WWWからの観光情報の抽出と比較

呉, 小斌
九州大学大学院システム情報科学府

廣川, 佐千男
九州大学情報基盤研究開発センター

殷, 成久
九州大学情報基盤研究開発センター

中藤, 哲也
九州大学情報基盤研究開発センター

他

出版情報：情報処理学会九州支部若手の会セミナー, 2010-09. Information Processing Society of Japan
バージョン：
権利関係：

WWW からの観光情報の抽出と比較

呉小斌*　廣川佐千男*[2]　殷成久*[2]　中藤哲也*[2]　田畑義之*[2]

九州大学大学院システム情報科学府*　　九州大学情報基盤研究開発センター*[2]

＜あらまし＞何か調べ事をするときには、検索エンジンを使うようになっている。しかし、一般の検索エンジンで網羅的に調べられたかどうかは分らない。我々は、特定の分野・対象の情報を WWW 上から網羅的に収集し、収集データにもとづく専門的検索エンジンの構築について研究を行っている。本研究では、収集すべき情報として、特定の地域の観光情報を取り上げ、観光情報専門のサイトに記されている情報と、一般検索エンジンを用いて収集した情報を比較する。それにより、人手により収集された専門サイトと、一般検索エンジンから得られる情報の範囲と質の比較を行う。

＜キーワード＞　検索エンジン,情報の抽出, 順位相関係数,データ収集,　地域の観光

１．はじめに

情報検索のための技術的な研究は、インターネットの普及を契機として、質・量ともに飛躍的に進歩しており、インターネットによる情報収集が可能になっている。何か調べ事をするときには、検索エンジンを使うようになっている。しかし、一般の検索エンジンで網羅的に調べられたかどうかは分らない。我々は、特定の分野・対象の情報を WWW 上から網羅的に収集し、収集データにもとづく専門的検索エンジンの構築について研究を行っている。本研究では、収集すべき情報として、それにより、人手により収集された専門サイトと、一般検索エンジンから得られる情報の範囲と質の比較を行う。

最近外国人観光客が連続的に大きく増加した。特に中国の富裕層に対するビザの交付が大幅に簡略化されたことで、中国人観光客の数がますます伸びることが期待できる。しかし、外国旅行には、言葉や習慣・文化などの違いによって旅行者を不安にする場面がある。旅行情報の問題はそれらの情報の信憑性であり、まことしやかな「間違った情報」がたくさんあるほか、特に多いのは、今は使えない古い情報である。つまり、インターネット上の情報の正確性は保証できない。したがって、旅行者に品質良く、かつ正確な観光情報を提供する必要がある。

図１．九州観光推進機構のイベント情報

観光は単なるスポットを見るだけではなく、イベントにも参加できることにより、文化の理解につながり、意味のある旅行になると考えられる。地域の観光活性化を支援するため、本稿では、九州観

光のイベント情報を取り上げ（図１）、九州観光推進機構の Web サイト(http://www.welcomekyushu.jp/)に記されている情報と、Google(http://www.google.co.jp/)検索エンジンを用いて収集した情報を比較する。

本研究は、外国からの観光客、あるいは外国からの観光客にサービスを提供する組織をユーザーとして考え、どのような観光情報が求められ、どのような形で観光情報を提供したらよいかを検討するための研究である。本稿は、そのための第一ステップであり、現在 Web でどのような観光情報がどのようなサイトで提供されているか、という予備的調査を行ったものである。日本の観光についてのあらゆる情報を網羅的に収集するのではなく、必要とされる個別の情報に限定することで、効率よい収集と提供を目指す予定である。この研究を行うことで、旅行者に旅行情報が正しく伝えられ、九州への旅行者を増やすことに繋がる事が期待できる。

2. 観光イベント情報の収集

本研究で取り扱う観光情報として、九州地域のイベント情報を選択した。イベント情報を収集するため、まず最初に次の方法でイベント名の一覧を作成した。九州観光推進機構の Web サイト（以下、welcomekyushu）に掲載されている「九州・沖縄のイベント一覧」のページから 906 件の情報を取得した。このイベント一覧には、「ざる豆腐」等の食べ物名や「しゃくなげ公園」等の地名が入っていたため、イベントとしての開催日が書かれているものだけを抽出し、316 件のイベント名を取り出した。更に、イベント名に書かれた開催年に関する文字列（「2009 年」など）を取り除き、リスト中で結果的に同一となったイベント名を削除した。これにより、312 件からなるイベント名リストを作成した。収集したイベント名の一部を表 1 に示す。

表１：九州・沖縄のイベント名リスト中の一部 20 件

十五夜ソラヨイ，鏡山スカイスポーツフェスティバル，串木野浜競馬，「小城」ホタルの里ウォーク，とす弥生まつり，べっぷ鶴見岳一気登山大会，宮崎神宮大祭，美山窯元祭り，小倉祇園太鼓，四十九所神社「やぶさめ祭り」，白地楽，幸若舞，うすき竹宵，筑前いいづか雛のまつり，牛深ハイヤ祭り，武雄の荒踊り，門司みなと祭，鹿児島カップ火山めぐりヨットレース，南大隅町ねじめドラゴンボートフェスティバル，山鹿灯籠浪漫・百華百彩

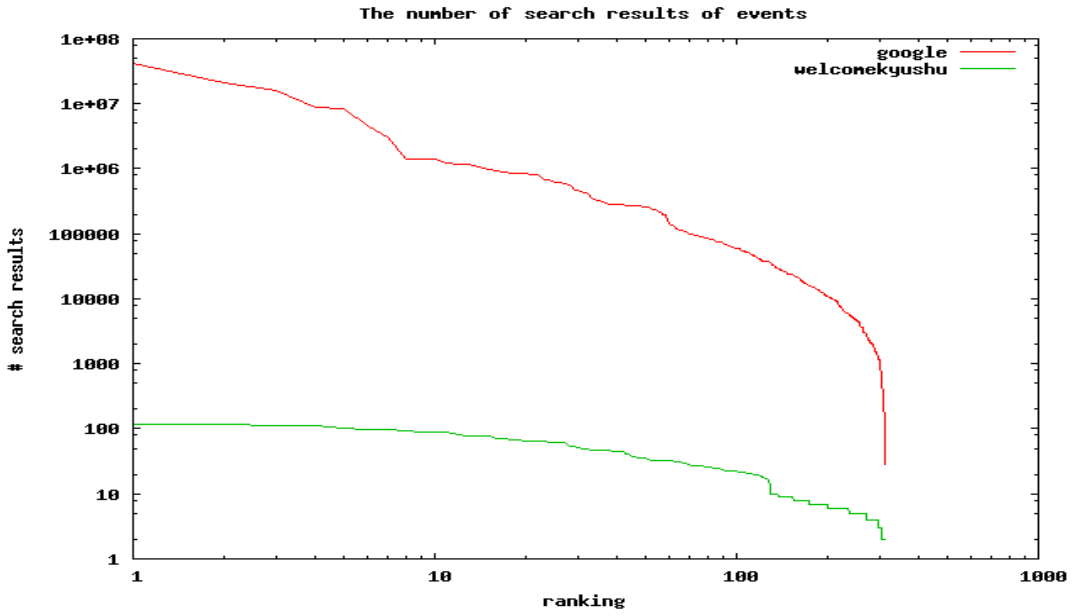

図２．各イベントについて検索結果の件数

これらのイベント情報に関して、WWW 上の一般的な Web ページで扱われている情報量、及び観光専門サイトで取り扱われている情報量を比較するために、検索エンジンによる検索結果の件数を取

得した。検索エンジンとしてはGoogleを、観光専門サイトとしては九州観光推進機構を選択した。まず、作成したイベント名リストの各イベント名をキーワードとしてGoogleに対して検索を行い、取得したページ中の検索結果件数を一般的なWebページ上の情報量とした。同様に、"site:演算子"を用いてGoogleの検索範囲をwww.welcomekyushu.jpに限定して検索結果の件数を取得し、welcomekyushuの持つ情報量を観光専門サイトの情報量とした。

一般的なWebページにおける情報量は、キーワードとしたイベント名によって、29件から241,000,000件まで、観光専門サイトの情報量は同じくイベント名によって2件から603件であった。このイベント毎の検索結果の件数の多いものから順にプロットしたものを図2に示す。

3．順位相関での分析

3.1 ケンドールの順位相関係数の計算

我々はイベント情報をケンドールの順位相関係数[Abdi 2007]で比較した。ケンドールの順位相関係数を計算する手順は以下の4つのStepから成る。

Step 1. 2変数x, yについて、測定値の数をnとする。

Step 2. 変数 Xと変数Yについて小さい方から順位をつけ，同順位があるので平均順位をつけ、変数 X について小さい順に並べ変える。

Step 3. 同順位があるので、変数 X、変数Y における同順位の個数をnx 、ny、同順位の大きさを$t_i, t_j(i = 1, 2, \ldots, nx ; j = 1, 2, \ldots, ny)$としたとき，次式$(Tx , Ty)$で計算される。

$$Tx = \sum_{i=1}^{nx} \frac{t_i(t_i-1)}{2} \qquad Ty = \sum_{i=1}^{ny} \frac{t_j(t_j-1)}{2}$$

Step4. $Y_i\big(\mathrm{i} = 1, \ 2, \ \ldots, \mathrm{n} - 1 \big)$にいて、$Y_i < Y_j$の個数を$P_i$, $Y_i > Y_j$ の個数をQ_iとする $(j = i + 1, \ i + 2, \ \ldots, \ n)$。

$\sum P_i$ は 2 変数の順位の方向が一致する回数，$\sum Q_i$ は 2 変数の順位の方向が逆方向に一致する回数なので，$\sum P_i - \sum Q_i$ は順序の一致性の指標である。(同順位がない場合には $(Tx = Ty = 0)$となる)

このような準備のもと、ケンドールの相関係数r_kは次式で定義される。

$$r_k = \frac{\sum_{i=1}^{n} P_i - \sum_{i=1}^{n} Q_i}{\sqrt{\frac{n(n-1)}{2} - Tx}\sqrt{\frac{n(n-1)}{2} - Ty}}$$

3.2 結果と検定

本稿では312件のイベントについて、Googleによる順位とwelcomekyushuによる順位を比較するので、n=312となる。手順に従って求めた相関係数は0.20744であった。更に有意性検定を行い、有意確率P=4.616e-08を得た。有意水準5%（α=0.05）で検定を行うと、$P<\alpha$であるので帰無仮説は棄却された。すなわち、有意な相関があるといえる。

4．順位に差があるイベント

312 件の各イベントについて、Google での順位をx 軸、welcomekyushu での順位をy 軸にプロットし たのが、図3 である。強い相関に対応する直線A から離れたイベントは、Google とwelcomekyushu で の順位が大きく異なる。それらがどのようなイベントであるかを分析する事で、以下の 4 項目が分か った。

1. 領域Xに存在する点の数はケンドールの順位相関係数の値を上げる効果がある。点の数が 多くなったら、相関係数が高くなる。この図からも弱い相関であることが確認できた。
2. 領域Yに存在する点はGoogle で上位、welcomekyushu で下位である。これらの点に代表されるイ

ベント名は全国どこにでもありそうなイベント名であった。例えば､｢大人歌舞伎」や「甘酒 まつり」といった名称のイベントなどである（表２）｡

3. 領域Zに存在する点はGoogle で下位、welcomekyushu で上位である。これらの点に代表されるイベントは九州に特有のものであった｡イベント名に地名が入っているものも多い｡例えば｢山 鹿温泉祭」や「宝泉寺ホタル祭り」｢桜島納涼観光船」といったイベントである（表３）｡

4. 領域Yの周囲にあるイベントの数は、領域Zの周囲にあるイベントの数よりも多い。これは、welcomekyushu が地域性の強いポータルサイトであることを示している。

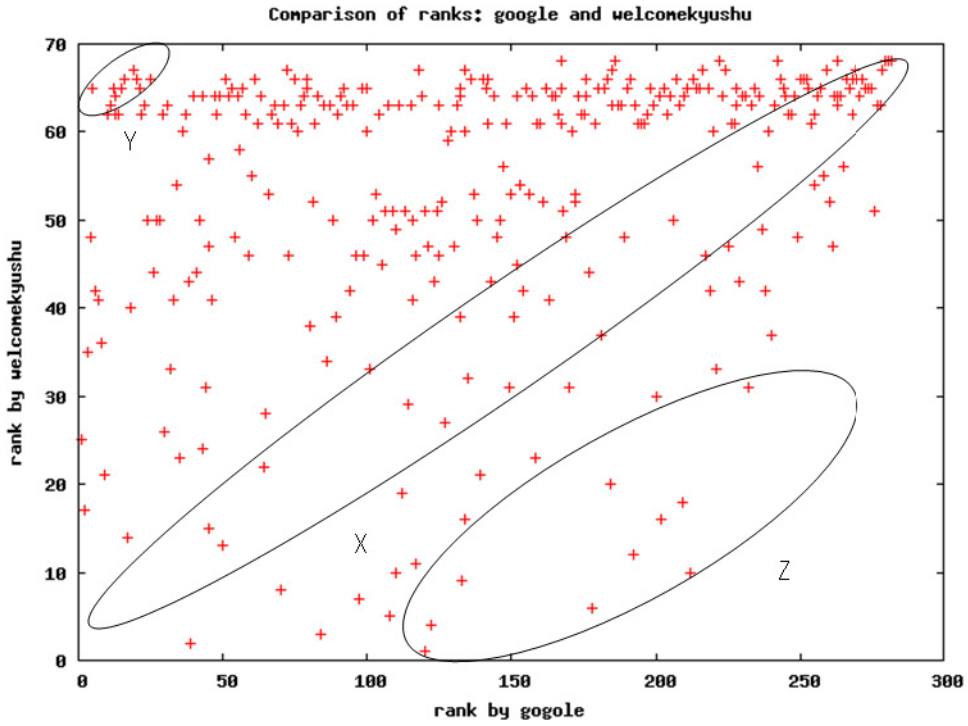

図３．イベントランキング

一般的な Web ページにおける情報量と、観光専門のサイトの情報量の違いを見るために、イベント名毎の件数の順位を比較した。表２は一般 Web ページの検索結果数の順位が相対的に低かったもの、表３は逆に観光専門サイトの検索結果数の順位が相対的に高かったものである。

地名を含んだイベント名が一般的な Web ページで検索される傾向を見て取る事ができる。

表 2：Web ページの順位が低く、welcomekyushu の順位が高い 10 件

順位	*rg*-rk*	rg*	rg	rk	g	k	イベント
1	*63.79*	1.21	5	65	8710000	5	竹ん芸
2	*62.42*	4.58	19	67	858000	3	おしろい祭
3	*62.14*	3.86	16	66	935000	4	みそ五郎まつり
4	*62.11*	2.89	12	65	1240000	5	甘酒まつり
5	*61.38*	3.62	15	65	965000	5	鮒市
6	*61.18*	4.82	20	66	836000	4	八天神社例大祭
7	*60.87*	3.13	13	64	1180000	6	大人歌舞伎
8	*60.35*	2.65	11	63	1400000	7	出の山ホタル恋まつり
9	*59.97*	6.03	25	66	625000	4	サン・サン・さんわフェスティバル
10	*59.94*	5.06	21	65	819000	5	くも合戦

表 3：Web ページの順位が低く、welcomekyushu の順位が高い 10 件

順位	*rg*-rk*	rg*	rg	rk	g	k	イベント
303	***23.07***	32.07	133	9	25700	93	皿山まつり
304	***24.37***	44.37	184	20	10200	65	早岐茶市
305	***24.94***	55.94	232	31	3970	47	長串山つつじまつり
306	***25.42***	29.42	122	4	32700	112	薩摩のひなまつり
307	***27.94***	28.94	120	1	36300	603	山鹿温泉祭
308	***32.40***	50.40	209	18	6100	67	仙酔峡つつじ祭り
309	***32.71***	48.71	202	16	6790	72	筑後吉井おひなさまめぐり
310	***34.30***	46.30	192	12	8980	83	宝泉寺ホタル祭り
311	***36.92***	42.92	178	6	10900	101	桜島納涼観光船
312	***41.12***	51.12	212	10	5950	90	人吉球磨は、ひなまつり

表 2、3 の各項目は下のとおりである。

rg*-rk: Google 順位と welcomekyushu 順位の差

rg* ： Google での順位の正規化（1-68）

rg ： Google での順位（1-282）

rk ： 九州観光推進機構サイトの順位（1-68）

g ： Google での件数

k ： welcomekyushu での件数

5．関連研究

従来から、Web や新聞記事からの情報収集と抽出のテーマとして、観光を扱った研究がある。例えば、[斉藤大内 2001]では、観光情報を客観的に定義するため、WWW 文書からキーワードを抽出し、共起関係を可視化するキーマップを提案し、北海道や沖縄を観光をキーワードした検索結果から得られる具体的なキーマップを示している。[金城大内 2001]では、「北海道 観光」というキーワードの検索結果として得られる上位 100 件のページにおいて TITLE,HREF,COLOR で囲まれるキーワードを分析し、「イベント」や地域名など、観光に関連のあるキーワードが多く現れることを示している。個人の好みに応じた旅行プランを提供するためには、観光情報を収集し更新する必要があるという観点から、 [Esparcia 2010] は、観光情報推薦システムの add-on 機能として情報抽出エージェント(IE)と、それを分類する情報分類エージェント(IC)の有用性を示している。目的サイトから、特定のパターンに従って、サービスの名前、場所、値段、時間、期間などを抽出することで、第三者によるサービスについての情報も提供できることなどを示している。

近年、一般の利用者による情報発信の量が飛躍的に増加し、観光情報についても旅行ブログの価値が認められるようになっている。旅行関連業者から提供される情報だけでなく、一般の利用者による具体的な体験談や評判情報に注目が集まっている。[石野他 2009]では、ブログの個別エントリに含まれ旅行、ツアー、出発や地名など 416 個の素性を使った機械学習により、旅行ブログの判定法を提案している。また、抽出すべき観光情報として、同一文中に共起する地域名と土産物の対を考え、提案手法による 17,268 件の旅行ブログの全ての文(8 万文)、一般ブログ、一般ウェブから任意に選んだそれぞれ 8 万文にから 482 個の地域名と土産物の対を抽出する実験を行い、抽出性能を比較している。[Hao 他 2010]では、地域を特徴付けるキーワードを旅行ブログから抽出する手法を提案している。具体的には、ブログのインデックス化における単語*文書行列において、まず単語を local topic と global topic に分け、単語*文書行列を local topic*文書行列と global topic*文書行列の重み付きの和として分解し、さらに、local topic*文書行列を local topic*location 行列と location* 文書行列の積として分解する Location Topic(LT)モデルを提案している。米国の観光地についての 10 万件の旅行ブログに適用し、有効性を示している。

これまではキーワード検索が観光情報の主な利用のされ方だったが、自然言語による検索や、関連情報の提供や推薦が考えらえるようになっている。 [Ruiz-Martinez 2009]では、すでに構築された観光

についてのオントロジーを、自然言語で利用するためのインターフェースを提案している。Murcia 地方の観光情報について、アトラクション、ゴルフコースがあるホテル、教会の近くにあるホテル、レストランなどの 20 通りの検索について、提案システムにより自然言語のまま検索したものと、SPARQL 検索コマンドとして書き直して検索したものと比較し、提案手法の有効性を示している。[小作他 2004]では、観光地に不慣れなユーザーでも容易に観光コースを作れるように、観光イベントの周期性、不定期な観光イベントでも記述形式や共起する単語などの類似点を利用し、ユーザーの検索に対する検索拡張として適切な観光イベントを提示する方式を提案している。3 万件の新聞記事について提案手法を適用し、奈良県の観光イベント 1425 件を抽出する実験を行い、有効性を示している。

最近ではさらに、Web や雑誌による観光情報提供だけでなく、様々なメディアによる観光情報の効果が検討されるようになっている。例えば、[井出 2010]では、映画や TV ドラマなどのロケ地やモチーフなんどの観光情報が、特定地域を訪れる客数の変化にどのような影響を与えたかを分析している。特に、テレビに関する分析の結果、一過性の誘客よりも持続性のあるプロモーションの重要性を指摘している。

6. まとめと今後の課題

本研究においては、welcomekyushu から観光イベント情報を収集し、312 個イベントのリストを作った。これらのイベント情報に関して、WWW 上の一般的な Web ページで扱われている情報量、及び観光専門サイトで取り扱われている情報量の順位差を比較し、弱い相関があることを示した。

また、Google における順位と welcomekyushu における順位が大きく異なるイベントを分析し、このイベントが地域性強い観光ポータルサイトであることを確認できた。

なお、今後我々は２つ以上の九州観光ポータルサイトのイベント情報を集めて、地域性強い観光ポータルサイトを見つける手法を提案する。このために、イベントを沢山収集し、イベントのリストの作る方法の検討が必要となる。また、収集するイベントの同一性の判定方法も考える必要があること等の課題があり，今後はこれらの点を踏まえて研究を行う。

参考文献

[Abdi 2007] H. Abdi, Kendall rank correlation. In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage,2007.

[Esparcia 2010] S. Esparcia, V. Sanchez-Anguix, E. Argente, A. Garcia-Fornes, V. Julian, Integratiing Information Extraction Agents into a Tourism Recommender System, Proc. HAIS2010, Springer LNAI 6077, pp.193-200, 2010.

[Hao 他 2010] Q. Hao, R. Cai, Ch.Wang, R. Xiao, J.-M. Yang, Y. Pang, L. Zhang, Equip Tourist with Knowledge Mined from Travelogues, Proc. WWW2010, pp.401-410, 2010.

[井出 2010] 井出明, コンテンツ誘発型ツーリズムの現状と問題点, 日本知財学会, 2010.

[石野他 2009]石野亜耶, 難波英嗣, 田熊遥, 尾崎貴紘, 小林大祐, 竹澤寿幸,ブログからの観光情報の自動抽出, 電子情報通信学会第 15 回 Web インテリジェンスとインタラクション研究会，pp.19-23, 2009.

[金城大内 2001]金城伊智子, 大内東, 北海道観光情報のための Web データ分析に関する研究, 電子情報通信学会技術研究報告 DE2001-07, pp.99-104,2001.

[Ruiz-Martinez 2009]J. M. Ruiz-Martinez, D. Castellanos-Nieves, R. Valencia-Garcia, J. T. Fernandez-Brieis, F. Garcia-Sanchez, P. J. Vivancos-Vincente, J. S. Castejon-Garrido, J. B. Camon, R. Martinez-Bejar, Accessing Touristic Knowledge Bases through a Natural Language Interface, Proc. PKAW2008, Springer LNAI 5465, pp.147-160, 2009.

[小作他 2004]小作浩美, 内山将夫, 井佐原均, 河野恭之, 木戸出正継, 新聞記事コーパスでの単語出現特徴を利用した観光イベント情報の検索支援, 人工知能学会論文誌, Vol. 19, No. 4, pp.225-233, 2004.

[斉藤大内 2001]斉藤一, 大内東, 観光情報に関する概念形成のための WWW 文書の可視化方法の検討, 電子情報通信学会技術報告 DE2001-07,pp.261-267, 2001.