ウェブページにおける情報の鮮度「未来」「現在」「過去」の判別

Judgment of Web Pages Based on the Freshness of Information — Future/Present/Past

テーマ：インターネット技術とその応用
指導教員：松本 章代

教養学部　情報科学科
1057213　亀和田 徹

1.　研究背景および目的

2010 年 7 月にマイクロソフト社が発表した「検索エンジン利用に関するアンケート」[1]によると「現在の検索で不満に思う点」の 2 位は「新しい情報と古い情報が混ざっている」（41.5 %）であった [2]．本研究ではこの点に着目する．

たとえば,「映画 公開日」と検索するとき，この先公開される映画の公開日（未来），今公開されている映画の公開日（現在），現在公開は終了している映画の公開日（過去），のように様々な映画の公開日情報が混在した検索結果となる．また，現在の時点で見ると最新の情報だったとしても，日が経ってしまうと，その情報も過去の情報になってしまう．そして，必ずしも検索者は新しい情報のみを求めている訳ではなく，今までに公開された映画の情報を知りたい場合もある．

このような問題を解決するためには，検索時に未来・現在・過去を選ぶことができればよいと考える．そこで本研究では，ウェブページの情報の鮮度に応じて未来・現在・過去のそれぞれに分類を行うことを目的とする．

2.　鮮度の判別手法

ウェブページの情報の鮮度を決める最も分かりやすい手がかりは，ウェブページ内に表記される日付だと考えられる．ここでいうウェブページの情報の鮮度とは，ウェブページの内容に，いつの時点で起こった出来事なのか，が記載されているかによって決定するものである．たとえば，ウェブページ上に二年前に起きた出来事について書かれた内容があり，新たに分かった付加情報が三日前に追記されたとする．その場合，三日前をウェブページの鮮度として表すのではなく，あくまで内容に応じるため，このウェブページは二年前に起こった出来事とし，現在から見たら鮮度が落ちた，過去のウェブページに判断する．

このように考えたうえで，ウェブページに記載される，日付を鮮度判定に用いることにする．日付表記される文字列を正規表現によって抽出を行う．日付が複数表記される場合には，ウェブページの情報の鮮度を適切に表す日付を選び，判定を行わなくてはいけない．

そこで，ウェブページ上に記述される日付周辺の語に着目する（先行研究において，この語は日付周辺語 [3] と呼ばれるため，本研究でも以下，日付周辺語と呼ぶ）．たとえば，図 1 に示した，日付周辺語「最終更新日」はそのウェブページの更新日を表しているものだと考えられ，未来を表す語だとは考えられない．検索を行う日時によって，現在と過去のどちらかに分類される．このように，未来・現在・過去の分類に有効な日付周辺語をあらかじめ人手により，設定する。すなわち，抽出した日付周辺語がプログラム内に設定された語と一致したときに，その日付周辺語が持つ性質と周辺にある日付の時系列により，未来・現在・過去に分類する．

[1]現在，このアンケート結果が記載されたウェブページは削除されている。しかし，WayBackMachine という過去にあった，削除されたウェブページの状態を閲覧できるウェブサービスによってアンケート内容の確認ができる [1].

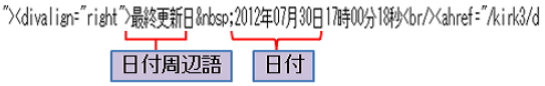

図 1.　日付周辺語の例

3.　研究の流れ

3.1　研究対象データの収集

未来・現在・過去の分類に有効な日付周辺語をあらかじめ見つけ出しておく必要がある．その日付周辺語を収集するために，ウェブ検索の際に，未来・現在・過去の情報が混在した検索結果になるように，実際に確認を行い，検索語を 10 個設定する．これらの検索語を用いて Google にて検索を行う．検索結果（検索結果を表示する際，一番最初に表示されるウェブページから順に数えて 100 件× 10 個の検索語＝計 1000 ページの HTML ファイル）をダウンロードし，それらを研究対象のデータとする．

3.2　日付抽出プログラムの作成

先行研究における日付抽出プログラムの精度を改善するため，3.1 節で得る研究対象データを分析し，日付だと思われる部分を抽出するプログラムを作り直す．主な表記は，年月日という語を用いる日本語，.（ドット）や/（スラッシュ）による記号である。また 8 桁の連続した数字が西暦・月日を表す場合や，英語による表記にも対応し，正規表現によって抽出を行う．

3.3　先行研究の日付周辺語抽出法の再検討

先行研究における日付および日付周辺語の抽出には，HTML ファイル内に記述されるタグに着目をしている．タグを境界とすることで，日付を含む文字列を抽出していた．この抽出法では，文字を装飾する font タグや外部サイトへのリンクタグ等も対象となってしまう．タグの種類に応じて境界とするタグ，境界としないタグを分けるかどうか検討の必要があると考えた．そのため，先行研究に対し，本研究では，境界とするタグについては，ある一定の範囲を示すブロックレベル要素と呼ばれるタグを設定する．ウェブページから閲覧したときに，見出し・段落・リスト・テーブルで表記されるものがそれに該当し，一般的に前後は改行された状態で表示される。ある一定の範囲を抽出できることから，日付と共に日付周辺語が抽出されやすいのではないか，と考えられるため，この方法で日付および日付周辺語の抽出を行う．

3.4　正解データの作成

3.1 節において述べた，HTML ファイル（検索結果表示のウェブページを閲覧した際に，削除されたウェブページを除く計 836 ページ）を用いて，ウェブページの情報の鮮度判定のテストを行うための正解データを二種類作成した後，統合し，作成する．

一つ目は，ウェブページごとに検索した時点から見て，未来・現在・過去のどの時点の情報を表すのか，実際にウェブページを閲覧し，評価をつける．また，時点を判断する際に，未来・現在・過去の情報が 1 つの

ウェブページ	時点	鮮度	日付	日付周辺語
1.html	過去	表す	2013 年 06 月 01 日	発売
1.html	過去	表さない	2013/05/21	投稿日
1.html	過去	表さない	2013/04/01	日時
2.html	現在	表す	2013/07/01	公開日
3.html	未来	表す	2013/12/01	販売日
3.html	未来	表さない	2013/07/01	更新
4.html	過去	表さない	2013 年 06 月 01 日	時間

図 2.　正解データ作成例（イメージ）

ウェブページ内に混在しており，判断が困難と思われるウェブページも確認される．こうしたウェブページには判断困難という分類を行う．

二つ目は，3.3 節で述べた手法により得られた，日付および日付周辺語とそれを含むウェブページ名を出力する．これに，一つ目の正解データの項目である「時点」を分類する際に有効となった日付および日付周辺語が抽出されるか否か，抽出文字列が情報の鮮度を表すか否かの判断を行い，結果を表に書き加える．

この二つの正解データを統合させ，ウェブページの情報の鮮度を分類する結果およびその判断材料となった日付と日付周辺語が記述される表を作成する（例：図 2）．

3.5　判別精度の評価実験

3.5.1　実験準備

3.4 節において作成した正解データの「鮮度」項目の「表さない」とされるものは，鮮度判定に必要ではないことから，行ごと削除しておく．さらに，ウェブページの情報の鮮度を表す日付を一つに決める必要がある．これは，時点の判断を現在にしたにも関わらず，鮮度を表す日付に過去の日付を用いてしまうと，判別結果に影響をおよぼす問題点がある．ウェブページ毎に情報の鮮度を表す文字列が複数個抽出される場合，検討の余地はあるが，ウェブページの内容から視覚的に重要度が高いと感じられる日付に優先度を持たせる．その判断が困難であれば，一番新しい日付を選出するという判断基準を設け，他の行は削除する．

こうして，鮮度を表す日付および日付周辺語が抽出されるウェブページのみが残り，さらに，ウェブページの情報の鮮度を表す時点・日付・日付周辺語は一つに限定される．このウェブページを対象に未来・現在・過去に分類する情報の鮮度判定を行う．

3.5.2　判別分析

作成した正解データを統計解析ソフト SPSS による判別分析手法を用いて，鮮度判定の分類を行う．分類結果からは，元の正解を与えているデータすべてに対して分類した，訓練データの結果が出力される．この判別結果をデータ A と呼ぶ．その一方で，交差検定の手法により，鮮度判定の分類を行う結果も出力される．この判別結果をデータ B と呼ぶ．ここで行う交差検定は，1 つのウェブページには目的変数を与えていない状態とし，そのウェブページ以外から得られた判別係数を用いたうえで，目的変数が与えられていない 1 つのウェブページに対して，分類結果を出力する．その方法を，すべてのウェブページにおいて，他のウェブページからの独立変数の判別係数を用いて，結果を出力する．

3.5.3　評価実験結果および考察

評価実験の結果を図 3 と図 4 に記す．第 1 群が未来，第 2 群が現在，第 3 群が過去，第 4 群を判断困難としている．正判別率はデータ A が 84.9%，データ B が 68.6%となった．交差確認によって正判別率が減少した

（単位：ページ）

		判別された群				
		第 1 群	第 2 群	第 3 群	第 4 群	合計
実際の群	第 1 群	12	0	5	1	18
	%	[66.7]	[.0]	[27.8]	[5.6]	[100.0]
	第 2 群	0	2	3	0	5
	%	[.0]	[40.0]	[60.0]	[.0]	[100.0]
	第 3 群	9	0	137	3	149
	%	[6.0]	[.0]	[91.9]	[2.0]	[100.0]
	第 4 群	0	0	7	6	13
	%	[.0]	[.0]	[53.8]	[46.2]	[100.0]
						正判別率 = 84.9%

図 3.　データ A（訓練データ）

		判別された群				
		第 1 群	第 2 群	第 3 群	第 4 群	合計
実際の群	第 1 群	7	0	9	2	18
	%	[38.9]	[.0]	[50.0]	[11.1]	[100.0]
	第 2 群	0	2	3	0	5
	%	[.0]	[40.0]	[60.0]	[.0]	[100.0]
	第 3 群	25	1	117	6	149
	%	[16.8]	[.7]	[78.5]	[4.0]	[100.0]
	第 4 群	4	0	8	1	13
	%	[30.8]	[.0]	[61.5]	[7.7]	[100.0]
						正判別率 = 68.6%

図 4.　データ B（交差確認データ）

原因として，データ A で過去と判断した 16 ページが未来に，未来と判断した 4 ページが過去に，それぞれ分類されていることが挙げられる．現在に分類される境界を超えて未来・過去に分類されてしまうため，現在に分類する独立変数が有効でないと考えられ，各群のウェブページ数を増やして検討する必要があると思われる．

4.　まとめ

ウェブページの情報の鮮度には日付，日付周辺語が関わっていると考え，これらの情報から判別分析により未来・現在・過去に分類することを試みた．結果は 68.6%と精度向上の余地がまだあることがわかった．問題点は，日付抽出時に，未来を表すウェブページは月ごとに区切られる場合が多く，それらは今回の日付抽出から除外されてしまうことである．よって，それらの日付を抽出させることで，正解データに与えられるウェブページ数が増えることが予測できる．また，正解データ作成時に，未来・現在・過去の分類を行う人手での判断基準を別に定めることにより，分類数に偏りがないようにする．

今後は，第三者に正解データ作成を要請し，ウェブページの情報の鮮度を判断してもらう必要がある．客観的な判断基準を得ることで，鮮度判定の信頼性が増すと考える．

参考文献

[1] “検索エンジン利用に関するアンケート”，Bing, http://web.archive.org/web/20100704051117/http://keyword.jp.msn.com/bing/summary.htm(2010).

[2] 検索不満 6 割は結果空振り，GarbageNEWS.com, http://www.garbagenews.net/archives/1466626.html(2010).

[3] 松田奈々：ウェブページにおける情報の鮮度を判定するシステムの構築，東北学院大学卒業論文 (2013).