# ユーザーのスケジュールを用いた Web ページ推薦

和田 潤也 † 大石 哲也 †† 峯 恒憲 ††† 長谷川 隆三 †††
藤田 博 ††† 越村 三幸 ††† 蒋 偉 ††
† 九州大学工学部電気情報工学科
†† 九州大学大学院システム情報科学府
††† 九州大学大学院システム情報科学研究院

## 1 はじめに

近年 Web ページの数は急速に増加している．その膨大な情報の中からユーザが本当に必要とする情報を見つけ出すため，協調フィルタリングによる推薦 [1] やユーザの検索語拡張 [2] など，これまでも多くの Web ページ推薦に関する研究が進められ実用化されてきた．

しかし，これまで「時間」を視点にした Web ページ推薦に関する研究は十分に行われていない．我々は時間を考慮することでタイムリーな Web ページ推薦ができると考え，ユーザの趣向を判断する材料として時間要素を含むスケジュールに着目した．スケジュールにはユーザの嗜好だけでなく，ユーザが情報を必要とするタイミングが含まれている．

スケジュールに基づく Web ページ推薦として，スケジュールデータに対する関連 Web ページの推薦とスケジュールデータそのものの推薦の 2 つを提案する．前者についてはスケジュールデータからユーザの意思を抽出する研究として関連単語抽出アルゴリズムについての実験を行い，その効果を示している [3]．しかし，固有名詞や日付表現といった固有表現に弱い，スケジュールデータに直接的に含まれない語は検索語として利用されにくい，などの課題があった．また，後者については，他ユーザのもつスケジュールデータをユーザの興味に合わせて推薦するためにスケジュール同士の関連度を求める必要がある．

そこで，本稿では「スケジュールデータに対する自動タグ付加システム」について提案する．まず，第 2 節で提案手法の流れや詳細を説明し，第 3 節で具体例を，第 4 節で結論を述べる．

## 2 自動タグ付加システム

まず，「スケジュールデータ」とは最低限以下の要素を含むスケジュールの 1 項目を指すものとする．

- 件名
- 日付

Web page recommendation by using the user's schedules
Wada Junya, Tetsuya Oishi, Tsunenori Mine, Ryuzo Hasegawa, Hiroshi Fujita, Miyuki Koshimura, Jiang Wei
Kyushu University

このスケジュールデータに対して以下の要素からなる「タグ」を付加することを考える．

- タグの名前
- 重み（0 から 10 までの整数値）

ここで，「タグ」とは，そのスケジュールデータに関連した Web ページや別のスケジュールデータを特定するために重要なキーワードのことである．例えば，ある曲の発売日に関するスケジュールデータがあったときには，その曲名やアーティスト名が付加したいタグとなる．

これを付加することで，関連単語抽出アルゴリズムを用いた Web ページ検索を行う際に最低限クエリに含めるべきキーワードを特定したり，同じタグの付いたスケジュールデータ同士は関連が深いとして協調フィルタリングによる推薦を行ったりといったことが可能となる．

### 2.1 システムの概要

このシステムは大きく 3 つのステップからなる．

1. スケジュールデータからのキーワード抽出
2. キーワードの拡張
3. タグの生成

以下でそれぞれのステップについて説明する．

### 2.2 キーワード抽出

まずスケジュールデータの件名からキーワードを抽出する．

高速形態素解析エンジン MeCab[4] を用いて，名詞や固有名詞のみを抽出する．ただし，連続して現れる名詞は 1 つのものとして扱うこととする．これは，スケジュールデータに関連する固有名詞などは複数名詞の組み合わせによって成り立つことが多いためである．ここで得たキーワードを一次キーワードと呼ぶことにする．

### 2.3 キーワードの拡張

次に，スケジュールデータに関連が強く重要な語であるにも関わらず，スケジュールデータの件名には含まれないようなキーワードを見つけだすためキーワードの拡張を行う．

ここでは，一次キーワードをそれぞれ，スケジュールデータの日付とともに Web 検索する．スケジュールデータの日付を用いる利点については具体例とともに後述する．この検索結果に対して関連単語抽出アルゴリズムおよび共起率を用いて単語の重み付けを行う．

このとき，検索に用いた一次キーワード（以下，検索キーワード）の付近に同じ単語がよく現れる場合，検索

キーワードそのものもスケジュールデータに関連が深いキーワードであると判断して重みを増加させる．また，検索キーワード付近に別の一次キーワードが出現している場合も重みを増加させる．逆に一次キーワードであっても，検索結果に一貫性がない（特定の単語が強く現れない）場合や，その他の一次キーワードと同時に出現しないキーワードである場合は，そのキーワードは一般性の高い単語であると判断し重みを減少させる．

こうして新たな重み付きキーワード（二次キーワードと呼ぶ）が得られると同時に，一次キーワードへの重み付けも行われる．

### 2.4 タグの生成

以上の過程で得られた2つのキーワード群の中には重要なものとそうでないものが含まれている．そこで，閾値以上の重みを持つキーワードのみを抽出しそのスケジュールデータに対するタグとして付加する．

### 2.5 ユーザフィードバック

前節までの処理によって付加されたタグの中には，ユーザにとって不要なものやタグとして適切でないものが含まれる可能性がある．そこで，それぞれのタグはユーザによって不適切と判断された場合，それらを平易に除外できるための手段を提供する．このフィードバックによって，あるキーワードがタグとして利用できるものであるかどうかといった学習を行うことが可能となる．

## 3 具体例

例えば，「『サクラ』の発売日」というスケジュールデータがあるとする．ここに「サクラ」はあるCDの曲名，それを歌うアーティストが「TAIFU」である．

このスケジュールデータの件名に含まれる名詞を抜き出すだけでは，「サクラ」および「発売日」のどちらがこのスケジュールデータにとってより重要な語（関連が深い語）であるかはわからない．さらに，件名からだけでは，このスケジュールデータがアーティスト「TAIFU」に関するものであることがわからない．

そこで，このスケジュールデータに対して曲名である「サクラ」や，その曲を歌うアーティスト名の「TAIFU」をタグとして付加することを考える．これにより，「TAIFU」の公式Webサイトや「TAIFU」に関連した様々なWebページを推薦することができるだけでなく，「『TAIFU』ライブツアー」といった「TAIFU」に関連した別のスケジュールデータをこのユーザに推薦することも可能となる．

### 3.1 キーワード抽出

「『サクラ』の発売日」より名詞および固有名詞，すなわち「サクラ」および「発売日」を一次キーワードとして抽出する．この時点では「サクラ」と「発売日」はいずれも対等な重みであり，どちらがよりスケジュールデータに関連したWebページを探すのに役立つキーワードであるかはわからない．

### 3.2 キーワードの拡張

一次キーワード「サクラ」および「発売日」それぞれについて，そのスケジュールデータが持つ日付（ここでは1月25日と仮定する）を用いて「サクラ　1月25日」および「発売日　1月25日」の2つのクエリでWeb検索を行う．

日付とともにWeb検索を行うのは，「サクラ」のみで検索を行った場合に樹木のサクラに関連したWebページなどが多数ヒットしてしまい，このスケジュールデータが「TAIFU」に関連したものであることを特定するのが困難だからである．しかし，「1月25日」のようなスケジュールデータの日付情報をクエリに含めることで「TAIFUが1月25日に新曲『サクラ』を発売！」といった「TAIFU」の単語を含むWebページが多数ヒットするようになる．これによって，Web検索の結果が一般名詞の「サクラ」に関連したものではなく固有名詞としての「サクラ」に関連したものとなり，「TAIFU」といった関連単語を見つけやすくなるのである．

これに対して，「発売日　1月25日」の検索結果については1月25日を発売日とする他のCDや書籍など「TAIFU」とは関連のないページが多数ヒットする可能性がある．そのため，他の曲名や書籍名なども二次キーワードとして選ばれる可能性は高い．しかし，それらのキーワードが検索結果全体に占める割合は「サクラ　1月25日」の検索結果に占める単語「TAIFU」の割合に比べて低くなる．そのため「TAIFU」に比べ重みが小さくなり，タグの生成過程で取り除かれることになる．

### 3.3 タグの生成

このようにして一次キーワード群より「サクラ」、二次キーワード群より「TAIFU」がタグとして発見でき，「『サクラ』の発売日」に対して「サクラ」「TAIFU」のタグが付加される．

## 4 おわりに

スケジュールを用いたWebページ推薦のためのスケジュールデータに対する自動タグ付加システムについて検討した．これによって，スケジュールデータに強く関連したWebページの検索・推薦やスケジュールデータ同士の関連性によるスケジュールの推薦などが可能となる．

また，タグには重みづけが行われているため，そのスケジュールデータに関連する各種公式Webサイトの発見が容易になり，ユーザのブックマークをスケジュールに合わせて動的に変化させるといったことも可能になると考えられる．

## 参考文献


[1] Amazon.co.jp, http://www.amazon.co.jp/

[2] Google, http://www.google.com/

[3] 大石哲也, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸, 倉元俊介, 永田廣人：ユーザのスケジュールを考慮したWeb検索手法, JAWS2007, 2007

[4] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/