タスク型WWW検索システム構築のための観光イベント単語と時間情報の関係抽出

小作 浩美†‡ 内山 将夫† 井佐原 均† 河野 恭之‡ 木戸出 正継‡
†通信総合研究所 ‡奈良先端科学技術大学院大学
{romi,mutiyama,isahara}@crl.go.jp {kono,kidode}@is.aist-nara.ac.jp

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 [1]．例えば，論文情報を収集し提示しているCiteSeer[1]や，いろいろな商品の比較情報を掲載しているMySimon[2]などが有名である．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．

我々は，WWW情報をより使いやすく，効果的に利用するため，検索結果をタスクに合わせて統合できるタスク型WWW検索システムが必要であると考えている．WWW上には観光に関する情報が大量に存在し，対話システムやEC（電子商取引）の分野でも観光タスクにおける研究が注目されている．そこで，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 [2]．このシステムは，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）に合わせ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．

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 [3, 4]．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．

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．

2 観光イベントの時間情報

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，特定の時期に出現すると考えた．つまり，日々更新されるテキストデータから周期的に現れる単語を抽出すれば，イベントに関係する単語を抽出できる可能性がある．さらに，その単語の具体的に出現する日時（テキストの持つ発行日など）を抽出できれば，溝渕らの研究における，時間情報のうち，時点と周期時間を自動抽出するための規則化が可能である．

周期性のあるイベント情報とイベントの行われる時期が抽出できれば，ユーザがシステムを利用している時を基準時点とし，その基準時点と比較を行うことで，将来行われるイベントを推測することも可能となる．この推測により観光コースの推薦が可能となると考える．もちろん，観光イベントには，周期性のないものも存在するが，ここでは，周期性のあるイベント単語を抽出することを目標とする．さらに，イベント情報には，そのイベントが行われる日付や季節（「冬期オリンピック」や「春季大会」）が共起して現れる場合が

[1] http://citeseer.nj.nec.com
[2] http://www.mysimon.com

表 1: 月毎に抽出された「京都」と「山鉾巡行」の出現頻度

単語	年	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
京都	1991	378	417	339	651	675	551	500	442	407	402	554	690
京都	1992	591	625	750	652	464	503	626	704	512	486	498	533
京都	1993	533	450	549	713	453	533	483	566	430	509	507	461
京都	1994	385	319	467	494	536	766	664	692	586	584	593	616
京都	1995	793	925	602	577	474	598	620	713	461	641	551	588
京都	1996	621	578	610	591	659	801	716	703	714	922	790	755
京都	1997	849	738	809	706	818	715	965	869	861	1526	1308	1841
京都	1998	773	709	1084	842	699	666	814	1039	638	954	903	740
京都	1999	819	755	911	823	834	776	868	749	696	936	847	904
京都	2000	864	991	820	807	766	770	775	785	693	799	830	757
京都	2001	726	597	849	861	827	833	1224	803	670	754	852	800
山鉾巡行	1991	0	0	0	0	0	3	8	0	0	0	0	0
山鉾巡行	1992	0	0	0	0	1	6	10	0	0	0	0	0
山鉾巡行	1993	0	0	2	0	0	4	13	0	0	0	0	0
山鉾巡行	1994	2	0	0	0	4	4	17	2	0	0	0	0
山鉾巡行	1995	0	0	0	0	1	8	11	0	0	0	0	0
山鉾巡行	1996	0	0	0	0	0	1	13	0	0	0	1	0
山鉾巡行	1997	0	0	0	0	0	2	10	0	0	0	0	0
山鉾巡行	1998	0	0	1	0	0	1	7	0	0	0	0	0
山鉾巡行	1999	0	0	0	0	0	1	8	0	0	0	0	0
山鉾巡行	2000	0	0	0	3	0	1	9	0	0	0	0	0
山鉾巡行	2001	0	0	0	0	0	1	12	0	0	0	0	0

あると考え，各月（1月，2月といった単語）や季節単語（春，夏，秋，冬）との共起性についても調査し，より正確にイベント単語を抽出することを目指す．

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[5]．そして，名詞の出現頻度を大まかな周期性を見るために月毎に算出した．その例を表1に示す．表1において,「山鉾巡行」は，京都で行われる祇園祭の中心的イベントであり，我々が抽出したいと考えている観光イベント単語である．一方,「京都」は観光地ではあるが，イベント単語ではない．

このような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．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．

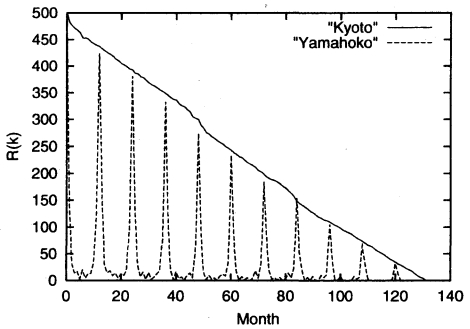

図 1: 自己相関サンプル結果 1（京都と山鉾巡行）

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（表1）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[6]，式1で表される．

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

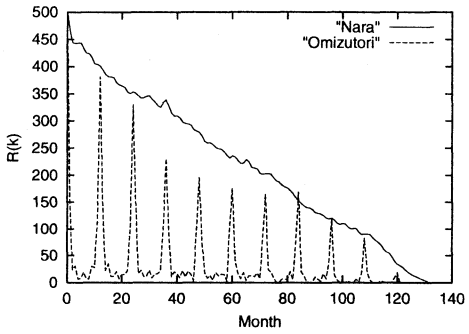

図 2: 自己相関サンプル結果 2（奈良とお水取り）

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，仮の周期を周期性が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，周期性が全くない．

出現数が 11 回以上あり，各年に必ず出現しているとして抽出された名詞群 10 万 3 千語のうち，周期性のある単語として抽出されたものは約 1 万 5 千語あった．その中には，12ヵ月ごとに「お盆休み」「びわ湖毎日マラソン」「日本シリーズ」や 48ヵ月ごとに「W 杯出場」などがある．例を表 2 に挙げる．

3.3 共起性の抽出実験

イベント情報には，開催地，開催日が含まれるため，1月，2月などの単語と共起して現れる可能性がある．また，イベントの名称には,「春期大会」などと，季節単語を利用したものが多く見られるため，季節単語とも共起して現れる可能性がある．そこで，本節では，各月（1月，2月，3月などの単語）と春，夏，秋，冬の 4 つの季節単語との共起頻度の高い単語を抽出する．

共起頻度は，対数尤度比 [7] の大きいものとした．単語 v と単語 w の対数尤度比 λ とは，v と w の 2 単語が従属とした場合と独立とした場合との最尤推定量による尤度比であり，式 2 で表される．2 単語が従属している度合が強いほど大きい値を取る [8].

$$\lambda = 2\sum_{i,j} f_{ij}\left\{log\frac{f_{ij}}{F} - log\frac{f_i f_j}{F^2}\right\} \qquad (2)$$

ただし，$f(v,w)$ を単語 v と w が同時に出現した文書数，$f(x)$ を単語 x が出演した文書数，F を全文書数とする．

実験結果をの一部を表 3 に示す．

表 2: 周期性のあると判断された単語例

周期	抽出された単語例
2 から 4ヵ月	千秋楽，振興自治宝くじ，出げいこ
6ヵ月	アマ本因坊対抗戦，最高殊勲選手賞，驚見賞
12ヵ月	アユ解禁，サクラ開花，岸和田だんじり祭，祇園祭
24ヵ月以上	カンヌ国際映画祭，東京モーターショー

表 3: 各月に共起頻度の高い単語例

1 月	阪神大震災，新年，昨年末，大相撲初場所，正月
2 月	兵庫県南部地震，阪神大震災，長野五輪，衆院予算委会
3 月	オープン戦，大相撲春場所，選抜高校野球大会
4 月	統一地方選，バレー特集，桜，センバツ高校野球，入学式
5 月	新社長，大相撲夏場所，6 月就任
6 月	フランス W 杯，サミット，株主総会，衆院選
7 月	都市対抗野球，アトランタ五輪，大相撲名古屋場所，選挙
8 月	全国高校野球，夏，夏休み，甲子園，終戦記念日
9 月	大相撲秋場所，シドニー五輪，台風
10 月	日本シリーズ，秋季高校野球，米国同時多発テロ
11 月	APEC，大相撲九州場所，米英アフガン攻撃
12 月	全国高校駅伝，全国高校ラグビー，師走，大蔵原案

4 考察

周期性のある単語の抽出実験においては，興味深い単語が抽出できた．特に，有名な観光イベント単語（だんじり祭，山鉾巡行など）だけでなく，観光目的として曖昧に利用されやすい単語（山開き，桜開花，キンモクセイなどの花の名前など）も抽出できており，イベント単語抽出の一条件として利用できると考える．また，出現頻度の大小にかかわらず，周期性のある単語が抽出できることが分かった．

イベント単語の周期については，自己相関関数を利用する事で，現在は月単位ではあるが，12ヵ月ごとなど，具体的な数値を得る事ができた．しかし，周期性のある単語すべてが，イベント単語ではない．より詳細な期間についても調査し，より周期性の高い単語を抽出し，イベント単語の抽出精度をあげる必要がある．

共起性のある単語については，季節単語（春，夏，秋，冬）を利用し調査を行っていたが，常識的な事実の場合は，季節単語が明記されないため，共起性からは常識的なイベント単語が抽出できないことがわかった．例えば，本来夏に行われると考えられる「水泳大会」が，冬に行われるイベントとして抽出される．これは，一般的に「水泳大会」が夏に行われるために,「夏期水泳大会」とわざわざ記述することは少なく，一般的でない場合には「真冬の水泳大会」と季節を明示することが多いためと考える．

月単語との共起性においては，スポーツ欄に関係するイベント単語（名古屋場所，九州場所，天皇杯など）が抽出された．毎月，何が行われるか把握するには利用できる単語であると考えるが，我々の求める観光イベント単語とは，若干異なるものである．一方，イベント単語として抽出された単語との共起頻度の高い単語を抽出する事で，イベントの行われる場所を把握することができる可能性がある．例えば，周期性からイベント単語として抽出された「山鉾巡行」が掲載されている記事から，共起している単語を抽出すると,「京都」という開催地名が抽出できると考えられる．共起情報の利用については，さらに検討する必要があると考える．

今回の実験では，名詞だけに着目してイベント情報の抽出を試みたため，取り扱えなかったイベント情報が存在する．例えば,「鹿の角切り」や「桜の通り抜け」のような「～の」による修飾を受けて，ある固有のイベントとして表記されるものや,「食べる会」「歩く会」のように動詞＋名詞の組み合わせを利用した言葉を利用したものである．詳細な出現頻度を調査していないが，観光イベントをより正確に抽出するためには，取り扱える必要があると考える．さらに，同音異義語の取り扱いについては注意が必要である．例えば，奈良の観光として有名な「鹿の角切り」があるが，これは茶筌による形態素解析の結果として，名詞（鹿）＋助詞（の）＋名詞（角）＋名詞（切り）に分割される．今回の実験では名詞の連続にだけ着目したため,「角切り」が調査候補となる.「角切り」は,「若鶏肉の角切り」や「野菜の角切り」といった料理に関係する記事にも出現し，周期性が抽出できない．調査記事を分野に分けて調査する必要があると考えられるが，WWWデータの検索において，分野にデータを分類することは，負荷が高い作業であるため，他の方法を検討する必要がある．

5 おわりに

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，各月における共起単語の抽出においても時間軸に関して特徴的な単語が抽出できた．

今後は，抽出結果をより詳細に検討し，これらの単語を観光旅行の期間と照らしあわせ，時期的にマッチしている場合に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を検討していく予定である．

今回の実験結果中には，イベントと関係のない単語も抽出されている．不要な単語の除去についても検討する予定である．イベント単語を抽出する方法を確立し，この知識を有効に利用した観光コース作成支援システムの実現を目指したい．

参考文献

[1] 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).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] 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.

[7] T. E. Dunning, “Accurate methods for the statistics of surprise and coincidence”, *Computational Linguistics*, 19(1),pp.61-74, (1993).

[8] 内山将夫, 井佐原均, “情報検索パッケージの実装”, 情処研報告 FI2000-63-8, (2001).