Webからの明示的・暗示的な将来情報の抽出 (O)

金澤　健介† AdamJatowt†,†† 小山　聡††† 田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻
†† MSR IJARC Fellow
††† 北海道大学大学院情報科学研究科複合情報学専攻
E-mail: †{kanazawa,adam,tanaka}@dl.kuis.kyoto-u.ac.jp, †††oyama@ist.hokudai.ac.jp

あらまし Web 上では計画や予測などの数多くの将来に関する情報が記述されており、それらは新たな予測や意志決定に大変有用である．しかし、将来情報を自動的に判断することは簡単ではない。一見将来情報のようであっても、時間の経過により、実は将来情報ではなくなっている場合がある。これらは有効期限が切れた情報であり、実際は過去を参照している。このような情報が、ユーザによる将来情報の判断を難しくしている。そこで本研究では、将来情報を自動的に抽出する 2 種類の手法を提案し、実験・比較を行った。既存のサーチエンジンを用いて Web から時間表現を持つ将来情報と持たない将来情報を抽出した。前者を明示的な将来情報と呼ぶ。明示的将来情報は将来のある時間を参照しており、絶対的な将来の時間表現を含んでいる。後者の確かな時間表現を含んでいないものを暗示的な将来情報と呼ぶ。本研究では、明示的な将来情報を用いて、暗示的な将来情報を抽出することである。
キーワード 将来情報検索、時間情報, 情報抽出，データマイニング

Extracting Explicit and Implicit future-related information from the Web.(O)

Kensuke KANAZAWA†, Adam JATOWT†,††, Satoshi OYAMA†††, and Katsumi TANAKA†

† Department of Informatics, Faculty of Enginieering, Kyoto University
†† MSR IJARC Fellow
††† Division of Synergetic Information Science, Graduate School of Information Science and Technology, Hokkaido University
E-mail: †{kanazawa,adam,tanaka}@dl.kuis.kyoto-u.ac.jp, †††oyama@ist.hokudai.ac.jp

Abstract There is a lot of future-related information in the Web such as schedules, plans and expectations. This information can be useful for forming predictions by users, decision making and for other purposes. However, the automatic detection of such information is not a trivial task. Given information may refer to the already occurred events, even though, on surface, it may appear to be related to future events. The information is then out-of date, actually referring to the past. This situation happens because of the rapid decay of the information freshness and validity. In this paper, we propose methods for extracting future-related information using existing search engines. We determine time-referenced information and non-time-referenced information from the Web using search engine indices. The former is called explicit future information as it refers to particular future time points and contains absolute future-referring time expressions. The latter is called implicit future and does not contain any explicit and credible future-referring time expressions. We focus on estimating the futureness of implicit future-related information using the explicit future-related information.
Key words Future-related information retrieval, Temporal Information, Information Extraction, Data Mining

1. はじめに

将来に関する情報を知りたいという要望は強い。古くから予言や占いなどが将来を知るために行われてきている。学問分野では、未来学という分野において、今後起こりうる未来や確かな未来、望ましい未来を求めようとしている。このように多く

の人々が意思決定や新たな将来の推測を行うために、将来情報を求めている。また、実際に Web 上には数多くの将来に関する情報が存在する．様々な人々が，将来起こりうるイベントや変化，傾向のような将来に関する多様な計画や予測を，Web 上の文書に書いている．例えば，あるメーカの新製品の発売に関する計画や，今年度のワールドシリーズの優勝チームの予測などである．我々の行った簡単な実験において将来情報を含む文書が Web 上でどの程度あるかを調べた結果，検索結果に対して約 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．これらは新たな将来予測や意志決定に大変有用である．

このように Web には大量の将来情報が存在しており、検索のニーズがあるにも関わらず、現在将来情報を検索できる有効なシステムや、手法は存在しない。そこで、我々は 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目的とする。本研究ではその将来情報の活用支援の準備として、Web 上の将来情報の抽出に関する手法を提案する。Web 上には多数の将来情報が存在するが，ユーザが将来情報を探し出すことは多くの場合難しい。Web 上には最近書かれた情報のみではなく、過去に書かれた古い情報も残されている。これらのものは書かれた時点においては将来のこととして書かれているが、読まれている時点では時間の経過によって将来に関する情報ではなくなっている場合がある。このような記述された時間と読まれている時間の差異により、将来情報の選択が難しくなっている。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将来情報と呼ぶ。また、将来情報かどうかを測る値として将来度を定義する。注意すべき点として、将来度は情報の対象とする時間が将来であるかのみを考慮したものであり、情報の確かさや鮮度とは異なるものである。状況の変化などにより確かさが低くなった計画や予測でも、将来情報となりうる。また、“will” などの将来時制の有無のみでは、将来情報かどうかを判断できない。時制は著者もしくは話者にとっての時間の前後を表したものであり、読者にとっての時間の前後関係を表したものではない。図 1 で将来情報についての説明をする。横軸は文章中に書かれている時間である。書き手にとっては文書 A, B はともに将来に関する情報である。そのため、文章 A,B はともに将来の事として書かれている。一方、読者にとっては文書 B のみが将来情報であり，文書 A は過去の情報となっている．

また、ユーザの将来情報選択を難しくしている要因として、対象となる時間が記述されていない情報の存在がある。対象となる時間に関する記述がある場合にはユーザは将来情報かどうかの判断が簡単に行える。しかし、多くの将来情報では対象となる時間を明示的に表しておらず、述べられていることに対する詳しい知識がないとユーザは将来情報かどうかを判断できない。将来情報のうち将来の時間表現を持つ情報を明示的な将来情報と呼び、時間表現をもたない情報を暗示的な将来情報と呼ぶ。

これまでに定義した語を整理するために図 2 を示す。本研究では主に時間表現のない文から暗示的な将来情報を自動的に抽出する手法について述べる。明示的な将来情報を考慮することで、判断の難しい暗示的な将来情報の抽出を行っている。

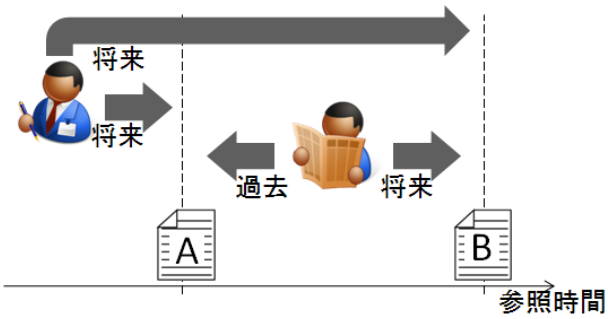

図 1　時間の経過による将来情報となる対象の変化

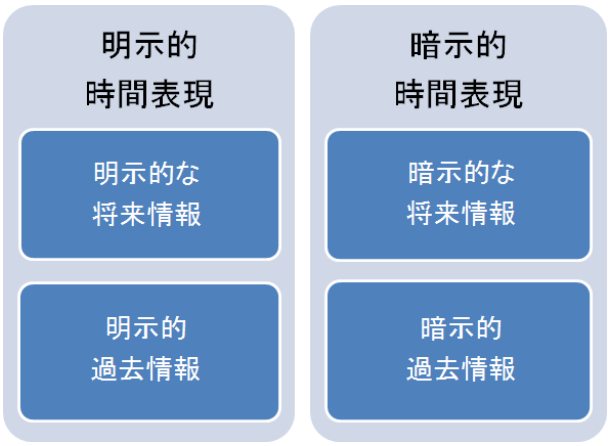

図 2　時間表現による情報の区分

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 章においては将来情報の抽出手法について述べる。最後に第 4 章において結論と今後の課題を述べる。

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]。この研究では明示的な将来情報のみを対象にして、クラスタリングにより主な将来のイベントを発見した。暗示的な将来情報についてはこの研究においては考慮していない。

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に関する情報を参照する典型的な特徴を選択し、サポートベクタマシンを用いて学習を行っている。

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。この研究においても、明示的な将来情報のみを対象としている。

木村ら [8] は，人物の歴史年表を提示するための Web マイニング手法を提案している．彼らの手法の論点は，同名の人物の曖昧性の除去，日付表現の抽出と正規化，日付表現に対するその人物の情報の抽出である．日付表現を表す日本語の言語パターンを作って，日付表現を抽出している．彼らの目的は人物の過去の情報を得ることであり，本研究では将来の情報を得ることを目的としている。

山本らは [6]Web 上での自然言語での質問に対して信憑性の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. 将来情報の抽出

本章では将来情報の抽出についての提案手法について述べる。まず、初めに明示的な将来情報の抽出手法について述べる。次に、暗示的な将来情報抽出の準備として、将来情報特徴語の抽出を行う。将来情報特徴語とは、将来情報の判断の指標となる語である。最後に得られた将来情報特徴語を用いて、将来度を定義し、暗示的将来情報を抽出する。また、本研究では文を単位として情報の抽出を行っている。

3.1 明示的な将来情報の抽出

明示的な将来情報であるかは，将来の時間を表す表現の有無で判断する．そのために文から時間表現を抽出し，時間表現の示す時間を明らかにする必要がある．

文から時間表現を抽出するには，辞書を用いる．時間表現を抽出する辞書を事前に作り，それを用いて文中の時間表現の抽出する．ここで，時間表現が示す時間には粒度に違いがある．例えば，日単位や月，季節，年などである．本研究では，簡単のために年単位で考え，より細かな時間を対象とした場合については今後の課題とする．

次に，時制表現の示す時間を特定する方法について述べる．文に含まれる時制表現には大きく分けて，絶対的な時制表現と相対的な時制表現がある．絶対的な時制表現はその表現部分だけで時間が特定できるもので，例えば “2012 年 9 月 3 日” や “2012 年” などである．相対的な時制表現の例として “2 年後” や “9 月 25 日” が挙げられ，その時制表現のみでは時間の特定が行えない．特定を行うには，相対的な時制表現が参照する別の時制表現を得る必要がある．そのため，本研究では相対的な時制表現を文の作成された時間を視点と考え，時制表現の示す時間を特定した．例えば，“10 年後” といった表現が 2006 年に作られた文に表れた場合，その表現が示す時間は 2016 年である．この手法は常に正しい結果を得られるわけではないが，我々の行った簡単な実験では 8 割程度の精度があった．

3.2 将来情報特徴語の抽出

将来情報に特徴的な語の抽出を行う。前項の手法において抽出した明示的な将来情報を基にして、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 を将来情報を含む文に特徴的な語とする。同様にして明示的な将来情報を含まない文にのみ有意に多く出現するような語 t を求め、将来情報を含まない文に特徴的な語とする。

以下に、具体的な手法を示す。

（1） すべての対象となる文中より明示的な将来情報を含む文を抽出し、初期の将来文集合とする。

（2） 将来文に現れる語 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に特徴的な語、低い語を非将来情報に特徴的な語とする。

（4） 同様にして、すべての文中より過去の時間情報を持つ将来情報を含む文を抽出し、過去に特徴的な語を抽出する。

（5） 将来情報に特徴的な語のうち、過去に特徴的な語に含まれるものを除く。

（6） 将来情報および非将来情報に特徴的な語を、将来情報特徴語とする。

これにより得られた語集合を将来情報の判定に用いる。手法において過去の特徴語を求めているのは、時間表現を含む文に特徴的に表れる語を除くためである。時間表現を持つ文の数は時間表現を持たない文の数に比べて大きな差異があるため、明示的な将来情報を用いるだけでは時間表現を用いた場合に頻出する語が抽出される。しかし、過去情報に特徴的に表れる語を除くことによって、時間表現を含む文に特徴的に表れる語を除くことができる。

また、本研究ではクエリ依存の特徴語と非クエリ依存の特徴語を求めた。クエリ依存の特徴語では、一つのクエリによって得られた結果中の文を対象として、抽出を行っている。非クエリ依存の特徴語では、複数の一般的なクエリによって得られた結果中の文を対象としている。非クエリ依存の特徴語では、“future” や “plan” 等の一般的に将来の情報によく現れる語が抽出される。一方、クエリ依存の特徴語では、クエリの将来の話題に関する語がよく現れる。例として “toyota” というクエリにおいては “hybrid” や “launch” などが得られる。

3.3 将来度の定義と暗示的将来情報の抽出

暗示的な将来情報の抽出手法として、カイ二乗値を用いた手法を提案する。文中に含まれている将来情報特徴語のカイ二乗値の合計を用いることで、将来度を測った。

まず、カイ二乗値による将来情報の抽出について述べる。将来情報に特徴的な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。同様に、非将来情報に特徴的な語が多く含まれている文は将来情報について述べている可能性は低い。そのため、将来情報特徴語の文中の出現頻度により、文が将来情報であるかどうかが判断できる。文 S 中での将来情報特徴語の出現頻度を文 S の将来度 $F(S)$ 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。

$$F(S) = \frac{\sum_{t \in Term(S) \cap t \in FT} X(t) - \sum_{t \in Term(S) \cap t \in NFT} X(t)}{|Term(S)|} \quad (1)$$

$Term(S)$ は文 S に含まれる語集合、$X(t)$ は将来情報特徴語 t

のカイ二乗値、FT は将来情報に特徴的な語集合、NFT は非将来情報に特徴的な語集合である．実験的にしきい値を定め、将来度 $F(S)$ がしきい値より大きい文を将来情報と判断する．

4. 実　　験

本研究では文を対象として将来情報の有無を判定しており、文の収集には Yahoo! Search Engine API により得られた上位 50 件の結果中のスニペットを用いている．この際に、一般のクエリでは結果中に将来情報を含む文が少ないため、クエリ非依存の特徴語から人手によって選択した語を用いてクエリの拡張を行う．また、検索データは 2009 年 1 月時点のデータであり、英語の文書のみを対象としている．

簡易実験として、機械学習を用いた手法、カイ二乗値を用いた手法を用いて将来情報の抽出の再現率・適合率の比較を行う．また、比較手法として将来の時制を含む場合に将来情報として抽出する手法を設定する．クエリとして、“toyota”, “japan”, “energy” の 3 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　各手法における将来情報の抽出精度

	適合率	再現率	F 値
時制による抽出	20.6%	90.0%	0.335
カイ二乗値による抽出	43.9%	66.7%	0.529

カイ二乗値のしきい値は、10 としている．時制による抽出手法では再現率が高いが、適合率は低い．F 値が最も高いのは、カイ二乗値による抽出手法の場合である．本実験より、カイ二乗値による将来情報の抽出が最も精度が高いことが示された．しかしクエリ数が少ないため、さらなる実験によって提案手法の有効性を示すことが今後の課題として挙げられる．

5. 結　　論

本論文では、Web から将来情報を発見する手法について述べた．著者と読者の立場による時間の差異を比べることにより、将来度の定義を行った．将来度は、読者の立場においても将来に関することである．つまり、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．

本研究は暗示的な将来情報抽出の予備研究であり、考慮していない点がある．まず、暗示的な将来情報の発見に明示的な将来情報を用いている点である．提案手法においては明示的な将来情報と類似している暗示的な将来情報しか発見されない．明示的な将来情報と暗示的な将来情報の話題が異なる場合においては、提案手法の精度は低下する．また、過去の情報が将来情報と類似している場合にも、提案手法の精度は低下する．次に、直近に起こったイベントは明示的将来情報においてまだ将来情報として記されている場合が存在し、過去のものだと判断されないことがある．

本研究では、将来情報検索において再現率を増加させることを目的として、明示的・暗示的将来情報の抽出手法を提案した．本手法の適用領域として、検索エンジンの改良や Web ページ閲覧中における将来情報かどうかの提示、一部の Web マイニングの適合率の改良等が挙あげられる．

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特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，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），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「Mining and Searching Web for Future-related Information」，文部科学省科学研究費補助金若手研究 (B)「時間変化するオブジェクト情報の Web からの収集と管理方式の研究」(研究代表者：小山 聡）によるものです．ここに記して謝意を表します．

文　　献

[1] R. Baeza-Yates: “Searching the Future”, In Proceedings of the ACM SIGIR Workshop MF/IR 2005, 2005

[2] P. Brun, H. Kawai, K. Kunieda, and K. Yamada: “ChronoSeeker: Future Opinion Extraction and Classification”, In Proceedings of 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, Oxford University Press, 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”, In Proceedings of the 9th ACM/IEEE-CS JCDL 2009, pp.115–124, 2009

[5] M. Pasca: “Lightweight Web-Based Fact Repositories for Textual Question Answering”, In Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp.87–96, 2007

[6] 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

[7] 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

[8] 木村 塁，小山 聡，田中 克己: “Web からの人物事典生成のための経歴情報の自動収集”，日本データベース学会 Letters，vol.5, No.2，pp.29–32，2006