6S-2

嗜好情報と文書構造を利用したWebページの局所的フィルタリング

牛田　恵利†　　永森　光晴‡　　杉本　重雄‡

筑波大学情報学群情報メディア創成学類†　　筑波大学図書館情報メディア研究科‡

1.はじめに

Web の普及による情報量の増加によって、様々な情報に簡単にアクセスできるようになった一方で、閲覧者にとって不適当な情報を取得する可能性が高くなったともいえる。例えば、録画しておいたスポーツの試合結果を視聴する前にWeb上で偶然目にしてしまい落胆することがある。また反対に、大量の情報の中に必要な情報が埋もれてしまい、見逃しているという事もある。そこで、本研究ではWeb閲覧者の嗜好に基づいて、Webページ中から情報を選択し、適切な形で表示する手法を提案する。

2.URL フィルタリングとキーワードフィルタリング

URLフィルタリングとは､特定のURLを持つWebサイトを閲覧できないようにする技術である。しかし、URLフィルタリングでは、あるWebページを閲覧する際に、当該ページに含まれる一部の情報を閲覧者の嗜好に基づいて閲覧できないようにすることは難しい。一方、Yahoo!あんしんねっと等で用いられているキーワードフィルタリングでは､HTML文書中に不適切な言葉が含まれている場合、その言葉だけを伏せ字に変換して表示をすることが可能である。しかしながら、伏せ字にした言葉の前後の文脈から内容の判断ができてしまうという問題点がある。

本研究では、これらの問題を解決するために、Web ページの中で閲覧者の嗜好に適合する部分を局所的にフィルタリングする手法を提案する。

3.局所的フィルタリング手法

3.1 嗜好情報の記述

嗜好情報とは、Web 閲覧者が情報収集する際に持っている興味や関心を表した情報である。例えば、「サッカー日本代表の試合結果は1/28までは見たくない」や「ワールドカップ開催中は日本代表選手のインタビュー記事は見逃さないようにする」といったものである。このような嗜好情報を1)興味関心を表すキーワードリスト､2)興味関心に一致する部分をブラウザ上でどのような形で表示したいのか、3)フィルタリングを行う期間という三つの要素で表現する。本稿では、1)と2)に焦点を当てる。例えば、嗜好情報を({サッカー日本代表、試合、結果}、見たくない)というように表現する。

3.2 嗜好情報に関連する領域の特定

図1上部は、嗜好情報を表したキーワードリスト「サッカー日本代表、試合」に該当する部分のみを墨塗りした例である。図のように、墨塗りしただけでは、前後の文脈から容易に試合結果を推測できてしまう事が分かる。局所的フィルタリングを実現するためには、図1下部のように嗜好情報の該当部分に関連する領域までフィルタリングする必要がある。関連する領域とは、例えば、嗜好情報が該当した部分の段落や、見出しとその本文という内容の固まりである。本稿では、ノードが深い位置にあるほど特定の話題について書かれている[1]という文書構造の特徴を利用して関連領域を導きだす局所的フィルタリング手法を提案する。ここで扱うノードとは、HTML文書を構成する要素の事を指す。以下に関連領域を求める手順を示した。

(1) キーワードへの重みの付与

キーワードリスト中で最初に登録されたキーワードほど嗜好をよく表していると考えられ

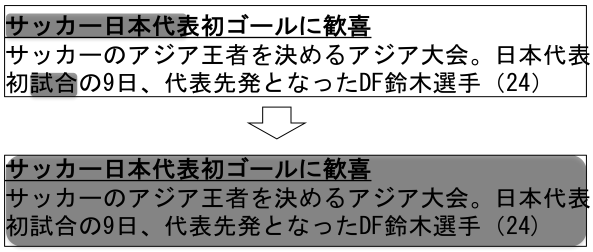


図1　フィルタリング例

" A Spot Filtering Method for HTML Documents based on its Tree Structure and User Preferences."

†Eri Ushida.School of Informations.U of Tsukuba

‡Mitsuharu Nagamori.Shigeo Sugimoto. Graduate School of Library,Information and Media Studies.U of Tsukuba.

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

る[2]。キーワードそれぞれに重要性を示す重みを付与する。

(2) 嗜好情報に該当する部分の特定

HTML 文書の中の全テキストノードの中から、キーワードリスト中のいずれかのキーワードを含んでいるノードを抜き出す。

(3) 嗜好情報に関連する領域の導出

キーワードリスト中のキーワードに該当した複数のノード同士をグループにする。ノードが深い位置にあるほど特定の話題を扱っているので、ノード同士が深く近い位置に存在するほど大きな値をとる下記の式からノードグループごとの関連度スコアを算出する。そして、最小共通祖先ノード(LCA)以下の部分木を関連領域とする。

$$score(\text{嗜好情報該当ノード、}LCA) = \frac{depth(LCA)*m}{\sum_{i=1}^{m} depth(n_i)} * \sum_{i=1}^{m} weight(n_i)$$

ここで depth(n_i)は根ノードからノード n_i までの距離を表し、LCA は最小共通祖先ノードである。weight(n_i)はノード n_i の重みを表している。

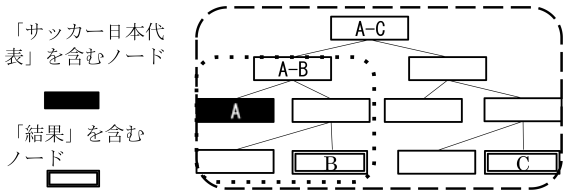
(4) フィルタリング箇所の決定

(3)のスコアに基づいて、一定以上の関連度であればフィルタリングを行う。

図 2 に関連領域の導出例を示した。

4. 評価実験

本手法に基づいたプロキシサーバを作成し、評価実験を行った。実験は、Yahoo!ニュース、Google ニュース、mixi ニュースから無作為に選んだ 23 ページを対象に、キーワードリスト({J リーグ、移籍}、見たい)({市川海老蔵、事件}、見たくない)等の 13 種類の嗜好情報を用いて行った。フィルタリングの正解箇所を予め数えておき、システムによるフィルタリング結果を人が A)正解 B)不十分 C)超過 D)不正解で評価した。



$score(AB, A-B) = \frac{2*2}{3+4}*(0.9+0.1) = 0.57$　　$score(AC, A-C) = \frac{1*2}{3+4}*(0.9+0.1) = 0.29$

図 2　関連領域導出例

表 1　実験結果

対象サイト(ページ数)	適合率(%)	再現率(%)	F 値
Yahoo!ニュース(9)	67.5	59.5	63.2
mixi ニュース(4)	88.2	57.6	69.7
Google ニュース(10)	55.5	44.6	49.5

実験結果を表 1 に示した。表 1 の結果より約 6 割の精度でフィルタリングをすることができた。一方、全てのサイトで適合率は高いが再現率は低い結果となった。これは、嗜好情報に該当する部分を完全には取得できていないためである。例えば、キーワード「J リーグ」を用いると、「日本プロサッカーリーグ」と記述されている記事は表記の違いから同義語であるにも関わらず取得できないという場合である。また、Google ニュースの適合率、再現率が低い理由は、記事を表す見出しと本文が、構造上遠い位置関係にあることが挙げられる。

5. おわりに

本稿では局所的フィルタリングを行うための手法を提案し、評価実験を行った。深く近い位置にあるノードからは、高い関連度が算出され、最小共通祖先ノードから関連領域を導き出す事ができた。また、約 6 割の精度でフィルタリングを行うことができた。今後の課題は、(1) 表記の違いを考慮した、嗜好情報に該当する部分の取得精度の向上、(2)見出しと本文といった文書構造を考慮した関連領域の取得、が挙げられる。今後、既存のオントロジーを利用し、嗜好情報の同義語や関連語を取得しキーワードリストを拡張させる事を考えている。また、文書構造上遠い位置関係にある本文と見出しに対しては、heading タグで表記される事が見出しである場合が多いので、階層を辿る事で該当するタグを探し、関連領域とすることを検討している。また、HTML 要素の表示位置情報を利用して、関連領域を導き出すことも考えている。

参考文献

[1]長尾真.岩波講座ソフトウェア科学 15 自然言語処理.岩波書店.1996.

[2]大澤快至.検索キーワードとその関連語への重みづけによる重要な Web ページの上位配置.法政大学　学術機関リポジトリ. http://rose.lib.hosei.ac.jp/dspace/handle/10114/2774?mode=simple .2008.

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.