2B1-08

観光コース作成支援に向けたイベント情報抽出とその組織化

Extraction and Organaization of Event Information for Computer Assisted Production of Tourist Routes

小作浩美[*1*2] Hiromi itoh Ozaku　山本英子[*1] Eiko Yamamoto　内山将夫[*1] Masao Utiyama　井佐原均[*1] Hitoshi Isahara　河野恭之[*2] Yasuyuki Kono　木戸出正継[*2] Masatsugu Kidode

[*1]通信総合研究所 Communications Research Laboratory　[*2]奈良先端科学技術大学院大学 Nara Institute of Secience and Technology

We are developing a task-oriented search system that assist users to find information from the Internet. In the task of makeing trip plans, it is important to have knowledge of tourist resorts for recommendations and judgement standards that fit itineraries. We think event information and time imformation as event schedules at some tourist resourts are the key information for the recommendations and evaluation when users make trip plans.

In this paper, we reports an examination to automatically extract the relation between event terms and time information. We can automatically extract event terms that have certain periodicy and cooccurrence relations.

1. はじめに

インターネットの普及に伴い，電子化されたテキストが入手しやすくなっている．WWW では，電子化されたテキストを効率良く利用するため，様々な技術が研究開発されている．特に，目的を限定し利用できる情報を必要十分な量，組織化して掲載している目的指向型 WWW サイトは注目されている [古関 2001]．しかし，実際の検索状況を考慮すれば，目的指向型サイトは使いやすいものとは言いがたい．そこで，我々は，WWW 情報をより使いやすく，効果的に利用するため，検索結果をユーザの要求に沿って統合できる「タスク型 WWW 検索システム」を構築を目指している．我々の言う「タスク」はいくつかの処理を経て達成される作業である．ユーザが行っている WWW 検索は，まさに「タスク」であり，数回の検索処理，情報抽出処理，情報統合処理を経て，検索要求を満たす．

実際，タスクには様々なものがあるため，我々は観光コースを作成するタスクを取り上げ，タスク型 WWW 検索サイトの一例として観光コース作成支援システムの構築を行っている [小作 2001]．本システムは，WWW や ML から観光に関係する情報を収集し，時間情報や地理情報により情報を組織化する． ユーザの要求（旅行期間や趣向）に沿った観光地候補を提示し，観光コースの作成を支援する事を目指している．

観光コースを作成するタスクにおいて，観光情報を推薦するための知識とそのコースが実現可能か評価することが重要である．推薦のための知識と評価には，観光地で行われるイベント情報とそのイベントが行われる時間情報を利用することが効果的である．

本稿では，イベント情報とそれに関連する時間情報の自動抽出のために行った実験について報告すると共に，その組織化について考察する．

2. 観光イベント の時間情報

一般に時間情報とは，年月時分を数字で示したものであると考えられる．しかし，実生活の中では，具体的な数字で表される情報よりも，曖昧な言葉で示される情報が数多く存在する． 時間情報と曖昧な言葉や単語の関係知識は一般常識として取り扱われることが多い．コンピュータ上に常識知識を構築することは，重要な研究課題の 1 つである．そのため，時間情報を切口に，常識知識を組織化するための研究がなされている [小畑 2001, 溝渕 1999]．どちらの研究も概念的な時間情報の獲得に有益な考察を行っているが，時間に関連する単語を手動で辞書に登録する必要があるなど，時間情報の自動抽出や利用は難しい．

我々は，観光イベントが周期的に行われること，またある特定の季節に行われる事が多いため，そのイベントに関連する単語は周期的に出現するあるいは，特定の季語（春季や冬季など）の単語と共起して出現すると考えた [小作 2003]．周期性あるいは、共起性を持った単語を抽出できれば，時間情報の抽出を自動化できる可能性がある．

次章で，周期性と共起性のある単語を抽出するために行った実験について報告する．

3. イベント 単語の抽出実験

3.1 実験方法

データベースからある単語の周期性を抽出するためには，検索データ内に発行された日や更新された日が具体的に記述されている必要がある．また，数年にわたり収集されたデータである必要がある．そこで，我々はこの実験に毎日新聞 11 年分の記事を利用した．

前処理として，以下の作業を行った．毎日新聞 1991 年から 2001 年，11 年分の各記事を形態素解析し，発行日情報と名詞を抽出する． 形態素解析には茶筌を利用した [松本 2002]．そして，名詞の出現頻度を利用し周期性を見るために，月毎と週毎に算出した．その出現頻度データを利用し，時間軸（発行日）に対して特徴的な単語を抽出する[*1]．

3.2 周期性の抽出実験

「奈良の山焼き」などのように毎年決まった時期に行われるイベントは，新聞記事でも周期的に現れる可能性がある．単語の出現頻度の周期性を抽出する実験を行った．周期性の検出には，各単語について，11 年分，132ヵ月分の出現頻度あるいは，574 週分の出現頻度を入力とし，自己相関関数を利用した． 自己相関関数 $R(k)$ は，信号の周期の検出に用いられ [江原]，式

連絡先: 小作浩美:romi@crl.go.jp, hiromi-i@is.aist-nara.ac.jp
[*1]独立行政法人 通信総合研究所
〒 619-0289 京都府相楽郡精華町光台 3-5
[*2]奈良先端科学技術大学院大学
〒 630-0101 奈良県生駒市高山町 8916-5

*1 なお，11 年分の記事から抽出した名詞のうち，11 年間の出現数が 11 回以下（ 1 年に 1 回以下）の単語は削除した．連続する名詞は 1 単語として登録すること，数詞のみの名詞は削除するなど，茶筌の結果を一部変更している．

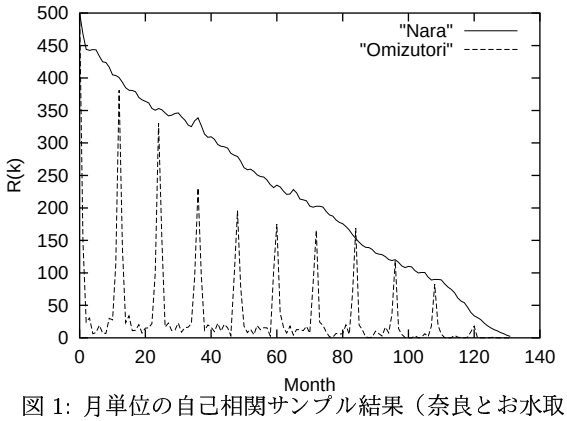

図 1: 月単位の自己相関サンプル結果（奈良とお水取り）

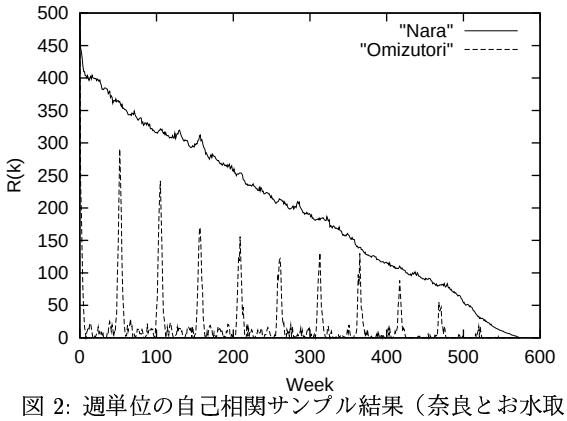

図 2: 週単位の自己相関サンプル結果（奈良とお水取り）

1 で表される．

$$R(k) = 1/N \sum_{n=0}^{N-1} x(n) \times x(n+k) \quad (1)$$
$$(k = 0, 1, 2, ..., N-1)$$

$x(n)$ はある単語の第 n 月，あるいは週の出現頻度を示す．ある単語の自己相関関数の結果において，k が 1 以上の時の $R(k)$ の最大値を取り，その時の k の値をその単語の仮の周期とする．その仮の周期を定数倍した時の自己相関関数の値がピーク値であれば，その周期をもつ周期性のある単語として抽出する．

自己相関関数のサンプル結果を図 1 と図 2 に示す．ただし，見やすさのため，$R(k)$ の値を $R(0) = 500$ となるようにして表示している．これらの図からわかるように「山鉾巡行」と「お水取り」のようなイベント単語は 12ヵ月ごと，52 週ごとの周期性が見て取れる．それに対して，それぞれのイベントが行われる「京都」「奈良」の単語は，月，週単位に関係なく、周期性がない．

3.3 共起性の抽出実験

共起性の抽出実験では，我々は補完類似度を利用した．補完類似度とは文字認識の分野で有効とされている類似尺度であるが，この類似度は事物の出現パターンをベクトルで表したときに二つの事物のパターンの包含状況を測ることができ，事物間の一対多関係を推定する問題にも適用されている [山本 2002]．

二つの事物ベクトルを $F = f_1, f_2, ...f_n (f_i = 0\ or\ 1)$，$T = t_1, t_2, ...tn (t_1 = 0\ or\ 1)$ としたとき，補完類似度 $S_c(F,T)$ は次のように定義される．

$$S_c(F,T) = \frac{a \times d - b \times c}{\sqrt{T \times (n-T)}} \quad (2)$$

ただし，$a = \sum_{i=1}^{n} f_i \times t_i$, $b = \sum_{i=1}^{n} (1-f_i) \times t_i$,
$c = \sum_{i=1}^{n} f_i \times (1-t_i)$, $d = \sum_{i=1}^{n} (1-f_i) \times (1-t_i)$,
$a+b+c+d = n$，$T = \sum_{i=1}^{n} t_i$ である．

周期性のある単語を軸にし，複数の共起する単語の関係（一対多関係）を推定し，周期性のあるイベント単語とそのイベントの開催日時や開催場所の単語の関係がどのようになっているか，調査した．これは，その関係の特徴を利用し，開催日などの情報が自動抽出可能であるか調査しようとしたものである．

周期性のある単語の「山鉾巡行」と補完類似度を計算してみたところ，「長刀鉾」「先頭」や「囃子」「コンチキチン」などが抽出された．これは、イベントの説明に使われている単語である．また，「府警」「調べ」や「天国」「歩行」などの単語も抽出され，このイベントの際には交通規制がなされているのがわかる．周期性のある単語ではないが，奈良で有名なイベントの「角切り」と補完類似度を調査したところ，「鹿」「奈良公園」などイベントに関係する単語が抽出された．一方で、「塩」「コショウ」など「角切り」を料理の単語で利用している場合の関係も抽出された．

4. 考察と今後の予定

周期性のある単語は，月単位の場合と週単位の場合を比較すると，週単位の方が明らかに多い．政治経済関係の単語（予算書，予算案など）や事件に関係する単語（事故者数，事故隠しなど）は週単位にした場合，周期性があると出力される傾向が強い．これは，ある程度短い周期で委員会などが開かれており，それに伴い新聞記事として報告されているためと考えられる．また，事件などは週末に起こり易い傾向があるのではないかと考えられる．しかしながら，これは新聞記事特有の出現パターンである可能性があり，イベント単語とは，違うものとして取り扱う必要がある．

「角切り」は料理関係の記事に現れるため，周期性は取れないが，共起関係から地名や鹿などの特徴的な単語と共起しているため，補完類似度を利用することでイベント単語として抽出できると考える．同音異義語の取り扱いは，補完類似度による関係抽出で取り扱える可能性がある．

メイリングリストなどの情報が詳細化されているデータにおいて周期抽出実験を行い，新聞記事の場合との比較することで，イベント単語の条件が明確になると考える．さらに，これらの条件を利用しイベント単語を含む記事の傾向を利用すれば，推薦知識として組織化できると考える．

我々は，旅行コース作成タスクにおいて，時間情報に着目し，支援することを検討している．本稿では，時間情報を得るための手がかりとしてのイベント単語の抽出実験について報告した．本実験では，出現頻度が少なくても，周期的に現れるイベント単語を抽出できた．また，共起単語の抽出においてもイベント単語と特徴的な関係を持つ単語が抽出できた．

今後は，抽出結果をより詳細に検討し，推薦情報の検索キーとして利用するなどシステムへの組み込みと実推薦データの構築を進める予定である．

参考文献

[古関 2001] 古関義幸, 福島俊一, “新世代検索ポータル技術”, 情報学シンポジウム, (2001).

[小作 2001] 小作浩美, 河野恭之, 木戸出正継, “観光コース作成支援を題材としたユーザビリティの考察”, ヒューマンインタフェースシンポジウム, (2001).

[小畑 2001] 小畑陽一, 渡部広一, 河岡司, “単文の名詞と動詞から時間／季節を判断するメカニズム”, 信学技報 AI2000-56, (2001).

[溝渕 1999] 溝渕昭二, 住友徹, 泓田正夫, 青江順一, “日本語時間表現の一解釈法”, 情報処理学会論文誌, vol.40, No.9, pp.3408-3419, (1999).

[小作 2003] 小作浩美, 内山将夫, 井佐原均, 河野恭之, 木戸出正継 “タスク型 WWW 検索システム構築のための観光イベントと時間情報の関係抽出”, 言語処理学会第 9 回年次大会, pp.663-666, (2003).

[松本 2002] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 “日本語形態素解析システム 茶筌 Version 2.2.9 使用説明書”, (2002).

[江原] 江原義郎, “ユーザズデジタル信号処理”, 東京電機大学出版局.

[山本 2002] 山本英子, 梅村恭司 “コーパス中の一対多関係を推定する問題における類似尺度”, 言語処理学会, Vol.9, No.2 pp.45-77, (2002).