D-008

目標と作業との紐付けによるウェブ閲覧履歴の検索支援

Making Links between Goal and Tasks to Retrieve Information from Web Browsing History

樋口 賢治 † Kenji Higuchi
原田 史子 ‡ Fumiko Harada
島川 博光 ‡ Hiromitsu Shimakawa

1. はじめに

日々のネットサーフィンにおいて，我々は多くの偶発的な情報を得ている．ネットサーフィンを続ける中でウェブブラウザの閲覧履歴は膨大な量になり，単純な履歴検索では偶発的に得た情報の再利用は困難である．

我々は目標をもって具体的な作業を実施している．ネットサーフィンにおいても，いくつかの目標をもってウェブ上での情報閲覧や値の入力をおこなっている．得られた情報によって，新たな目標や，より細かい粒度の作業が生まれこともある．もし作業ごとにウェブページを整理できれば，ある目標をもってネットサーフィンしているときに，以前同じ作業を実施していたときのウェブページを利活用できるようになる．

ウェブページのもつ特性を定量的に表現する手法として，ベクトル空間モデル [1] がある．ベクトル空間モデルでは，文書の持つ特性を文書中に出現する索引語の $tf \cdot idf$ 値をもとに多次元空間上のベクトルとして表現する．閲覧履歴を分類する手法として，長野らは，時系的な近さとウェブページ本文の類似性をもとに，要求変化に着目した閲覧履歴を分類する手法を提案した [2] ．これらの手法をもとにして，さまざまなウェブページ分類手法が提案されているが，ウェブページを作ごとに分類し，それを履歴検索に応用することは実現できていない．そこで本稿では，目標と作業との紐付けによるウェブ閲覧履歴の検索手法を提案する．

2. ゴールとタスク

ある目標を達成するためには，いくつかの実現手段としての作業を伴う．本稿では目標をゴール，実現手段としての作業をタスクとそれぞれ呼ぶ．例えば，ゴールが「就職したい」だとすると，タスクは「インターンシップに参加登録」や「志望企業にエントリ」などとなる．いま,「志望企業にエントリ」したいと思いネットサーフィンをしているとしよう．このとき,「インターンシップに参加登録」しようと閲覧したウェブページが参考になる．

多くのウェブブラウザは閲覧履歴を保存する機能を有する．これは，ネットサーフィンにおいて，過去に閲覧したウェブページがしばしば役立つためである．馬らは，ScrapBook・履歴・ブックマークをそれぞれ横断的に検索する手法を提案した [3] ．しかし，閲覧履歴とともに検索結果も膨大になり，また利用目的に応じ分類されないため，目的のウェブサイトの発見が困難になるという問題点がある．つまり，閲覧履歴よりタスクを推測し，同じゴールのための閲覧履歴の参照を容易にする仕組みがあれば，履歴検索における負荷の軽減につながり，現行のタスクに関連する閲覧履歴を利活用できるようになる．

†立命館大学大学院 理工学研究科
‡立命館大学 情報理工学部

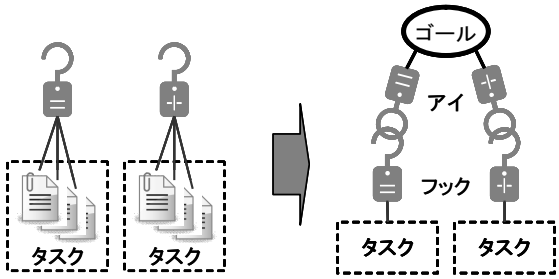

図 1: フックによるゴールとタスクの紐付け

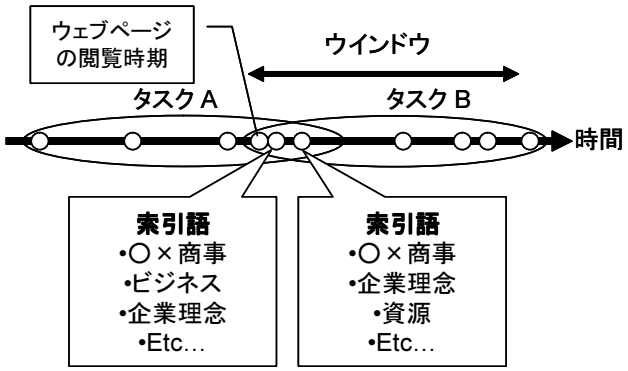

図 2: 索引語の共起と収集時期によるクラスタリング

3. 閲覧履歴の整理

3.1 ゴールとタスクとの紐付け

本稿では，ゴールとタスクとの紐付けによる閲覧履歴の検索手法を提案する．まず，図 1 左のように閲覧したウェブページをタスクごとに分類する．つぎに，タスクに紐付けのためのフックを作成し，図 1 右のようにそのフックとゴールのアイとを紐付ける．フックとは，タスクの特性を示すものであり，ひとつのタスクに複数存在することがある．履歴活用時には，現行のタスクのフックよりゴールを判別し，ゴールに紐付けられた関連するタスクの閲覧履歴を提示する．また，アイとはゴールに紐付けられるタスクを判別するものであり，ひとつのアイに複数の対応するフックが紐付けられる．これにより，膨大な履歴情報の中から，ゴールとそのタスクとに関連の強い履歴を絞り込んで検索可能となる．

3.2 タスクごとのクラスタリング

同じゴールのもとにはさまざまな性質のタスクが複数存在し，それらのタスクの中には互いに似た性質のものが存在すると考えられる．つまり，同じゴールの中でも，現行タスクと似た性質の関連タスクを優先的に参照できるようになると，現行タスクにおいて利活用できる閲覧履歴を得やすくなる．

本手法では，ウェブページの索引語の共起と収集時期とを考慮し，ウェブページをタスクごとにクラスタリングする．ネットサーフィンにおける人の要求は一定時間で移り変わる [4] ．つまり，近い時期に閲覧するウェブページは，同じ要求のタスクに分類できると考えられる．しかし，時間的な近さのみを考慮しウェブページを分類する場合，タスクが流動的に変化すると，変化の前と後のタスクの区別は困難になると考えられる．そこで本手法では，図 2 に示すように，一定サイズの時間ウィンドウを設け，ウィンドウ内におけるウェブページの索引語の共起確率によってウェブページ間の類似度を図ることで，要求の変化を判別する．

まず，ウィンドウサイズをもとにウェブサイトの閲覧時期を区切る．つぎに，ウィンドウ内における索引語の共起確率を求める．そして，このウィンドウを徐々にシフトし，高確率で共起しあう索引語が共起しにくくなる時点を，要求の変化点とみなす．この時点をもとにウェブページをタスククラスタとして分類する．

3.3 フック語による紐付け

タスクにフックを，ゴールにアイを作成し，それぞれを対応させることでゴールとタスクとを紐付けする．初期状態のゴールにはアイが存在せず，複数回にわたる手動でのゴールとタスクとの紐付けによってフックの傾向を学習し，対応するアイとして登録していく．本手法では，ゴールとタスクとの紐付けに，タスククラスタの特性を強く示す索引語を用いる．この索引語をフック語と呼ぶ．ゴールのアイには手動で紐付けられたタスクのフック語が登録されることになる．

タスククラスタの特性を強調するために，以下の 2 点の性質をもつ索引語をフック語として抽出する．

- タスククラスタ中の多くのウェブページに出現する
- 文書中の出現頻度が高い

よって，これらの性質を満たす索引語 t をフック語として抽出する指標 $h(t)$ を式 1 のように定義し，閾値を超える索引語をフック語として同定する．なお，閾値の設定によっては 1 つのタスクに複数のフック語が存在することになる．

- $D = \{d_i|$ タスククラスタに含まれるウェブページ $\}$
- $D_t = \{d_{ti}|D$ 中の索引語 t が出現するウェブページ $\}$
- $tf(t, d_i) = \{d_i$における t の tf 値 $\}$

$$h(t) = \frac{|D_t|}{|D|} \cdot \frac{\sum_{k=1}^{|D_t|} tf(t, d_{tk})}{|D_t|} = \frac{\sum_{k=1}^{|D_t|} tf(d_{tk})}{|D|} \quad (1)$$

3.4 フック語による関連タスクの抽出

現行タスクにおいて有益な閲覧履歴を得るために，現行タスクのフック語をもとにゴールを判別し，関連タスクを抽出する．まず，現行タスクにおけるタスククラスタをもとに現行タスクのフック語を抽出する．つぎに，図 3 に示すように，フック語をもとにゴールを判別し，現行タスクと同じアイで紐付けられたタスクを関連タスクとして抽出する．以上の手法により，膨大な履歴情報の関連タスクによる絞込みが可能となる．

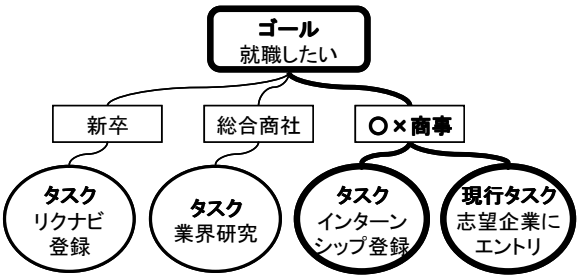

図 3: 関連タスクの抽出

4. ベクトル空間モデルによる履歴検索

本研究では，抽出された関連タスクに属するウェブページに，ベクトル空間モデルを用いることでランクを付け，一覧として提示する．これにより，検索クエリを用いることなく，現行タスクにおいて関連度の高い閲覧履歴を参照できるようになる．

まず，各ウェブページの特性に応じて多次元ベクトルを設定し，そのベクトルをもとにタスククラスタの要求をベクトルとして求める．すべてのウェブページには索引語を軸とするベクトルが設定される．ここで，タスククラスタはベクトルを持つウェブページのクラスタであるため，タスクの要求はタスククラスタの重心として表現できる．この重心をタスクベクトルと呼ぶ．同様にして，現行タスクのタスクベクトルも同定できる．以上により，ベクトル間の類似度を距離などの何らかの指標で計算し，現行タスクの要求に近いウェブページから順に表示可能となる．

ここで，検索性向上のために，ウェブページのベクトルの補正を考える．ウェブページには索引語の tf 値をもとにベクトルが設定されるものとする．これは，タスクベクトルの値をより多く出現する索引語の軸に偏らせるためである．ページのベクトル値をタスクベクトルにより近い値へと補正することで，最初に設定された tf 値によるベクトルに関係なく，現行タスクと近い要求のもとに閲覧したウェブページがランクの上位に出現するようになる．

5. おわりに

本稿では，ゴールとタスクとの紐付けによるウェブ閲覧履歴の検索手法を提案した．今後は本手法の有用性の検証を行なう予定である．

参考文献

[1] G. Salton, et al., “A vector space model for automatic indexing,” Commun. ACM, vol.18, no.11, pp.613–620, 1975.

[2] 長野ほか，“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.

[3] 馬ほか，“Web ブラウザの scrapbook・履歴・ブックマークを横断的に検索可能なツールの開発と評価,” 電子情報通信学会技術研究報告. ET, 教育工学，vol.106，no.583，pp.93-98，2007.

[4] 長野ほか，“コンテキストを変化させる閲覧履歴の抽出,” 人工知能学会全国大会 (第 22 回)，2008.