RO-007

Web ページの分類と閲覧時間を利用したコンテンツフィルタリング

Contents filtering using the classification and reading time of Web page

大井　彩香† 寺田 実† 丸山 一貴‡
Ayaka Oi Minoru Terada Kazutaka Maruyama

1　背景

近年 Web によって提供される情報量は増加し, 誰もが容易に様々な情報を取得することができる. 反面, 教育上不適切な情報の氾濫, ネット依存による Web の長時間利用などの問題がある. 文部科学省の調べでも, インターネットに氾濫する様々な有害情報は, 価値観やモラルにも悪影響を及ぼすおそれがあり, 生活上の支障を引き起こすネット中毒ないし依存症の事例もあると述べている*1.

氾濫する情報の中には暴力描写, 出会い系サイトなど子供の教育上, 全面的に遮断を推奨されているサイトがある一方, 長時間の利用を避けるべきだけのサイトがある. 本研究ではこのように Web には制限されるべきページとして遮断すべきサイトとそうでないサイトの 2 種類があることに着目する.

2　目的

Web の閲覧に制限をかけるシステムとしてフィルタリングソフトが存在する. これは主に不適切な情報へのアクセスの制限, 長時間の利用を防止するためにサイト閲覧の時間制限を目的としている. 主な機能は以下の 3 つである.

- カテゴリごとの制限
  データベースに登録されている URL とそれに対応するカテゴリを用いる. 閲覧を禁止するカテゴリと禁止しないカテゴリを選択し, 閲覧を制限する. なお, データベースはソフト提供側の目視作業でページが分類及び更新される.
- ホワイトリストとブラックリストによる制限
  ホワイトリストに閲覧を許可する URL, ブラックリストに閲覧を禁止する URL をソフトを実際に利用する管理者があらかじめ個別に登録し, それぞれのリストを用いて閲覧を制限する.
- 時間制限
  Web ブラウザを開ける時間を曜日と時間帯, もしくは 1 日の利用時間で制限する.

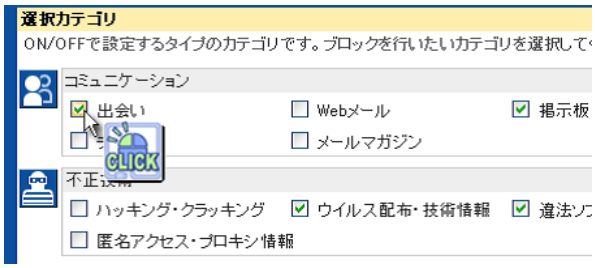

図 1　カテゴリ選択によるフィルタリング*2

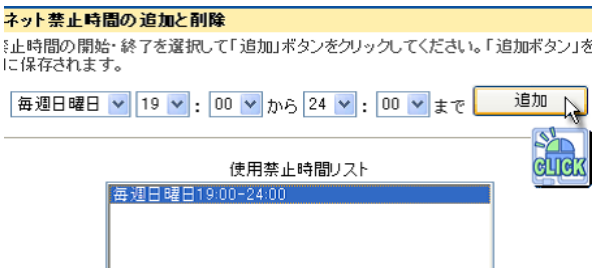

図 2　閲覧の禁止時間*2

しかし関連システムでは有害情報の遮断を特に重視しているため, 次のような問題点があげられる.

- カテゴリ制限の場合
  カテゴリごとの完全遮断行うことになり, エンターテイメントなど一概に有害とは言えない情報の完全遮断は過度な規制となり得る.
- 時間制限の場合
  Web 全体の利用時間で行われ細かい指定は出来ない. この場合時間がどのように使われているかが問題になる.

そこで本研究では Web 全体の利用時間ではなくカテゴリごとの閲覧時間について着目する.

カテゴリごとの閲覧時間に応じてアクセス制限をかけることで過度な規制を防ぎつつ有効的な時間制限を行い, Web の長時間の利用防止を目的とする. さらに, 閲覧状況を可視化することで閲覧意識の自発的な改善・抑止の可能性を追求する. これにより本来の目的とは関係のない Web ページを過多に閲覧することを防止する.

† 電気通信大学,The University of Electro-Communications
‡ 東京大学 情報基盤センター, Information Technology Center,The University of Tokyo

*1 インターネットの有用性と危険性, http://www.mext.go.jp/a_menu/sports/ikusei/030301b.htm.

*2 i-フィルター Active Edition, http://www.nifty.com/webfilter/help/index.html.

3　関連研究

3.1　ページ分類に関する研究

日々増え続ける Web ページに対応するにはページ分類は自動かつ短時間で出来ることが望ましい．Web ページを自動分類する方法は主に次の 3 種類である．

- ページ内の単語解析による分類 [1][2]
  ページ内から名詞を抽出および解析を行うことでページを分類する方法．
- リンク構造による分類 [3][4]
  リンク元ページを利用してページを分類する方法．
- 協調フィルタリングを用いた分類 [5]
  固定した興味を持つ仮想ユーザに対して協調フィルタリングによる推薦を行うことでページを分類する方法．

3.2　コンテンツフィルタリングに関する研究

コンテンツフィルタリングに関する研究については主にレイティングデータの作成に関するものと遮断するデータに関するものの 2 種類がある．

- レイティングデータを利用したフィルタリング [6]
  テキストを解析し，有害情報の候補を有害情報にかかわる単語や文書を用いて効率的に検出する研究．検出された有害情報を目視作業で確認し，レイティングデータを作成する．
  西埜らの Harmful content block-Prototype(HCB-P) はこのレイティングデータを使い Microsoft Internet Explorer 5.0(IE5.0) に組み込まれているフィルタリング機能を利用して有害サイトを遮断する．
- 流通ポリシに基づいたフィルタリング [7]
  情報通信におけるセキュリティの向上を図る研究．予め定めておいた流通ポリシ情報に基づき流通の可否を決めてゲートウェイ装置で制御する．なお流通ポリシとはセキュリティレベルや実行アクション (情報の転送，廃棄など) の対からなるフィルタリングルールのリストである．

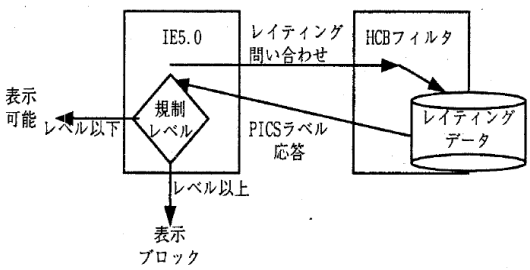

図 3　IE5.0 と HCB-P フィルタのインタフェイス [6]

4　提案手法

4.1　概要

Web ページ内の単語解析による分類を行い，カテゴリごとに閲覧時間を記録する．そしてカテゴリごとの完全遮断ではなく，あるカテゴリに対し一定時間以上閲覧した場合にアクセスを制限する．またカテゴリごとの閲覧状況を可視化し，閲覧意識の自発的な改善・抑止効果を期待する．

4.2　Web ページの分類

Web ページを閲覧するたびにそのページの分類を行う．カテゴリごとに関連する用語をキーワードとして予め設定しておき，ページ内に出現したキーワードの重要度をそれぞれ計算し，カテゴリごとに合計を求める．その合計が最も高いカテゴリに分類する．キーワードには Yahoo!カテゴリ[*3]に登録されたカテゴリ名を用いる．

4.2.1　tf-idf

単語の出現頻度 tf と逆出現頻度 idf の 2 つの指標を用いて文章中の単語の重要度を計算するアルゴリズムである．ページ内に出現したキーワードの重要度を計算する．

$$tfidf = tf \cdot idf \tag{1}$$

4.3　アクセス制限

カテゴリごとの閲覧時間によって制限をかける．カテゴリごとに閲覧時間の上限を管理者が指定し，あるカテゴリに対し一定時間以上の閲覧した場合にフィルタリングによりアクセス制限をかけて閲覧を規制する．閲覧時間はブラウザの該当ウィンドウがアクティブになっている時間とする．

4.4　可視化

閲覧時間をカテゴリごとにグラフ化し，現在の閲覧状況を常に表示させておく．閲覧状況を常に視覚的にとらえることで閲覧の抑止効果を期待する．

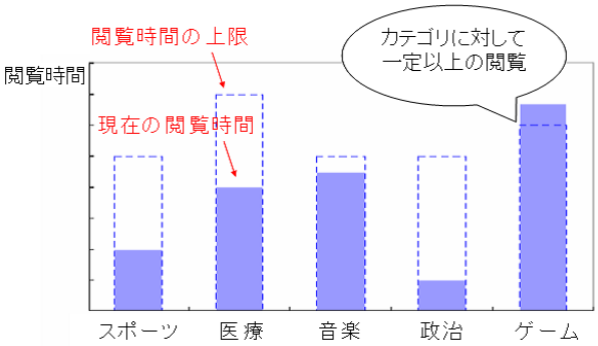

図 4　カテゴリごとの閲覧時間による制限

[*3] Yahoo!カテゴリ, http://dir.yahoo.co.jp/.

5 実装

5.1 概要

ページ分類とコンテンツフィルタリングを行うためにプロキシサーバを使用する．プロキシサーバにおいて分類と閲覧制限に達した場合のページの書き換えを行い，ブラウザは分類及び書き換え後の Web ページを表示する．カテゴリごとの閲覧時間の計測は別スレッドにおいて行い，ページ分類後に今までの閲覧時間をプロキシサーバが取得する．この取得したカテゴリごとの閲覧時間に応じて閲覧の制限をかける．なお，1 つのマシンで完結させるためにプロキシサーバは同一の PC 上で構築する．

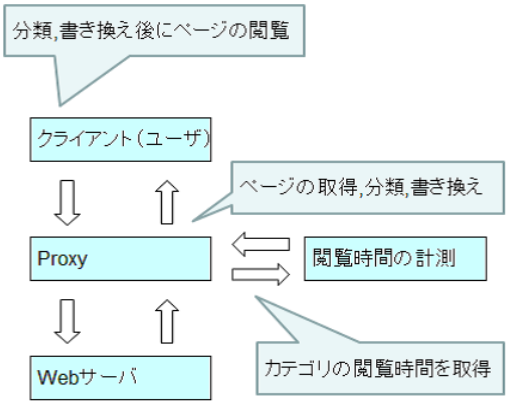

図 5 構成図

5.2 ページ内の単語解析による分類

日々増え続ける Web ページに対応するにはページ分類は自動かつ短時間で出来ることが望ましい．そこで本研究では分類のためのデータベースは使用せず，Web ページの分類を行う．ページ内に出現したキーワードの重要度を用いて閲覧したページを分類する．なお分類するカテゴリはゲーム，政治，医療，音楽，スポーツの 5 つとする．

1. 分類する Web ページから単語を抽出し，ページ内に含まれるカテゴリのキーワードの tf-idf 値を求める．
2. ページ内の単語の数が 10 個以下だった場合，もしくはどのカテゴリともマッチしなかった場合はページ内の単語数が少ないとみなして，分類するページのリンク元のページの単語を足して再度分類を行う．
3. tf-idf 値をカテゴリごとに加算していく
4. tf-idf 値の合計が最も高いカテゴリに分類する．

リンク元のページは Yahoo!JAPAN の link コマンド[*2]を使用して検索する．分類にかける時間を短くするためにリンク元のページを参照するのは最高 5 ページとした．それでも分類できなかった場合は未分類とする．

5.3 閲覧時間の計測

閲覧時間は別スレッドにて WindowsAPI を用いてウィンドウを常時監視し，ブラウザの該当ウィンドウ がアクティブになっている時間を閲覧時間として計測し，カテゴリごとに時間を蓄積しておく．アクティブになっているウィンドウのタイトルを WindowsAPI で取得し，分類後にタイトルに付加したカテゴリ名 (図 6 丸枠) からどのカテゴリを閲覧しているのか判断し，その時間を計測する．

図 6 ページタイトル名の後に分類結果のカテゴリ名を付加

5.4 コンテンツフィルタリング

プロキシサーバを構築し，プロキシサーバを通ったデータを取得し変換することでコンテンツフィルタリングを行う．まず制限時間を予め設定する．プロキシサーバを通して取得したデータの内，コンテンツタイプが html のものを取得し，第 5.2 節の方法で閲覧ページの分類を行う．分類されたカテゴリの閲覧時間を別スレッドから取得し，上限に達しているかどうかを調べ，上限に達していた場合には書き換えたページをユーザに送信して閲覧を制限する．

なおプロキシサーバの構築には Webrick [*3] を使用する．Webrick は HTTP サーバの機能を提供する Ruby のライブラリである．

図 7 game カテゴリの閲覧上限に達したときの画面

5.5 閲覧状況のフィードバック

閲覧状況を視覚的にとらえることでユーザの自発的な閲覧意識の改善・抑止を図るため，カテゴリごとの閲覧時間を可視

[*2] Yahoo!JAPAN での検索時のコマンドの 1 つ．指定したページにリンクするページを検索できる．

[*3] webrick - Ruby リファレンスマニュアル，http://www.ruby-lang.org/ja/man/html/webrick.html.

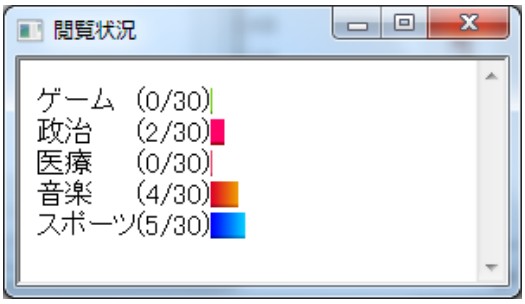

図 8　閲覧時間のグラフ化によるフィードバック（分単位）

化する．通常の閲覧を阻害しないために，邪魔にならない程度の小さいグラフをディスプレイの右下に常に表示するようにする．

6　評価実験

被験者 10 名に 10 分間自由に Web 閲覧を行ってもらい，システムの Web ページの分類率と総合的なアンケートにより提案手法を評価した．

6.1　分類結果

被験者には自分が閲覧したページをゲーム，政治，医療，音楽，スポーツ，その他 (未分類) のいずれかに分類してもらった．この結果を用いて式 (2) より分類率を計算したところ，表 1 を得た．

$$\text{分類率} = \frac{\text{システムが被験者と同じカテゴリに分類できた数 A}}{\text{被験者に分類してもらった数 B}} \tag{2}$$

表 1　カテゴリごとの分類率

カテゴリ	ゲーム	政治	医療	音楽	スポーツ	未分類
A	12	16	16	10	14	3
B	22	20	19	17	14	16
分類率 [%]	54.5	80.0	84.2	58.8	100.0	18.8

ゲームカテゴリと音楽カテゴリはあまり高い結果を得られなかった．原因としては，文字が少なく画像が多いことから単語解析による分類がうまくいかなかったと考えられる．これを改善するためにリンク元 URL を利用したが，まだ不十分であったと言える．また未分類ページの分類率が低いのは，５つのカテゴリ以外のページを閲覧しても無理やり５つのどれかに分類してしまうことが考えられる．

6.2　アンケート結果

提案手法について有効性等の意見を求めたところ，次のような意見が上げられた．

既存フィルタリングソフトと比較したときの本システムの有効性を尋ねた結果，「自由度の高いブラウジングが出来て良い」，「ゲームやスポーツなど一部のカテゴリには時間制限をする方法は有効」などの意見が出た．

また改善すべき点には，処理速度（5〜30 秒程度）の向上，分類精度の向上，適切なカテゴリ選択と時間の選択を支援，などが上げられた．他にも「自分の閲覧状況を把握するのにグラフは有効である」との意見がある一方，「画面端は普段見ないのでグラフは目につかない」との意見も出た．また複数のカテゴリに属するページの対応についての検討についても意見があった．

7　結論

7.1　結論

カテゴリごとの閲覧時間による制限は柔軟なフィルタリングを可能にし，完全遮断が必要なカテゴリにも対応できることから，カテゴリごとの時間制限は有効である．また，分類精度が上がればカテゴリごとの時間制限はより有効である．

閲覧状況のフィードバックに関して，グラフ化するだけでは閲覧状況を意識するには不十分であり，自発的改善のためには情報の提示方法が重要との知見を得た．

7.2　今後の課題

分類精度と処理速度の向上を図る．他の手法での分類を利用することを検討し，通常の閲覧を阻害しないためにも特に処理速度の向上は必須であると考える．またグラフ表示について，閲覧の邪魔にならない程度に常に目にできる位置に表示するために簡易的な表示方法を取ったが，閲覧意識を改善を図るために，視覚的に面白い表示方法や通常自分が意識しない情報の視覚化，閲覧状況の可聴化についても検討する．

参考文献

[1] 金村和美　他，“ Web ページの自動分類に関する一手法 ”，電子情報通信学会 OIS, pp.25-30, 2004.

[2] 土方嘉徳　他，“ アンカー関連テキストを用いた Web ページ分類方式の実装と評価 ”，第 21 回人工知能学会全国大会, 2007.

[3] 大西高裕　他，“ リンク構造を用いた Web ページ自動分類の精度向上法 ”，情報処理学会第 65 回全国大会, pp.3-161〜3-162, 2003.

[4] 山下 長義　他，“ リンク構造に基づいた WWW からのトピック抽出 ”，情報処理学会 MPS, pp.9-14,2008.

[5] 伊藤真宏，“ 協調フィルタリングを用いた Web ページの分類と推薦 ” 電気通信大学平成 20 年度卒業論文.

[6] 西埜 覚　他，“ 学校教育における有害な情報の検出とレイティング，フィルタリング技術の開発 ”，情報処理学会研究報告 コンピュータと教育研究会報告,pp.39-46,2001.

[7] 松岡 直樹　他，“ コンテンツプロファイルに基づくコンテンツ流通管理システムの開発 ”，電子情報通信学会技術研究報告 CS,pp25-30,2008.