Web 文書からの利用不可能オブジェクトの自動抽出

岡嶋穣、中澤聡、安藤真一
ＮＥＣ　共通基盤ソフトウェア研究所
{ y-okajima@bu, s-nakazawa@da, s-ando@cw}.jp.nec.com

1. はじめに

情報爆発の時代と云われるようになってから久しく、電子化された文書データにおいても、長期スパンで日々蓄積されている状況である。インターネット上の Web ページは勿論のこと、電子化された書籍や企業内文書等々、蓄積された莫大な文書群を信頼できる知識源として活用していくために、古い文書の中で時間的に劣化し、無効化した情報を検出し更新することの重要性が近年増大している。

文書自体の古さを検出するための研究が存在する。Jatowt ら[1]は、Web アーカイブのスナップショットを用いて Web 文書のコンテンツの古さを判定している。また他の関連技術として、松本ら[2]はニュースやショッピングなど情報が経時劣化しやすいジャンルのページを分類する研究を発表している。しかしこれらの研究は文書の更新時間やジャンルの古びやすさを見る技術であり、文書内のどの情報がどう劣化したかを具体的に判定するものではない。

本稿では、利用不可能になったオブジェクトを Web 上の文書から抽出しリスト化することで、文書内情報の経年劣化の問題にアプローチする。ここでいう利用不可能なオブジェクトとは、閉鎖した店舗や、終了したサービス、などである。利用不可能なオブジェクトを検出することができれば、そのオブジェクトについて言及している、文書中の情報の古くなった問題箇所を具体的に特定し、更新することが容易になると期待できる。

この目的のため、本稿では、オブジェクトが利用不可能となったことを示す表現を含む Web 文書の中から、利用不可能なオブジェクトの候補を抽出し、オブジェクトに関する記述の出現数の時間変化等を手がかりとしてオブジェクトの利用不可能性を判定する手法を提案する。

2. 提案手法

本稿で提案する利用不可能オブジェクトの抽出手法は以下の３段階からなる。１：終了表現を含むテキストの収集、２：利用不可能オブジェクト候補の抽出、３：利用不可能性の判定。以下、この各段階について説明する。

2.1 終了表現を含むテキストの収集

利用不可能オブジェクトについて書かれたテキストを収集するために、我々は終了表現を設定する。終了表現とは、「閉館」や「閉園」、「休刊」や「閉鎖」など、オブジェクトが利用不可能になったことを表わす特定の表現のことである。これらの終了表現を含むテキストには、利用不可能になったオブジェクトの情報が含まれることが期待できる。特に、ニュース記事にはこれらのオブジェクトが効率良く含まれているため、本稿では終了表現を検索クエリとしてニュース記事を検索し、検索結果文書中の終了表現を含む文を、利用不可能オブジェクトを含むテキストと見なして収集する。

2.2 利用不可能オブジェクト候補の抽出

終了表現を含むテキストの中から、終了表現の対象となっている、利用不可能オブジェクトの候補を抽出する。終了表現の対象になるオブジェクトは、前後に出現する名詞との構文的な関係を用いることで抽出できる。たとえば、終了表現として「閉館」を用いて検索し、「交通博物館が閉館することになった。」という記述が得られたとする。この文について構文解析を行い、

終了表現と格関係のある名詞を抽出することで、利用不可能になったオブジェクト候補として「交通博物館」という文字列が得られる。

2.3 利用不可能性の判定

前段階で、利用不可能になっていそうなオブジェクトの候補の集合を得ることができる。しかし、これらのオブジェクトはまだ候補であり、実際に利用不可能になっているとは限らない。たとえば、実際のニュース検索結果には以下のような事例が含まれている。

・日常的・一時的な利用不可能化:「閉館時間は午後６時です。」「今は冬季閉園している。」
・利用不可能化が未実現:「閉館の可能性も視野に入れ検討している。」
・利用不能化が取り止め:「昨年休刊になったが、強い要望に答え復刊することになった。」

これらの事例を取り除くため、オブジェクトが実際に利用不可能かを判定する必要がある。

そこで第３段階では、第２段階で得られた利用不可能オブジェクトの候補の各々について、そのオブジェクトが実際に利用不可能になっているかどうかを、第１段階で得られた文書以外の、オブジェクトを含んでいる文書を手がかりに用いて、判定する。

オブジェクトが利用不可能になっているかどうかを判定するための手がかり情報は、大別して、オブジェクトが利用不可能であることを示す肯定的情報と、利用可能であることを示す否定的情報の２つに分けられる。

オブジェクトが利用不可能であることを示す肯定的情報としては、終了表現を用いて、オブジェクトが利用不可能であることを説明している他の文書の記述が利用できる。たとえば、「閉館」で文書を検索して「交通博物館」が候補として得られたとする。このとき、さらに「交通博物館」で検索して、「交通博物館は先月で閉鎖しています。」のような、そのオブジェクトが利用不可能であることを表わす同様の記述を検索する。その数が多ければ、オブジェクト候補が利用不可能になっている確度は高いと言える。

一方、オブジェクトが利用可能であることを示す否定的情報としては、オブジェクトを実際に利用している文書中の記述が挙げられる。たとえば、「今日、交通博物館に行ってきました。」のような記述があれば、交通博物館が利用可能であることを示す手がかりとなる。また、より広く考えれば、オブジェクトの名称が文書中に出現しているということ自体が、オブジェクトが利用可能な状態であることを示唆する手がかりと見なせる。

オブジェクト名の出現数の時間的な変化も重要である。利用が不可能になったオブジェクトは、文書中で言及する価値も減退し、文書中での出現数も少なくなると考えられる。このため、作成日時が新しい文書でのオブジェクト名の出現数が、全期間での出現数と比較して減少しているかどうかが、オブジェクトが現在も利用可能かどうかを判定する手がかりとなる。

以上の考察から、我々は、利用不可能性を判定するための素性として、以下の３種類の素性に着目する。

終了表現数：オブジェクト名と終了表現が共起する総文書数
全期間出現数：検索エンジンに登録された全期間でのオブジェクト名の出現数
最近出現数比率：文書の作成時間を最近半年／最近１年間に限定したオブジェクト名の出現数の、全期間での出現数に対する比率

本提案手法では、これらの素性を用いて機械学習を行い、オブジェクトの利用不可能性を自動的に判定する。

3. 実験

本実験では、利用不可能になったオブジェクトの正解リストを人手で構築して、提案手法の有効性を検証した。本稿で実験対象とするのは、第３段階の判定正解率である。すなわち、構文解析等によって既に終了表現の対象である「交通博物館」などの利用不可能オブジェクト候補が抽出されているときに、それらのオブジェク

ト候補が実際に利用不可能になっているかどうかの判定を評価する。第２段階の抽出は、人間の作業者による手作業で代替する。第１段階のテキスト取得は一般的な検索手法で実現でき、また第２段階のオブジェクト候補の抽出は、構文解析や照応解析等の既存技術が利用できる。そのため、本稿では、第３段階の評価と考察に焦点を絞る。

3.1 評価データ

利用不可能オブジェクトの候補を得るための検索クエリとして、終了表現３種類「休刊」「閉園」「閉館」を定めた。これらを検索クエリとして、2008 年のニュース記事を検索して該当箇所の文を取得した。これらの文から、終了表現の対象となるオブジェクト名称を人手で抽出し、固有名詞か判定した。さらに固有名詞のオブジェクトについて Web 上で調査し、現在利用不可能かどうかを判定することで、正解データを作成した。得られた正解データの内訳を表１に示す。<記事数>は今回の評価実験のため終了表現で検索した記事の数、<オブジェクト候補数>はそれらの記事の中で「閉館」などの終了表現の対象となっていた語の数、<利用不可能>はその候補の中で実際に利用が不可能になっていたオブジェクトの数である。

表１．正解データ

	記事数	オブジェクト候補数	利用不可能
閉館	129	44	14
閉園	632	113	29
休刊	431	56	34
合計	1192	213	77

3.2 実験設定

学習とテストに用いた素性は前述の３種類である。時間情報を指定して出現数を取得するために、時間情報つきのブログデータを検索した。ニュース記事とブログを使い分けている理由は、オブジェクト候補を抽出するためには客観的な情報を効率よく探索できるニュース記事が、オブジェクト候補が実際に利用不可能か検証するためにはブログ等の時間情報つきの多数の文書が適しているためである。

３つの終了表現から得られたオブジェクトをまとめて全体で学習し判定した場合と、各終了表現ごとのオブジェクトの判定を個別に学習し判定した場合で実験した。学習器としては Support Vector Machine [3]を用い、20 分割交差検定を行った。さらに、学習とテストに用いる素性を変化させることで、各素性が判定結果に与える影響を調べた。

3.3 実験結果

実験結果を表２に示す。<全素性>は全ての素性を使った場合の正解率を、<×～～>という項目は、～～で表わされる特定の素性を使用しなかった場合の正解率を表わす。

表２．実験結果：正解率

	全素性	×終了表現数	×全期間出現数	×最近出現数比率
全体	0.79	0.76	0.79	0.73
休刊	0.78	0.78	0.78	0.65
閉園	0.84	0.78	0.78	0.83
閉館	0.71	0.71	0.68	0.71

全体について全素性を使って学習した場合の正解率は 0.79 となった。正解率向上のために最も寄与している素性は、最近出現数比率であり、正解率を 0.06 改善している。次に終了表現数が効いている。全期間出現数はほとんど効いていない。個別の学習では、終了表現ごとに効いている素性に差がある。休刊には最近出現数比率が効き、閉園には終了表現数が効いている。

3.4 考察

全体で最も有効だった素性は、オブジェクト名の最近の出現比率である。利用不可能なオブジェクトが言及されなくなるという時間的な変化が、利用不可能なオブジェクトを識別する手がかりとして有用であることが分かる。

判定結果の詳細を調査したところ、判定失敗

事例は以下の３つに大別できた。第一に、数十年前に閉園した遊園地など、遠い過去に利用が不可能になったオブジェクトである。こうしたオブジェクトは、インターネット普及後に出現数の時間変化が起こっていないため、誤判定しやすい。第二に、休刊になったあと復刊された雑誌など、一度終了したあとに再び利用が可能になったオブジェクトである。この種のオブジェクトは、一度終了した際に終了表現と多く共起するため利用不可能になっていると誤判定しやすい。第三に、一時的なイベントだけで注目され、イベント期間の間、出現数が増えたオブジェクトである。この種のオブジェクトは、イベント後ほとんど言及されなくなり出現数が減少し､利用が不可能になったと誤判定しやすい。

第一の誤判定を解消するためには､「１９８２年に閉園した」など、終了表現前後の日付表現を手がかりとすることで、インターネット普及以前の古い終了を識別することが考えられる。

第二の誤判定は､「再開」など、終了表現を打ち消すような表現との共起を調べることで解消できる可能性がある。

第三の、出現数の減少による誤判定が起きるようなオブジェクトは、そもそも文書中での出現数自体がわずかであることが多く、期限切れの判定を行うための手がかりを得ることが難しい｡検索対象となる文書集合を拡充することや、少ない出現文書の中の記述を詳細に調べることで対処することが考えられる。

4. まとめと今後の課題

本稿では Web 上の文書を元に利用不可能なオブジェクトを自動抽出する手法を提案した。人手で構築した正解データで利用不可能性判定正解率の評価実験を行い、オブジェクトの文書出現数の時間変化等を手がかりとした判定手法によって約８割の正解率を達成した。

手法の改善のためには､考察で挙げたように、３種類の誤判定の問題に対してそれぞれ対策を行う必要がある。特に、第３の誤判定の問題である、文書中の出現数が少ないオブジェクトの期限切れを精度良く判定するために、二つのアプローチを考えている。一つは、検索対象となる文書集合の拡充であり、ブログ以外の一般の Web 文書についても時間情報を付与する技術を開発している[4]。また、少ない文書のみから判断を行うため、時制やモダリティ等の文中表現の意味解析を強化して、出現数などを用いた統計的な判定と組み合わせる予定である。

さらに、時間情報を用いた素性の詳細化を検討している。オブジェクトと終了表現の共起する文書の作成日付を見ることで、オブジェクトの終了日時が推定できる。この推定された終了日時の前後で出現数が減少しているかを素性とすることで、元の終了表現の時点で実際にオブジェクトが利用不可能になっているかどうかをより厳密に判定することができる。

また、本手法の自動化にあたって、本稿では人手で行っていた第２段階の抽出を自動化した場合の正解率の評価を行い、全体での性能を総合的に評価したい。

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] Adam Jatowt, Yukiko Kawai and Katsumi Tanaka: Detecting Age of Page Content, Proceedings of the 9th ACM International Workshop on Web Information and Data Management (WIDM 2007), ACM Press, Lisbon, Portugal, pp. 137-144 (2007)

[2] 松本 章代, 他: 時間の経過により価値が減衰する情報を主体とするウェブページの判定, Web とデータベースに関するフォーラム 2009, 1B-3 (2009.11).

[3] Vladimir N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[4] 河合 剛巨, 他: 非定形文書を対象としたＷｅｂページの発信日付推定, 言語処理学会第 16 回年次大会