非定型文書を対象としたWebページの発信日付推定

河合剛巨†　中澤聡†　安藤真一†
†NEC 共通基盤ソフトウェア研究所
{t-kawai@bx，s-nakazawa@da，s-ando@cw}.jp.nec.com

1. はじめに

インターネット上の膨大なテキストを対象に時間的な観点からその変遷を分析する研究が多くなされている．しかし，そうした時間的な観点での分析を実現するためには，テキストを有する Web ページの発信日付を推定する必要がある.本稿では,こうしたニーズを背景として，特に類似の記述形式のページが同一サイト内に見受けられない非定型文書を対象とした Web ページの発信日付推定手法を提案する．

ここで,定型文書の Web ページとはサイト内等でページ間に共通の記述形式があり，発信日付の明示的な記述部分を抽出し易いページのことを指す．例えばブログやニュースサイト等であり，その形式の種類は限られるので，こうした Web ページの発信日付推定は容易である．

一方,多くの Web ページは記述形式が多彩であり，固定的なパターン等で発信日付を抽出するには個々のページに応じたパターンを多数用意する必要があるため，発信日付の特定が困難である．そこで本稿ではこうしたサイト固有の記述形式を事前に準備することが現実的でない Web ページを非定型文書と見なす．

定型文書のWebページの典型例としては上述のように,ブログやニュース記事などがあるが，それらはWeb全体の一部に過ぎないため，その他多くのWebページが非定型文書の対象となる．例えば日本語Webページでは，日本語Webページ[*1]中における国内ブログ記事数[*2]の割合は概算で約 1 割と想定される．実際には，複数の記事が 1 ページに存在する場合など記事数とページ数は単位が異なるが，ブログの割合は一部であることが分かる．

[*1] インターネットの全 Web ページ数を 2008 年 6 月時点で 912 億と推計し[1]，その中の日本語の言語分布率を童らの約 13%と仮定する[2]と，912 億中で少なくとも日本語ページは 100 億を超えると推察される．

[*2] 総務省の調査において,2008 年 1 月時点での国内ブログ全体の記事数は 13 億 5000 万件と推計されている[3]．

従って，定型文書に依存せずより多くの Web ページを対象にするには，なるべく汎用的に発信日付を推定する方法が必要となる．

そこで我々は，非定型文書とされたページに記述された複数の日付表現の発信日付らしさを判定し，各日付表現の判定結果に基づき Web ページとしての発信日付を推定する手法を提案し，評価実験により有効性を示す．

2. 関連研究

Web ページの発信時間の推定方法には，収集時間の利用，メタデータの利用，Web ページの記述情報を用いた方法が考えられる．

収集時間の利用は，クロールした収集時間に依存する問題がある[4]．また，発信タイミングに即時的に対応するなど収集の工夫が必要であり，大規模になると収集コストが高い[5]．

メタデータの利用は，RSS/Atom 等のフィードの更新時間情報を使えるが，全てのサイトがフィードを配信しているわけではない．

Webページの記述情報を用いた方法では，ブログや日記，掲示板サイトのように日付表現を伴う記事部分が連続して複数存在するようなWebページを対象とする手法を南野らは提案している[6]．Webページ中に記述された日付表現を複数発見し，日付表現のフォーマットやHTML中のXPathの情報が類似する日付表現を発信日付としDOMツリーを用いてヒューリスティックに発信日付と記事部分を抽出している．Wangらも同様の手法を提案している[7]．

ただし，1 つの Web ページ中に発信日付を有する記事が複数存在しない場合，つまり単一の記事のみから構成される Web ページは少なくない．上記の手法は，このようなページについては考慮されていないため対応できない．

一方で，文書中の時間表現からイベントに関する時間関係の推定を試みるTempEvalがある[8]．TempEvalのタスクでは文書自体の発信時間が付与されていることを前提としており，本研究とは補完的な関係にある．TempEvalでは

文書中の時間表現を使うが，記事の発信時間が不明の場合は十分に考慮されていない．

実際に単一の記事から構成されるページであっても複数の時間表現を含むことが多いため，時間表現を抽出するだけでは何れが発信日付であるかは不明である．

3. 提案手法

提案手法の説明を行う．まず，フィード対応 Web ページや定型情報の利用可能ページ等については既存の手法を用い，それ以外の Web ページを非定型文書とみなす．次に，非定型文書に対して発信日付推定を行う．

提案手法の流れを図 1に示し，以下にその詳細を述べる．

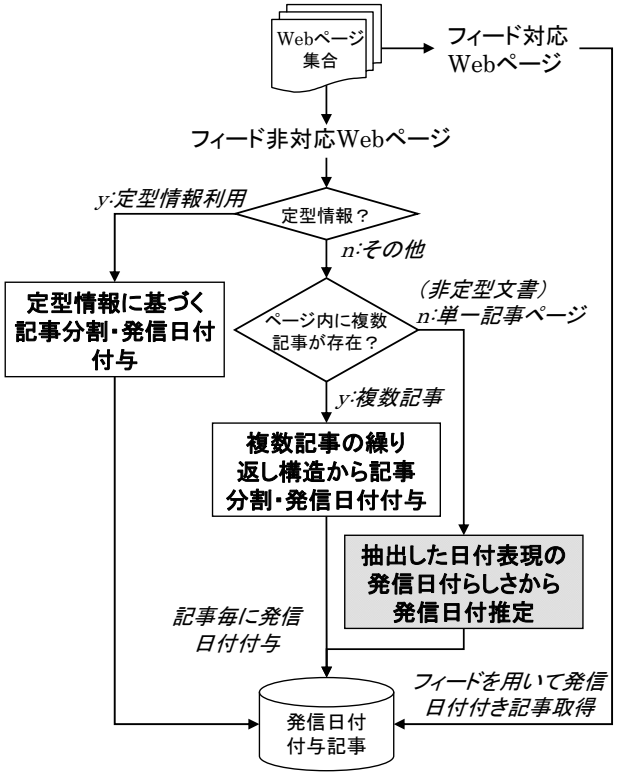

図 1. 提案手法の流れ

3.1. フィード，定型情報の利用可能ページ

まず，Web ページ集合に対し事前にフィードや定型情報を利用して発信日付が付与できるページを URL で弁別する．そして，フィード対応ページはフィードを用いて発信日付とし，定型情報の利用可能ページはそれらに基づき記事毎に分割し発信日付を付与する．

3.2. 複数記事ページ対応と非定型文書の抽出

次に，残ったその他のページに対して，ページ内の日付表現を複数抽出し，抽出した数のうち日付表現のフォーマットが共通する割合が一定以下のページを抽出して非定型文書とみなす．一定以上のページは複数記事が存在すると判断して，ページ内の記事レイアウトの繰り返し構造を用いて，南野らと同様に複数の記事別に記事部分と発信日付とをヒューリスティックに抽出する[6]．日付表現は，WebページのHTML中のテキストの文字コードや表記を正規化し，文単位に分割し形態素解析を行った後の解析結果から抽出する．

3.3. 非定型文書に対する発信日付推定

上記によって抽出した非定型文書に対し，個々の Web ページを単一記事からなる Web ページとみなして記事の発信日付を推定する．

発信日付の推定は 2 段階で行う．

最初に，Web ページ中から抽出した日付表現ごとに発信日付らしさを判定する．具体的には，日付表現ごとに発信日付となりうるか否かを判定する問題を考える．そのため，機械学習を用いて日付表現の発信日付判定器を構築する．

発信日付判定器には，SVM (Support Vector Machine)を用いた．SVM による判定器を適用することで，各ページから抽出した複数の日付表現に対して，個々に発信日付かどうかの判定を行う．カーネルは線形カーネルを使用した．

最後に，各日付表現の判定結果を用いて，ページ単位での発信日付を推定する．各ページ内において発信日付と判定された日付表現を日付が一致するものでまとめて，多数決を行い，最も多く発信日付と判定された日付表現の日付を発信日付とする．多数決の結果が同数の場合には，SVM の分離平面との距離を確信度として用い，最尤の結果を発信日付として採用する．

発信日付判定器に用いた素性を表 1に示す．

まず，抽出した日付表現のページ内の位置的な関係および，同一ページ内の複数の日付表現間の関係を考慮したモデル(L)を基本に考える．ページ内の位置的な関係を考慮するためには，前後の文数や形態素数を用いた．HTML レンダリング後の位置情報を用いる方法もあるが計算コストの問題から採用せず，テキストを用いて位置的な関係を考慮した．

表 1. 発信日付判定モデルの素性

素性	L	2M	2P	2MP	3MP
日付抽出部分の該当文の前に存在する文数	レ	レ	レ	レ	レ
日付抽出部分の該当文の後に存在する文数	レ	レ	レ	レ	レ
日付抽出部分の文内の前にある形態素数	レ	レ	レ	レ	レ
日付抽出部分の文内の後にある形態素数	レ	レ	レ	レ	レ
日付抽出部分の文内の前にある他の日付の数	レ	レ	レ	レ	レ
日付抽出部分の文内の後にある他の日付の数	レ	レ	レ	レ	レ
同ページ内の前にある他の日付の数	レ	レ	レ	レ	レ
同ページ内の後にある他の日付の数	レ	レ	レ	レ	レ
年月日の欠損がないか	レ	レ	レ	レ	レ
日付抽出部分の前後2形態素の原型		レ		レ	レ
日付抽出部分の前後2形態素の品詞			レ	レ	レ
日付抽出部分の前後3形態素の原型					レ
日付抽出部分の前後3形態素の品詞					レ

また，同一ページ内の複数の日付表現間の関係は，他の日付表現の存在と前後の相対位置関係とを文単位およびページ単位で考慮した．

モデル L に加え，周辺情報を考慮するために日付表現の周辺にある形態素の情報を用いた．

形態素の情報は前後 2 つ以内の，原型のみ(2M)，品詞のみ(2P)，原型と品詞を用いたモデル(2MP)の 3 種と，前後 3 つ以内の原型と品詞を用いたモデル(3MP)の場合を検討した．

4. 評価実験

評価実験では，非定型文書とみなす単一記事ページに対する発信日付の推定手法の有効性を検証した．

4.1. 評価用データの作成

評価用データは，非定型文書とする Web ページの人手による正解発信日付と機械的に抽出した日付表現とを比較することにより作成した．

評価用データの元となる Web ページは，評価用 4 トピックに関して収集したページから大手ブログサイトのページを除いた合計 3791 ページを用いた．HTML でないページは対象外とする．次に，収集した Web ページから非定型文書とみなす単一記事を含む Web ページを 721 ページ選別し，それらに対して発信日付を人手により判定し付与した． なお，この人手判定において，ページ中に初回投稿日と更新日の両方が記述されている場合は，初回投稿日を発信日付として採用している．

発信日付判定器の訓練および評価事例は，正解の発信日付を付与したページに基づいて作成した．具体的には，ページ単位に付与した正解の発信日付と抽出した日付表現とを比較し，最初に一致するものを正例とし，それ以外を負例とした．正確にはページ中に含まれる全ての日付表現に対して個々に適切な正例と負例を判断して付与すべきであるが，それは今後検討する予定である．

4.2. 評価実験

提案手法の評価実験では，まず，各日付表現の発信日付判定について評価した(実験 1)．

次に，その判定結果を用いて，ページ単位での発信日付推定の評価実験を行った(実験 2)．

実験 1. 各日付表現の発信日付判定

各日付表現の発信日付判定では，評価用データを用いて 5 分割交差検定を実施し評価した．

評価結果を表 2に示す．

再現率は全正解に占める正出力の割合であり，適合率は全正出力に対する正解の割合である．

表 2. 各日付表現の発信日付判定の評価結果

モデル	L	2M	2P	2MP	3MP
再現率	0.822	0.807	**0.832**	0.807	0.797
適合率	**0.804**	0.745	0.777	0.725	0.713

いずれのモデルの再現率も後述する実験 2 のベースライン手法のページ正解率より高く，発信日付判定器として有効であると考えられる．

個々のモデルの比較では，適合率は L が最も高い結果となった．日付表現の位置的な関係と，複数の日付表現間との関係を考慮した素性は効果が高いと言える．しかし，想定では前後の形態素情報を加えた方が高いと考えていたが，誤判定が増えている．この原因としては機械的に 1 つのみ正解とみなした評価データの作成方法に問題あると考えられる．再現率は 2P が最も高いので，適切な正解全てを用いることで改善できると考えられる．

また．前後にくる「投稿」や「発行」など発

信日付に特有の表現を考慮するため設定した 2M も本評価では寄与しない結果となった．前後 3 形態素のモデルについても，前後 2 形態素のモデルよりも劣る結果であるため，形態素の情報を素性に用いた場合，用いた素性が複雑になるに従い悪化している．実データでは「投稿/日時/：/2007～」のように 3 形態素で判断出来る事例も確認できているため，学習データ量の増加に加え，統一すべき素性はマージするなど素性作成の見直しによる改善の可能性がある．

実験 2．ページ単位での発信日付推定

次に，最終的なページ単位での発信日付の推定を行い，ベースライン手法と比較した．実験 1 の結果より判定器のモデルは L を用いた．

比較用のベースライン手法としては最初に出現する日付表現のみを発信日付として推定結果の出力とする方法を用いた．観察的にページ上部に出現する日付表現が正解になりやすい性質があることが理由である．

表 3にベースライン手法と提案手法の評価結果を示す．ページ正解率は，本手法により正解となったページ数が，全ページ数に占める割合である．出力精度は，本手法が発信日付の推定結果を出力したページのうち，推定結果が正解となったページ数の割合である．ページ出力率は，全ページに対して推定結果をどれだけ出力できたかを示す．

本手法では，ページ内に含まれる日付表現全てを発信日付でないと判定する場合があるため，入力全ページに対して発信日付の推定結果を出力するわけではない．一方，ベースライン手法では，１ページにつき必ず１つの出力があることからページ正解率と出力精度は同じである．

提案手法は出力ページ率が劣るにも関わらず，ベースライン手法と比べるとページ正解率が高く，正しい発信日付のページを多く推定できた．このことは提案手法がベースライン手法より高い出力精度が得られていることからも分かる．後段の推定処理が有効であると考えられる．

以上より，提案手法の有効性が確認できた．

表 3．ベースラインと提案手法の評価結果

	ベースライン	提案手法
ページ正解率	0.707	0.807
出力精度	0.707	0.949
出力ページ率	1.000	0.851

5. まとめ

本稿では，特に非定型文書を対象として，Web ページの発信日付を推定する手法を提案し，実験の結果，有効性を確認した．

今後は評価データの拡充，および評価データ量に応じた素性の見直しにより高精度化を図るとともに，定型文書に対する発信日付推定と統合した全体での評価に取り組む予定である．

謝辞

本研究は，独立行政法人情報通信研究機構(NiCT)の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」の成果である．

参考文献

[1] 山名早人，検索エンジンの信頼性，人工知能学会誌 Vol.23 No.6 pp.752-759 (2008).

[2] 童芳，平手勇宇，山名早人，全世界のWebサイトの言語分布と日本語を含むWebサイトのリンク・地理的位置の解析，DEWS2008 A2-3 (2008).

[3] 総務省情報通信政策研究所，「ブログの実態に関する調査研究」報告書 (2008).

[4] 豊田正史，喜連川優，日本におけるウェブコミュニティの発展過程，日本DB学会論文誌(DBSJ Letters) Vol.2 No.1 pp.35-38 (2003).

[5] 田村孝之，喜連川優，大規模 Web アーカイブ更新クローラにおけるスケジューリング手法の評価，電子情報通信学会論文誌J91-D(3) pp.551-559 (2008).

[6] 南野朋之，奥村学，なんでもRSS! － HTML文書からのRSS Feed自動生成，人工知能学会第 10 回セマンティックウェブとオントロジー研究会SIG-SWO-A501-03 (2005).

[7] J Wang, K Uchino, T Takahashi, S Okamoto, RSS Feed Generation from Legacy HTML pages, APWeb 2006 / Lecture Notes in Computer Science Vol.3841 pp.1071-1082 (2006).

[8] M Verhagen, R Gaizaukas, F Schilder, M Hepple, G Katz, and J Pustejovsky, Semeval-2007 Task 15: Tempeval Temporal Relation Identification, In Proc. of the 4th International Workshop on SemEval-2007. pp.75–80 (2007).