イベント追跡システムにおけるモニタリング対象ページの選別

神野　雅宏　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{mjinno, kshirai}@jaist.ac.jp

1 はじめに

1.1 背景

本研究の目的はウェブにおけるイベント追跡システムの構築である．イベント追跡システムとは，特定の商品，映画，音楽作品などについて，それに関する情報がウェブ上で発信されたかどうかを定期的に観測 (モニタリング) し，ユーザに通知するシステムを指す．例えば，映画の「ハリーポッター」の新作が作られるというニュースを聞いたとき，今後発生するであろう配役の発表，試写会，封切のイベントをいちはやく知りたいと思うことがよくある．このとき，イベント追跡システムは，ウェブを定期的に観測し,「ハリーポッター」の新作に関する新たなイベントが発生すれば，ユーザに自動的に通知する．ユーザは「ハリーポッター」の情報を得るために毎日ウェブを見続けることなく，知りたい情報を迅速に得ることができる．

ウェブページを定期的に観測し，更新をチェックするアプリケーションは数多く開発されているが [1, 2]，そのほとんどがページの更新の検出のみを対象とし，更新の内容については考慮していない．また，山田らはページ全体の更新ではなく，ページのある特定の部分が更新されているかどうかをチェックする PUM システムを提案している [4]．ただし，ページのどの部分の更新をチェックするかはあらかじめユーザが指定する必要がある．これに対し，本研究におけるイベント追跡システムは，モニタリングの対象となるウェブページを自動的に決定する点，ユーザが関心のある更新のみを通知する点に特徴がある．また，杉本らは同様のイベント追跡システムを構築することを前提とし，イベントを表わす語をウェブから自動的に獲得する手法を提案している [3]．

1.2 目的

本研究におけるイベント追跡システムの特徴のひとつは，モニタリングの対象となるページをあらかじめユーザが指定するのではなく，システムが自動的に決定することである．以下，システムが定期観測を行うページのことをモニタリング対象ページと呼ぶ．本論文では，イベント追跡システムに必要な要素技術のうち，ユーザの要求に合わせてモニタリング対象ページを決定する手法について主に述べる．

以下，2 節ではモニタリング対象ページを決定する具体的な手法について述べる．3 節では 2 節の手法の評価実験について述べる．4 節では，ページが更新されたとき，ユーザの知りたいイベントに関する情報が含まれているかを判定する手法について述べる．

2 モニタリング対象ページの選別

本研究におけるイベント追跡の最初の処理は，モニタリング対象ページを収集することである．ここで,「活性ページ」または「不活性ページ」という用語を以下のように定義する．

活性ページ　頻繁に更新され，新しい情報がよく掲載されるウェブページ

不活性ページ　更新があまり行われず，同じ情報を長期間掲載するウェブページ

イベント追跡に有効なのは活性ページであることは明らかであろう．したがって，ウェブページが活性ページであるか不活性ページであるかを自動的に判定し，活性ページのみをモニタリング対象ページとする．

活性ページは大きく分けて以下の 2 つのタイプがある．

1. ページ内追加型
   新しい情報を同一ページ内に追記するウェブページ．
2. リンク型
   図 1 のようにリンク元の親ページ (P) とリンク先の子ページ (C_i) があり，P 内のリンクとそのリンク先ページ C_i を追加または更新するウェブページ．このとき，親ページ P を活性ページとみなす．新しい情報は C_i にあり，また C_i 自体は更新されない不活性ページであることに注意しなければならない．ニュースサイトのトップページなどが該当する．

モニタリング対象ページを選別する際，これらの活性ページのタイプの違いを考慮する必要がある．

2.1 候補ページの収集

まず，ユーザの知りたいイベントが発生しそうなウェブページを収集し，モニタリング対象ページの候補の集

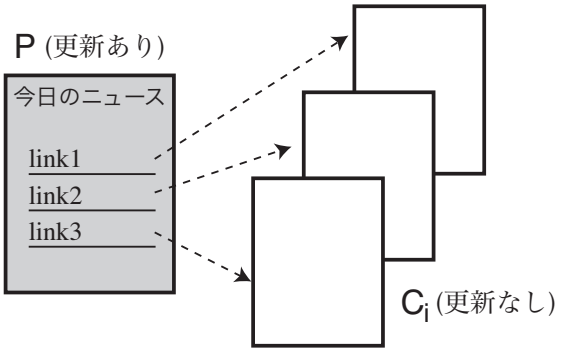

図 1: リンク型活性ページ

合 M' とする．本論文では，ユーザが知りたいイベントは以下の 3 種類のキーワードで表現されるとし，これらをイベント追跡システムに対する入力とみなす．

- 対象語
  イベントを追跡する対象を表わす一語．例えば「ハリーポッター」など．
- 分野語
  対象語が属する大まかなドメインを表わす一語．例えば「映画」など．
- イベント語
  ユーザに通知するべきイベントを表わす語．複数のキーワードを指定できる．例えば,「配役発表」「試写会」「封切」など．

既存の検索エンジンを利用し，上記のキーワードを含むウェブページを収集する．具体的には，検索エンジン goo[1] に以下のクエリーを入力する．

対象語 and 分野語 and (イベント語$_1$ or $\cdots$ or イベント語$_n$)

得られた検索結果の上位 N 件のページ集合をモニタリング対象ページの候補とする．3 節の実験では $N = 50$ とした．

ここで，リンク型の活性ページに対する注意が必要である．上記のようなキーワード検索を行った場合，リンク型活性ページの子ページ (図 1 の C_i) が検索結果として得られる場合がある．このページはユーザの知りたい情報を含む可能性が高いが，そのページ自体は更新される可能性は低い．したがって，このようなページはモニタリングの対象とはしない方がよい．むしろ，そのページにリンクを貼っている親ページ (図 1 の P) の方が頻繁に更新される可能性が高く，モニタリング対象ページの候補に加えるべきである．特に，親ページからのリンクが同一サイト内のページへのリンクであれば，その親ページはリンク型活性ページである可能性が高い．そこで，検索エンジンで得られたページがリンク型活性ページの子ページである可能性を考慮し，以下の操作を行う．検索の結果得られたウェブページの URL のパスが表わすファイルまたはディレクトリを 1 つまたは 2 つ除去する．その URL に実際のページが存在するなら，そのページをモニタリング対象ページの候補 M' に加える．例えば，以下の (a) の URL に対し，(b) や (c) の URL に対応するウェブページの有無をチェックする．

(a) http://www.aaa.com/bbb/ccc/index.html
(b) http://www.aaa.com/bbb/ccc/
(c) http://www.aaa.com/bbb/

2.2 活性ページのフィルタリング

2 節の冒頭で述べたように，モニタリング対象ページとしては不活性ページよりも活性ページの方がふさわしい．そこで，モニタリング対象ページの候補 M' の中から，活性ページを選別するフィルタリングを行い，最終的なモニタリング対象ページの集合 M を得る．あるページが活性ページか不活性ページかは，実際にモニタリングを行うことによって簡単に判別できる．しかし，あまりに多くのウェブページに対してモニタリングを行うことは効率が悪い．ここでは，モニタリングの効率化を図るため，実際にモニタリングを始める前に，ウェブページから得られる情報だけで活性ページか不活性ページかの判定を行う．

本論文における活性ページのフィルタリング手法を以下に示す．M' 中の各ウェブページに対して，次の 1. から 4. の順に活性ページまたは不活性ページの判定を行う．また，判定ができた段階で手続きを終了し，以後の判定処理は行わない．

1. タイムスタンプ
   ウェブページの最終更新時間 (タイムスタンプ) が得られる場合，最終更新時間がある一定の期間以内 (本論文では 2 日とした) なら，モニタリング対象ページとする．
2. URL の日付表現
   ウェブページの URL に日付を表わす文字列が含まれる場合がある．例えば，以下の URL の場合，“050207” は 2005 年 2 月 7 日を表わすとみなせる．

   http://www.aaa.com/bbb/ccc/050207.html

   このような URL はリンク型活性ページの子ページによく見られる．すなわち，新しい情報は日付を

[1] **http://www.goo.ne.jp/**

表 1: 実験に用いたイベント追跡タスク

イベント追跡の目的	分野語	対象語	イベント語
(a) 上原浩治の動向: メジャー移籍か残留か?	野球	上原浩治	移籍, 残留, 参加
(b) 中村俊輔が試合出場や得点	サッカー	中村俊輔	先発, 出場, 得点
(c) 小泉首相の行動	政治	小泉首相	発言, 辞任, 参拝
(d) トヨタの新車発売	車	トヨタ	発表, 発売, 販売
(e) 楽天イーグルスの活動	野球	楽天	獲得, 発表, 開始

表 2: 実験結果

	対象ページ	活性ページ			適合ページ		
	ページ数	ページ数	精度	誤り率	ページ数	精度	誤り率
(a) 上原	69 (121)	61 (90)	0.88 (0.74)	0.32	17 (23)	0.25 (0.19)	0.26
(b) 中村	62 (121)	58 (77)	0.93 (0.64)	0.25	36 (43)	0.58 (0.36)	0.16
(c) 小泉	58 (103)	49 (68)	0.84 (0.66)	0.28	15 (18)	0.27 (0.17)	0.17
(d) トヨタ	35 (115)	31 (74)	0.89 (0.64)	0.58	6 (11)	0.17 (0.10)	0.45
(e) 楽天	64 (100)	56 (80)	0.88 (0.80)	0.30	10 (13)	0.16 (0.13)	0.23
合計	288 (560)	255 (389)	0.89 (0.69)	0.34	84 (108)	0.29 (0.19)	0.22

ファイル名とした HTML ファイルに掲載し，そのページへのリンクを新たに貼ることによって新情報を発信する．したがって，このような URL を持つページ自体は今後更新される可能性が低いので，不活性ページとみなしてモニタリング対象ページの候補から除去する．ここでは，URL 内に現われうる日付表現のパターンをいくつか用意し，そのパターンにマッチするかどうかを調べる．

3. 更新を示唆する表現
   ウェブページ内に更新を示唆する表現があれば，そのページは活性ページであるとみなしてモニタリング対象ページとする．ここでは，更新を示唆する表現として以下の 8 種類の表現を用意した．

   最終更新∗，更新∗，what's new，last update，
   ∗一覧，∗ニュース，∗トピック，∗トピックス

   ∗ は任意の名詞を表わす．例えば,「最終更新日」や「社会記事一覧」なども更新を示唆する表現とみなす．
4. リンクのリスト
   以下のようなリストの存在を調べる．

   リンク 1 (*time*)
   リンク 2 (*time*)
   リンク 3 (*time*)

   time は「13:01」や「2/13」のような時刻や日付を示唆する文字列である．同一の型の時刻または日付表現を伴なうリンクが 3 回以上現われるなら，そのページはリンク型の活性ページであるとみなしてモニタリング対象ページとする．

なお，以上の処理で判定できないウェブページは全て不活性ページとし，M' から除去する．

3 評価実験

本節では，モニタリング対象ページの選別手法，特に 2.2 項で述べた活性ページのフィルタリングを評価する実験について述べる．まず，表 1 のように，追跡するイベントとして 5 つのタスクを設定し，分野語，対象語，イベント語を決めた．次に，検索エンジンで検索されたページと，そのページの URL のパスを削除して得られたページの集合を求め，モニタリング対象ページの候補 M' とした (2.1 項)．さらに，M' に対してフィルタリングを行い，最終的なモニタリング対象ページの集合 M を求めた (2.2 項)．

実験結果を表 2 に示す．「対象ページ」は提案手法で選別されたモニタリング対象ページを,「活性ページ」は活性ページを,「適合ページ」はイベント追跡を行った際に有用な情報が得られる見込みがあるページを指す．今回の実験では，モニタリング対象ページを選別してから一週間以内に更新があったページを活性ページとみなした．また，適合ページは人手で判断した．一方,「ページ数」はそれぞれのページの数を,「精度」はモニタリング対象ページの総数における活性ページまたは適合ページの割合を,「誤り率」は M' のうちフィルタリングによっ

て誤って削除された活性ページまたは適合ページの割合である．なお，( ) 内はフィルタリングを行う前の段階 (M') の値を示している．

表 2 から，フィルタリングによってモニタリング対象ページの数を約半分に減らすことができたことがわかる．また，対象ページ内における活性ページの割合 (精度) も，5 つのタスクの平均で 0.69 から 0.88 に向上した．これにより，本論文で提案する活性ページのフィルタリングがイベント追跡システムの効率化に有効であることが確認できた．一方，誤り率を見ると，活性ページが誤って削除される割合は約 35%であり，改善の余地がある．但し，適合ページに関していえば，誤り率は約 22%と低くなる．提案したフィルタリング手法はページの内容について考慮していないので，適合ページの誤り率が活性ページの誤り率よりも低いのは偶然でしかない．しかし，頻繁に更新されるページは常に新しい情報を発信しているため，イベント追跡システムにとって有効なページとなりやすい傾向がある．したがって，活性ページのフィルタリングは，イベント追跡システムに有効な適合ページの選別を間接的に行っているともいえる．

4 イベント発生の検出

我々は，モニタリング対象ページを定期的に観測し，ユーザの知りたいイベントの発生を検出する手法についても研究をすすめている．本節ではその概要について述べる．

まず，モニタリング対象ページを毎日観察し，更新の有無をチェックする．更新の有無は，ウェブページ自体をログとして保存し，その差分を検出することにより行う．具体的には UNIX の diff コマンドを用いる．次に，得られた差分のテキスト情報を解析し，イベントの発生があるかどうかを判定する．イベント発生の検出手法の概略は以下の通りである．

1. モニタリング対象ページがページ内追加型活性ページのとき，差分に含まれるテキストを文に分割する．一文の中に対象語とイベント語の両方が含まれていれば，イベントが発生したとみなす．但し，以下の場合は除く．
   - イベント語がサ変名詞で，かつ複合名詞の一部になっているとき．例えば,『発表する』がイベント語のとき，差分の中に「発表予定」という複合名詞があってもそれは『発表する』というイベントが発生したとはみなさない．
   - イベント語の後に「か?」「だろう」のような推測を示唆する表現や,「ない」「ぬ」などの否定を表わす表現があるとき．
2. モニタリング対象ページがリンク型活性ページであり，かつ差分の中にリンクがあるとき，以下の処理を行う．
   - リンクのアンカーテキスト内に対象語とイベント語の両方が含まれていれば，1. と同様に判定する．
   - リンクのアンカーテキスト内に対象語またはイベント語のどちらか一方があれば，リンク先のページについて 1. の判定を行う．この場合，差分ではなくページ全体を解析の対象とする．

上記の手法を実装し，表 1 に挙げた 5 つを始めとするいくつかのタスクについてイベント発生の検出を行い，評価する実験を行った．しかし，現在のイベント発生の検出手法はナイーブであり，ユーザの知りたいイベントの発生をうまく検出できない場合も多かった．今後，実験結果のエラー分析などを行い，イベント発生の検出手法の改良を行う予定である．

5 おわりに

本研究では，ウェブページを定期的に観測し，ユーザの知りたいイベントを発見して通知するイベント追跡システムについて述べ，定期観測を行うウェブページを自動的に獲得する手法について述べた．今後は，4 節で述べたように，ページの更新が行われたとき，ユーザの知りたいイベントの情報が新たに掲載されたかどうかを自動的に判定する手法を確立したい．

参考文献

[1] `http://soft.macfeeling.com/WebPatrol.html`.

[2] Santi Saeyor and Mitsuru Ishizuka. WebBeholder: A source of community interests and trends based on cooperative change monitoring service on the web. In *2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation (IECON-2000)*, pp. 1656–1661, 2000.

[3] 杉本浩和. Web 上でのイベントに注目した情報の自動追跡のための知識獲得. 修士論文, 北陸先端科学技術大学院大学, 2004.

[4] 山田誠二, 中井有紀. 対話的分類学習による Web ページの部分更新モニタリング. 人工知能学会論文誌, Vol. 17, No. 5, pp. 614–621, 2002.