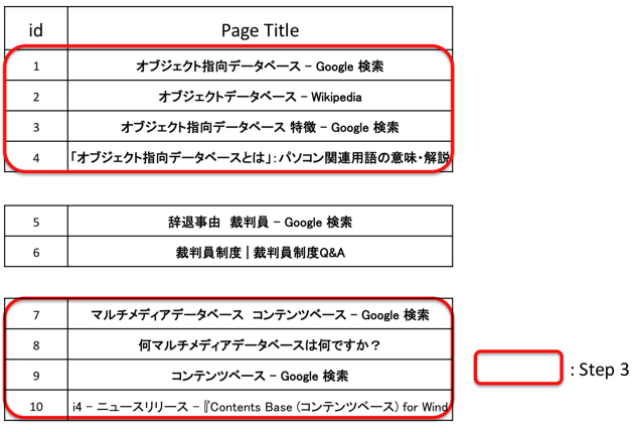

図 3: Step3 の具体例

このアルゴリズムを適用することで，上記の 10 件の履歴データから 3 つの検索タスク集合（id:1,2,3,4），(5,6），(7,8,9,10) が形成される．

3.2 Web 履歴のグループ化

ユーザをまたいで類似する Web 履歴をグループ化する際には，検索タスク集合内に含まれる検索キーワードの類似性に着目する．検索タスク集合に含まれる検索キーワードを属性に，そのキーワードの出現の有無を要素とした検索キーワードベクトルを形成し，比較するベクトル同士の成すコサイン値を類似度とする．検索タスク集合 i，j の検索キーワードベクトルをそれぞれ v_i，v_j とすると，求める類似度 $sim(v_i, v_j)$ を以下の式によって定義する．

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

上記の式によって算出された類似度の高い検索タスク集合同士を階層的クラスタリングの最短距離法を用いて併合する．そして，類似度の最大が閾値を下回った時点で併合を終了し，その時点で形成されている各グループを検索タスクグループとする．

[山口 09] では，履歴ページからキーワードを抽出してキーワードベクトルを作成しているが，検索タスクとは関係が少ないキーワードも抽出してしまうという問題から，本研究では検索キーワードを用いてキーワードベクトルを作成した．

4. 評価実験

提案手法を評価するために，本学の学生 5 名に 10 問の検索課題（表 1 参照）を与え，その際の Web 履歴を収集し，提案手法による Web 履歴の分類，検索タスクグループのクラスタリング終了条件の閾値を 0.5 に設定したグループ化を行った．また，全ての被験者に対して，それぞれの検索課題に対する事前知識の有無やその度合い，普段の検索活動についてヒアリングを行った．

表 1: 検索課題の内容

	検索課題
1	ソマリア沖における海上警備行動の目的と具体的な活動内容
2	クイックソートと比較した場合のヒープソートの特徴
3	地方の高速道路料金が休日（土日祝日）に「上限1000円」になる条件
4	SRAMと比較した場合のDRAMの特徴
5	課題3で取り上げた以外の高速道路の割引サービスの1つ
6	“フェールセーフ”と“フェールソフト”の違い
7	裁判員制度における裁判員選任手順
8	関係データベースと比較したときのオブジェクト指向データベースの特徴
9	裁判員法が定める辞退事由
10	マルチメディアデータベースに対するコンテンツベースの検索の利点と欠点

4.1 評価方法

提案手法による Web 履歴のグループ化結果に対する評価に，Adjusted Rand Index[Hubert 85]（以降，ARI とする）を用いた．[長野 08] でいわれているように，ARI は同一の分類対象を有する二つの分類方式の類似性を図るものであり，その値は主に 0〜1 の値をとり，1 で完全一致，0 でランダムによるクラスタリングの期待値となる．[長野 08] では，一方を提案方式による分類結果，一方を正解分類結果として ARI を適用することで分類方式の評価を行っている．本研究においても，一方を提案手法によるグループ化結果，もう一方を正解グループ結果として ARI 値を算出し，提案手法の評価を行った．

整理対象の履歴ページ総数を n，提案手法によるグループ結果と正解グループ結果で同じラベル付けされた履歴ページ数を n_{ij}，提案手法によるグループ結果で i とラベル付けされた履歴ページ数を $n_{i.}$，正解グループ結果で j とラベル付けされた履歴ページ数を $n_{.j}$ とすると，求める ARI 値は以下の式によって算出される．

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2} / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2} / \binom{n}{2}}$$

また，正解グループ結果は，被験者に明記してもらった各検索課題の開始時刻をもとに作成した．

4.2 実験結果

正解分類数が 10 に対して，提案手法により自動形成したグループの数は 36 であった．分類されたそれぞれのグループに対して，そのグループに最も多く含まれる正解データをそのグループのラベルとした．また，同一のラベルがついたグループが複数あったため，その中でも最も正解データ数の多いグループを採用し，それ以外のグループは未分類グループと定義し，ARI 値を算出した．算出された ARI 値は 0.54 であった．

4.3 考察

Web 履歴の収集に用いた検索課題 (表 1 参照) には，類似するトピックスの課題を複数用意したが，それらのタスクを混合することなくグループ化することができた．しかしこれは，それら類似するタスク同士を連続して取り組まないような課題順番にしたことがその要因として考えられる．例えば，検索課題 2 と 4 に対して,「比較」,「特徴」という検索キーワードを使用している被験者が複数いたため，これらの課題を連続して取り組んだ場合，その検索キーワードが履歴集合の前後間で一致する可能性が高く，提案手法ではそれら異なる検索内容を同じ検索目的としてしまうことが考えられる．

また，多くの Web 履歴が正しくラベリングできたにも関わらず，同じラベルの付いた異なる検索タスクグループが多数形成されてしまった．これに関しては，検索課題にその要因があると考えられる．例えば，その課題における閲覧ページ数に対して最も多くの検索タスクグループが形成された検索課題 3 は以下の様式で出題した．

- 地方の高速道路料金が休日（土日祝日）に「上限 1000 円」になる割引が 2009 年 3 月 28 日から全国的にスタートした．割引の対象となる条件を調べてください．

この検索課題に対して被験者が実際に使用した検索キーワードは,「高速道路料」,「高速道路料金」,「高速料金」,「1000」,「1000 円」,「割引」,「条件」,「割引対象」,「サービス」と多種類のものが使われた．これに対して，複数の検索タスクグループを形成しなかった検索課題 4 は以下の様式で出題した．

- SRAM と比較した場合の DRAM の特徴を調べてください．

この検索課題に対して被験者が使用した検索キーワードは,「SRAM」,「DRAM」,「比較」,「違い」と少数のキーワードしか使われなかった．実験結果と事前知識についてのヒアリング結果を照らし合わせると，事前知識の有無に関わらず，検索課題本文から検索キーワードを選出している傾向があり，そのため検索課題 4 よりも多種類の検索キーワードが思いつきやすい検索課題 3 のほうが多数の検索タスクグループを形成したと考えられる．したがって，検索課題 3 のように完全に一致するキーワードではないが，同じ意味，同じ目的で使われるキーワードに対応する工夫が必要であると考えられる．

また，検索課題本文のキーワードを使っても目的の情報が得られない場合には，閲覧ページ内に含まれるキーワードを新たな検索キーワードとして試すケースも見られ，履歴ページ本文のキーワードをもとにした動的に変化する検索キーワードへの対応も今後の課題として考えられる．

5. まとめ

本研究では，多ユーザの Web 履歴を同じ検索目的ごとに整理することで検索効率の改善を試みるシステムを提案した．またそのシステムにおける，検索キーワードの変化とその類似性に着目した多ユーザの Web 履歴の整理手法について評価実験を行った．その結果，類似する検索課題も混合することなくグループ化することができたが，同じラベルの付いた異なるグループが多数形成されてしまい，それらの要因について概観した．

今後の展開として，被験者の数を増やしたデータに対して履歴ページ全体のキーワードに着目した分類手法を適用し算出した ARI 値と本手法の値の比較を予定している．さらに，用意した検索課題に取り組んだ際の履歴データではなく，日常的に使用している Web ブラウザの履歴データに対する評価も検討している．

参考文献

[武田 08] 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, 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 (2008)

[Forbes 08] Forbes, What Are People Actually Doing On The Web ? , http://www.forbes.com/ (2008)

[山口 09] 山口雄大, 新美礼彦, 小西修: 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, 情報処理学会第 71 回全国大会講演論文集, 5P-4 (2009)

[鈴木 02] 鈴木俊輔, 山名早人: 時間間隔を用いた検索履歴のモデル化, 情報処理学会研究会報告. 情報学基礎研究会報告, Vol.2002, No.28, pp.103-110 (2002)

[Hubert 85] Hubert, L. and Arabie, P, Comparing partitions. Journal of Classification, pp.193-218 (1985)

[長野 08] 長野翔一, 高橋寛幸, 中川哲也: ユーザの要求変化に着目したウェブ閲覧履歴の分類方式，情報処理学会研究報告. 自然言語処理研究会報告，Vol.2008，No.90, pp.65-70 (2008)