ブログ記事自動抽出方式の検討

佐藤　和紀†　金井　敦†
†法政大学大学院　工学研究科

1．はじめに

近年，インターネット普及の飛躍的な高まりにともなって，個人情報の流出が大きな問題となっている．特に，個人情報保護法の施行を契機として，個人情報の保護に対する意識が社会的にも高くなり，企業における情報漏えい防止のための対策や技術の開発が行われている[1][2]．また，ブログやソーシャルネットワークサービス（ＳＮＳ）などのＣＧＭの利用も急速に広がり個人が自分の情報を発信することがごく普通の行動となってきている．同時に，ブログでは過失による個人情報露出について様々な問題が発生してきている．我々は安心かつ安全にインターネットを利用できるようにするためには，まずどの程度個人に関する情報が露出しているかを測定する技術が必要になってくると考えた．しかし，ブログの記述状況については把握できていないというのが実情であり，状況把握が望まれている．そこでブログにおける個人情報の露出状況を自動評価するための予備的な調査として，実際のブログ記事において具体的にどのような情報がどの程度露出しているのかを人手により調査した[3][4]．しかし状況を把握する場合，ブログはひとつひとつの記事が複数のページに分散しており全部を一度に見難い構成となっているため，予備調査の段階で人手による調査は非効率的で現実的ではないという課題が生じた．

そこでブログ記事を自動的に抽出し必要なデータのみをまとめることで，個人に関する情報の露出状況を効率的に行い，さらにその後の露出状況を自動評価する際に作業を円滑に進められるようにするための方式を検討した．本稿では，そのブログ記事の自動抽出方式について述べる．

2．ブログの構造

予備調査において，我々はオープンブログの YAHOO!ブログ[5]，goo ブログ[6]やＳＮＳの mixi[7]を対象に調査を行った．その結果，オープンブログに関しては多くの場合が同様の構造をしていたことがわかった．なお，ＳＮＳに関してはアカウントが必要なことと，制限を掛けられていて記事が見られないページがあることから今回は対象から除外した．ブログの構造イメージを図１に示す．

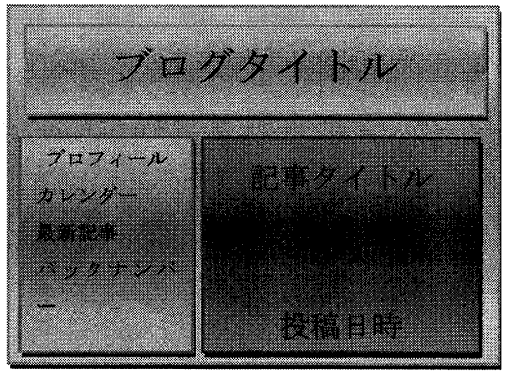

図１　ブログの構造イメージ

図１の様にブログは２段組み，または３段組みになっている．ブログ記事が表示されるメインエリアと，その他サブ情報が表示されるサイドエリア，ブログタイトルが表示されるブログ上部エリアの箇所に分類される．

どのブログにおいても共通で表示される情報には主に以下の項目が挙げられる．

・ブログ上部
　ブログタイトル
・メインエリア
　記事本文，記事タイトル，記事投稿日時，コメント，トラックバック
・サイドエリア
　プロフィール，カレンダー，最新記事，最新コメント，最新トラックバック，バックナンバー，カテゴリ

その他の情報は各サイト，ブログごとに異なる．

3．　目標

オープンブログに関しては調査した結果，個人に関する情報は主に記事本文で露出していたことがわかった．よってブログ記事の自動抽出をする際に，必要だと考えた項目はブログタイトル，記事本文，記事タイトル，記事の投稿日時，さらに基本情報としてプロフィールが挙げられる．

A study of method to extract blog information

Kazuki SATOH† and Atsushi KANAI†
†Graduate School of Engineering, Hosei University

これらのことから，まず対象ブログのＵＲＬを指定しプロフィール，記事タイトル，記事本文，記事投稿日時の情報のみを抽出する．そして，人手により調査していて生じた複数の記事を一度に閲覧する際の見難さを解決するためにそれらの情報を日付順に並べて表示させることを本方式の目標とする．

４．抽出方式

４．１　対象ブログ

本検討において記事の自動抽出を行う実験対象ブログは，gooブログとした．

４．２　gooブログ

前述したとおり，本方式においてブログから抽出する情報は記事タイトル，記事本文，記事投稿日時である．gooブログにおいて上記の項目はそれぞれ決められたHTMLのタグ内に記述されていることがわかった．まずソースコードから，各項目がどの様な場所に記述されているかを調査した．表１に各項目と，それらが記載されているタグを示す．

表１　抽出項目と記述場所

項目名	記載場所
ブログタイトル	<!-- TITLE BANNER -->内 ブログタイトル
記事タイトル	<!-- entry-top --> <a href="記事 URL">記事タイトル<a>
投稿日時	 投稿日時
記事本文	<!-- entry body --> 本文

４．３　抽出方法

抽出は，表１に示したタグを指定し各情報を抽出していく．基本的にブログのトップページには最新の記事が表示される。そこから過去の記事を辿る方法は以下の通りである．

各記事のページ最下部には“前の記事へ”と“次の記事へ”というリンクがあり，この部分のソースは次の記事へ となっているため，このリンクから過去の記事を辿ることとする．

また，本方式の検討には言語としてＰＨＰ５を使用して行った．

５．結果とまとめ

本稿では，ブログからブログ記事の自動抽出方式について検討した．検討結果としてプロフィール，およびブログ記事本文を抽出するという方式については抽出することができた．

今回，過去の記事を辿る方法として“次の記事へ”という部分を利用した．しかしこれは，最初に指定する記事を最新のものにする必要があることや，記事の件数を指定することになる．いつの記事を指定しても抽出できるようにすることや，日付指定をできるようにするにはカレンダーから過去の記事を辿る方法だと可能だと考えた．よってカレンファーから過去の記事を辿れるようにすることが課題となる．

さらに，最終的には全てのブログの抽出を行えるようにすることが目的なのでgooブログ以外でも抽出が行えるようにすることが今後の目標である．

６．参考文献

[1] NRI セキュアテクノロジーズ株式会社，“情報セキュリティに関するインターネット利用者意識2006” http://www.nri-secure.co.jp/news/2007/pdf/vol3-1.pdf

[2] 谷本茂明，廣田啓一，山本太郎，千田浩司，畑島隆，高橋克巳，金井敦，“次世代プライバシ保護サービスのコンセプト提案” 情報処理学会論文誌，Vol. 49, No. 7, pp. 2440-2455 (July 2008)

[3] 佐藤　和紀，安井　良介，針谷　友彰，金井　敦，廣田　啓一，谷本　茂明，“ブログにおける個人情報漏えいの状況調査”，情報処理学会研究報告，2008-EIP-043，Vol. 2009，No. 11，pp. 1－8 (2009)

[4] 佐藤　和紀，安井　良介，金井 敦，廣田 啓一，谷本 茂明，“ブログにおける個人に関する情報の露出状況調査”，情報処理学会シンポジウムシリーズ，IPSJ Symposium Series Vol.2009，No1，DICOMO2009 シンポジウム論文集

[5] “Yahoo!ブログ”，http://blogs.yahoo.co.jp/

[6] gooブログ，http://blog.goo.ne.jp/

[7] “mixi”，http://mixi.jp/