Web上の将来情報の分析による予測の信頼性評価

金澤 健介†, Adam Jatowt†, 田中 克己†

†京都大学大学院情報学研究科社会情報学専攻
E-mail: †{kanazawa, adam, tanaka}@dl.kuis.kyoto-u.ac.jp

1. はじめに

将来を予測したいという要望は強く，学術分野においても未来学という分野がある．また，Web上には数多くの将来に関する予測が存在する．我々の行った簡単な実験においては，Web上の約2割の文書が将来情報を含むと推測された．また，Baeza-Yatesの研究[1]によればGoogle Newsには将来情報を含む記事が5万件以上あるとされている．多様で多量の情報があるため，Webは将来情報の収集に大変有用であると考えられる．しかし，我々の知る限り，将来情報を検索できる有効なシステムや手法は現在存在しない．我々の研究の目的は，Web上の将来情報を検索・抽出・集約し，ユーザのWeb上の将来情報活用を支援することである．本研究ではその将来情報検索の準備段階として，Web上の将来情報の抽出手法を提案する．

Web上には新しいページだけでなく古いページも存在している．そのため，記述された時点では将来の事であるが，時間の経過により，読まれている時点においては将来に関する情報ではない場合がある．記述された時間と読まれている時間の差異が，ユーザの信頼できる将来情報の選択を困難にしている．本研究において，読まれている時点において将来について述べている情報を有効な将来情報と呼び，検索エンジンを用いて有効な将来情報を抽出する手法を提案する．有効ではない予測を排除することにより，現状にそぐわない予測を排除することができ，信頼しうる将来情報のみの抽出が行える．

本研究では有効な将来情報抽出のために，時間表現を含む将来情報を用いた．時間表現とは"2008年"や"10年後"といったある特定の時間を参照する表現である．時間表現を含む情報は参照する時間を明らかにすることにより，容易に有効性が明らかになる．時間表現を持つ将来情報より将来情報に特徴的な語を選択し，将来情報有効性の判定に用いる。また，本研究では文を対象として有効な将来情報の有無を判定した．

本論文は，以下の構成になっている．第2章において関連研究を示し，本研究との比較を行う．第3章では有効な将来情報の抽出手法について述べる．最後に第4章において評価実験を行い，第5章で結論と今後の課題を述べる．

Evaluation of Prediction Credibility by Analyzing Future-related Information on the Web
Kensuke KANAZAWA, Adam JATOWT and Katsumi TANAKA
Department of Informatics, Faculty of Engineering, Kyoto University

2. 関連研究

我々はこれまでの研究において，Web上での将来情報の集約を行った[4]．この研究では時間表現を含む将来情報のみを対象とし，クラスタリングにより主な将来のイベントを発見した．

Brun[2]らは，機械学習による文書中の将来情報を効率的に自動的に発見する手法について述べている．"Chronoseeker"と名付けられた提案手法では，テキスト文書中で将来を参照する情報の典型的な特徴を選択し，サポートベクタマシンを用いた学習を行っている．Baeza-Yates[1]は"将来検索"という考え方を提示し，将来検索エンジンの望ましい手法について述べている．この研究では，将来情報検索のための文書の索引づけ・検索アルゴリズムの概要が述べられている．各文書のインデックスには，時刻印とイベントがどの程度起こりうるかの確信度の組が与えられる．検索においては，ある将来の時間に直接関係する文書を返す．上記の2研究においても，時間表現を持つ将来情報のみを対象としている．

山本らは[5]Web上の自然言語での質問に対して信憑性がある答えを見つけるために，時間の分析を行っている．彼らの手法は，Webアーカイブを用いた情報の経過年数の評価に基づいている．本研究においてはWebアーカイブなどの外部に保管されたデータは用いない．

3. 有効な将来情報の抽出

まず，将来情報に特徴的な語の抽出手法について述べる．本研究では時間表現を持つ将来情報より，将来情報に特徴的な語の抽出を行う．すなわち，明示的な将来情報を含む文にのみ有意に多く出現するような語 t を求める．求めた語 t が将来情報を含む文に特徴的な語となる．以後、将来情報に特徴的な語を将来情報特徴語と呼ぶ。手法は以下のとおりである．

1. 対象となる全文中より時間表現を含む将来

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.

情報を抽出し，初期の将来文集合とする．

2. 将来情報集合に現れる語 t に対して“将来文集合と非将来文集合において語 t の出現確率が等しい”という帰無仮説に対してカイ二乗検定を行う．
3. 有意水準 α の検定において棄却された語 t のうち，将来文に含まれる確率が将来文に含まれない確率より高い語を将来情報特徴語とする．

上記の手法により得られた語集合を有効な将来情報の判定に用いる．

次に有効な将来情報抽出手法について述べる．将来情報特徴語が多く含まれる文ほど，将来情報について述べている可能性が高いと考えられる．そのため，将来情報特徴語の文中の出現頻度により，将来情報の有効性の判断が行える．文 S 中での将来情報特徴語の出現頻度を文 S の将来特徴量 $F(S)$ とする．この際に，特徴語の重みとして前項で求めたカイ二乗値を用いる．以下に式を示す．

$$F(S)=\sum_{t\in Term(S)\cap t\in FT} Xai^2(t)\Big/\left|Term(S)\right| \quad (1)$$

$Term(S)$ は文 S 中に含まれる語集合，$Xai^2(t)$ は将来情報特徴語 t のカイ二乗値，FT は将来情報特徴語である．実験的にしきい値を定め，将来特徴量 $F(S)$ がしきい値より大きい文を有効な将来情報と判断する．

4. 実験

文の収集には Yahoo! Search Engine API で英語のページを対象として得られた上位 50 件の結果中のスニペットを用いている．この際に，元のクエリでは結果中に将来情報を含む文が少ないため，クエリを一般的な将来を表す語で拡張を行い検索する．また，検索データは 2009 年 1 月時点のデータである．

実験として，提案手法を用いた有効な将来情報の抽出の再現率・適合率の比較を求める．また，比較手法として将来の時制を含む場合に有効な将来情報として抽出する手法を設定する．クエリは，"toyota", "japan", "energy", "obama", "windows"の 5 クエリを用いた．それぞれの手法における再現率・適合率は表 1 のようになった．

表 1 各手法における有効な将来情報の抽出精度

	適合率	再現率	F 値
時制による抽出	15.1%	67.6%	0.247
カイ二乗値による抽出	17.5%	83.7%	0.290

検定における有意水準 α は 5%，カイ二乗値のしきい値は，10 とした．

5. 結論

本論文では，Web 上の将来情報の信頼性評価の準備段階として，有効な将来情報を発見する手法について述べた．著者と読者の立場を比べることにより，将来情報の有効性の定義を行った．有効性とはイベントが実際に起こりうるかどうかではなく，ある情報が一般的に将来のこととして Web 上に書かれているかどうかを示している．本研究では有効性判断において考慮していない点がある．まず，時間表現を持つ将来情報を用いて有効性判断を行っている点である．そのため，時間表現を持つ将来情報と類似している情報しか発見されない．また，過去の情報が将来情報と類似している場合には，提案手法の精度は低くなる．

本手法の適用領域として，検索エンジンの改良や Web ページ閲覧中において有効でない将来情報の提示，Web マイニングにおいて適合率の改良が挙あげられる．

謝辞

本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，文部科学省科学研究費補助金（課題番号 18049041），NICT 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」（研究代表者：田中克己），および，マイクロソフト産学連携研究機構 CORE 連携研究プロジェクト，MSR IJARC Fellow によるものです．ここに記して謝意を表します．

参考文献

[1] R. Baeza-Yates: "Searching the Future", the ACM SIGIR Workshop MF/IR 2005, 2005

[2] P. Brun, H. Kawai, K. Kunieda, and K. Yamada. "ChronoSeeker: Future Opinion Extraction and Classification", 2009 IEEE/WIC/ACM International Conference on Web Intelligence, 2009.

[3] J. Hobbs and J. Pustejovsky: "Annotating and Reasoning about Time and Events", The Language of Time, pp.301-315, 2005

[4] A. Jatowt, K. Kanazawa, S. Oyama and K. Tanaka: "Supporting Analysis of Future-related Information in News Archives and the Web", the 9th ACM/IEEE-CS JCDL 2009, pp.115-124, 2009

[5] Y. Yamamoto, T. Tezuka, A. Jatowt, K. Tanaka: "Honto? Search: Estimating Trustworthiness of Web Information by Search Results Aggregation and Temporal Analysis", APWeb/WAIM 2007, pp.253-264, 2007

[6] **加藤誠, 大島裕明, 小山聡, 田中克己: "共起に基づく Web からの類似関係のブートストラップ抽出"**, DBSJ Journal, Vol.8, No.1, 2009

Copyright © 2010 Information Processing Society of Japan.
All Rights Reserved.