文書の記述内容に関連する日時表現の抽出

廣嶋 伸章　　別所 克人　　小池 義昌　　片岡 良治
日本電信電話株式会社 NTT サイバーソリューション研究所
{hiroshima.nobuaki, bessho.katsuji, koike.y, kataoka.ryoji}@lab.ntt.co.jp

1　はじめに

情報検索は様々な事物に関する情報の取得や行動の指針の決定などに幅広く利用され，人々の生活に必要不可欠なものとなっている．人々の検索意図は多岐にわたるが，その中には日時を考慮した検索を行いたいという状況が少なからず考えられる．例えば，旅行計画を立案する際には，旅行予定日の1年前の同じ時期に旅行先の候補地において行われたイベントについて知りたいという状況が考えられる．

「イベント 2009年3月10日」のように，通常のキーワードに加えて日時を表す文字列を指定することにより，従来の情報検索サービスを利用して日時を考慮した検索を実行することが可能である．しかしながら，このようにして実行した検索には以下のような問題がある．

1. 異なる形式で記述された日時にマッチしない
2. キーワードと日時との関連性が考慮されていない
3. 文書の記述内容と関係のない日時にマッチしている

このうち，第1の問題については，検索対象の文書に含まれる様々な日時のパターンについて解析を行って実際の日時を特定する手法が提案されており[1]，数値データとして扱うことにより問題が解決できると考えられる．また，第2の問題については，過去の研究においてキーワードと日時との関連性を考慮する手法を提案している[2]．そこで，本研究では，第3の問題に焦点をあてることとする．

Web文書には様々な日時が含まれている．Web文書に含まれる日時は，文書の記述内容に関連する日時および文書の作成・更新に関連する日時の2つに大別することができる．それぞれの日時の例としては，以下のようなものが考えられる．

文書の記述内容に関連する日時

- イベントの開催日時
- 製品の発売日時 など

文書の作成・更新に関連する日時

- 新着情報の掲載日時
- コメントの投稿日時
- 文書の最終更新日時 など

このような日時を含むWeb文書を対象として日時を指定した検索を実現するためには，文書の記述内容に関連する日時と文書の作成・更新に関連する日時を区別し，文書の記述内容に関連する日時のみを参照して適切な文書を提示する必要がある．そこで本研究では，Web文書から文書の記述内容に関連する日時表現を抽出する手法を提案する．

2　関連研究

Web文書から特定の日時表現を抽出する研究としては，文書の発信日時を取得する手法がいくつか提案されている．南野らは，Web文書の中からHTMLタグや日時表現のフォーマットの情報などをもとに同じタイプの日時表現を発見し，それぞれの日時表現に対する記事の開始位置と終了位置を求めることにより日時表現と記事の組を獲得する手法を提案している[3]．この手法は掲示板やブログのように日時表現を伴う記事が連続して出現する場合には有効であるが，そうでない場合には同じタイプの日時表現を発見できないため，有効に働かない．河合らは，複数記事の繰り返しではない非定型文書に対して，日時表現の文脈に関する素性を利用して2値分類を行うことにより，日時表現の中から発信日時を取得する手法を提案している[4]．この手法では「2011年1月10日」のように単体で日時が特定できる絶対的な日時表現のみが扱われて

いるが，文書の記述内容に関連する日時かどうかを判定する際には「昨日」のような相対的な日時表現も扱える必要があると考えられる．

3 提案手法

提案手法では，Web 文書からテキストを抽出し，テキストに含まれる日時表現を特定して，それらの日時表現が文書の記述内容に関連する日時表現かどうかを判定することにより，文書の記述内容に関連する日時表現の抽出を行う．以下では，それぞれの処理について説明する．

3.1 テキストの抽出

HTML のタグを除去し，テキストを抽出する．その際，それぞれのタグがテキスト中のどの位置に存在していたかの情報を保持しておく．

3.2 日時表現の特定

日時表現パターンをもとに，テキスト中に含まれる日時表現を特定する．日時表現としては,「2010 年 5 月 5 日」や「2010/05/05」のようにその日時表現だけで日時が特定できるものだけでなく,「5 月 5 日」のような年が省略された日時表現や「今日」「昨日」のようなパターンも日時表現として特定する．

3.3 文書の記述内容に関連する日時表現の抽出

文書中の各日時表現が文書の記述内容に関連する日時表現かどうかを判定するため，2 値分類の機械学習手法である Support Vector Machine(SVM) を用いる．SVM の素性として，Web 文書集合からランダムに抽出した約 2,000 件の文書を分析することによって得られた以下の日時表現に関する文脈の情報を利用する．

1. 日時表現の表記パターン

   着目する日時表現の表記パターン

   着目する日時表現の前方に出現する日時表現の表記パターン

   着目する日時表現の後方に出現する日時表現の表記パターン

2. 日時表現の周辺の単語

   日時表現の前方 5 単語

   日時表現の後方 5 単語

   日時表現の出現する文の文末 5 単語

   日時表現の出現する文の直前の文の文末 5 単語

3. 日時表現の文書中での位置

   日時表現が文書の先頭の文に出現するかどうか

   日時表現が文書の末尾の文に出現するかどうか

   日時表現が文頭に出現するかどうか

   日時表現が文末に出現するかどうか

4. 文書タイプ

   文書タイプ推定により得られた文書タイプ

5. HTML タグ

   日時表現が特定の HTML タグで挟まれているか

ここで，前方とは文頭の方向，後方とは文末の方向を表す．

日時表現の表記パターンを利用する理由は,「2010 年 6 月 15 日（火）」のように省略を行わずに記述されている場合は記述内容に関連することが多く,「2010.06.15」のように記述されている場合は記述内容に関連しないことが多いという傾向が見られたためである．日時表現の周辺の単語を利用する理由は,「開催」という単語が出現した場合は記述内容に関連することが多いという傾向が見られたためである．日時表現の周辺に記号類が存在する場合に手がかりが得られないことがあると考え，ここでは隣接する単語ではなく周囲 5 単語とした．日時表現の文書中での位置を利用する理由は，文書の最後の文に日時表現が出現した場合は記述内容に関連しないことが多いという傾向が見られたためである．文書タイプは文書の種別であり，本研究では「日記」「日記以外」の 2 つのタイプを想定している．文書タイプを利用する理由は,「日記」タイプの文書で「今日」という日時表現が出現した場合は記述内容に関連することが多いという傾向が見られたためである．文書タイプは正確に判定する手段が存在しないため，SVM を利用して文書タイプの分類を行った結果得られた推定値を用いる．文書タイプの分類では，

学習データを作成するコストを削減するため，URL に “blog” または “diary” を含む文書を「日記」タイプとし，URL のドメインが “.co.jp” で終わる文書 (ただし URL に “blog” および “diary” を含まない) を「日記以外」タイプとして学習を行う．HTML タグを利用する理由は，日時表現が dt タグに挟まれていた場合は記述内容に関連しないことが多いという傾向が見られたためである．

4 実験

提案手法の有効性を検証するために実験を行った．日時表現が記述内容に関連するかどうかの判定に関する実験および実験により得られた結果に対する日時表現の特定などを含めた誤りの分析を行った．

実験データとして，Web 文書集合からランダムに抽出した 643 文書に含まれる 4,310 の日時表現に対して 1 人の評価者により正解ラベルを付与した．このうち 3,459 の日時表現に関するデータを訓練データ，残りの 851 のデータを評価データとして利用した．日時表現のパターンなどは訓練データをもとに作成した．SVM のカーネル関数には線形カーネルを利用した．SVM のパラメータ C については，訓練データを 10 分割交差検定し，最適な値を求めた．

評価指標としては，正解率，F 値を利用した．

4.1 ベースラインとの比較評価

まず，ベースラインとの比較を行った．ベースラインとしては，日時表現の表記パターンおよび文書中での位置に関するルールを人手により記述し，ルールに基づいて記述内容に関連するかどうかを判定する手法を用意した．ルールの記述の際は，作成したルールで評価を行い，その結果判定を誤った事例に対して表記パターンや位置に関する共通点を見つけ，それをルールに追加するという操作を繰り返し行った．また，参考として，すべての日時表現に対して記述内容に関連しないと判定した場合の正解率を求めた．

評価結果を表 1 に示す．結果より，ベースラインと比較して提案手法のほうが正解率，F 値ともに高い値となっていることがわかる．ベースラインにおけるルールの作成は著者により行われたが，誤りの事例を観察しても共通点が見つからず，これ以上のルールを記述するのが難しいと思われる段階までルールの記述を行っているため，ベースラインに関する評価結果はルールベースの手法における限界に近い値であると考えられる．提案手法の性能はベースラインの性能を上回ったことから，提案手法はルールベースの手法に比べて有効に働くということがいえる．

表 1: ベースラインとの比較評価結果

手法	正解率	F 値
参考	0.694	-
ベースライン	0.853	0.772
提案手法	**0.895**	**0.824**

表 2: 各素性の効果に関する評価結果

素性セット	正解率	F 値
P+W+L	0.885	0.806
P+W +D	0.887	0.813
P+W +T	0.870	0.785
P +L+D	0.844	0.711
P +L +T	0.834	0.667
P +D+T	0.841	0.698
W+L+D	0.792	0.638
W+L +T	0.786	0.636
W +D+T	0.780	0.631
L+D+T	0.693	0.265
P+W+L+D	0.891	0.820
P+W+L +T	0.881	0.802
P+W +D+T	0.886	0.809
P +L+D+T	0.841	0.712
W+L+D+T	0.785	0.635
P+W+L+D+T	**0.895**	**0.824**

4.2 各素性の効果に関する評価

次に，提案手法で利用した各素性が有効に働いているかどうかを確認するための評価を行った．日時表現の表記パターン (P)，日時表現の周辺の単語 (W)，日時表現の文書中での位置 (L)，文書タイプ (D)，HTML タグ (T) の 5 つの素性セットをそれぞれ利用した場合と利用しなかった場合について比較を行った．

5 つの素性セットのうち 3 つ以上のものを利用した場合について評価を行った結果を表 2 に示す．結果より，すべての素性セットを用いた場合に最も高い正解率および F 値となっていることがわかる．提案したすべての素性セットが性能の向上に寄与していることが

表 3: 記述内容に関連すると誤判定された事例

パターン	件数
MM 月	6
平成 YY 年 MM 月 DD 日	5
季節/特定の時期	4
MM/DD(W)	4

確認できた.

また，5 つの素性セットのうち 4 つの素性セットを利用した場合の結果を比較すると，日時表現の表記パターンを利用しなかった場合に最も性能が低下していることがわかる．さらに，3 つの素性セットを利用した場合についても，日時表現の表記パターンを利用しなかった場合についてはすべて正解率，F 値ともに低い値となっていることがわかる．以上のことから，5 つの素性セットのうち最も性能に寄与している素性セットは日時表現の表記パターンであるといえる．同様に考えると，2 番目に性能に寄与している素性セットは日時表現の周辺の単語であるといえる．一方，性能に寄与しているもののその程度が低い素性セットは HTML タグであるといえる.

4.3 誤りの分析

提案手法によって得られた判定結果が正解と異なっていた事例について，誤りの分析を行った.

まず，提案手法では記述内容に関連すると判定されたが実際には関連しない誤り事例について，日時表現のパターンとその件数を調査した．そのうち全体の 10%以上を占めるパターンとその件数を表 3 に示す.

最も誤りの多かったパターンは,「6 月」のように月だけが単独で記述されたパターンであった．このパターンを持つ誤り事例について実際の文を確認すると,「6 月〜9 月」のように日時の範囲が書かれているものが多く見受けられた．今回の手法ではこのような日時の範囲を「6 月」と「9 月」という 2 つの日時表現として扱っており，記述内容に関連しているかどうかの判定の前に行っている日時表現の特定を正確に行えないことが誤りの原因であると考えられる.

次に，提案手法では記述内容に関連しないと判定されたが実際には関連する誤り事例について，日時表現のパターンとその件数を調査した．そのうち全体の 10%以上を占めるパターンとその件数を表 4 に示す.

最も誤りの多かったパターンは,「今日」「明日」のように日の相対表現を表すパターンであった．このパターンを持つ誤り事例について実際の文を確認すると，この誤りは長期間にわたる日記である 1 つの文書に集中していることがわかった．これは，文書に関する素性である文書タイプの判定を誤り,「日記以外」の文書であるとして扱われたために文書内で連続して誤りが生じたものであると考えられる.

表 4: 記述内容に関連しないと誤判定された事例

パターン	件数
日の相対表現	8
YYYY 年	7
平成 YY 年 MM 月 DD 日	6

5 おわりに

日時表現に関する文脈の情報を素性として 2 値分類を行うことにより，文書中の各日時表現が文書の記述内容に関連するかどうかを判定する手法を提案した.

今後は，日時の範囲などに対しても正しく日時の特定を行えるように改良を行うとともに，Web 文書が複数記事の連続で構成されているかを判別することにより文書タイプをより正確に判定する手法の検討を行いたいと考えている.

参考文献

[1] Han, B., Gates, D. and Levin, L.: From Language to Time: A Temporal Expression Anchorer, Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (2006).

[2] 廣嶋伸章，別所克人，小池義昌，片岡良治: 記述された日時の有効範囲を考慮した日時指定検索, WebDB Forum 2010 (2010).

[3] 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).

[4] 河合剛巨，中澤聡， 安藤真一: 非定型文書を対象とした Web ページの発信日付推定，言語処理学会第 16 回年次大会 (2010).