# ホテル情報とブログ情報のマッチングシステム

呉, 小斌
九州大学大学院システム情報科学府

曾, 駿
九州大学大学院システム情報科学府

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：情報処理学会研究報告. 2011, pp.1-4, 2011-03. 情報処理学会九州支部
バージョン：

# ホテル情報とブログ情報のマッチングシステム

呉　小斌[†2]　曾　駿[†2]　殷　成久[†1]
中藤哲也[†1]　廣川佐千男[†1]

旅行サイトにある、ホテルや旅館を利用した人のコメントは他の利用者に役に立つ情報である。しかし、外部のブログサイトには、そのホテル周辺の情報や、その時の旅行記など、もっと役立つ情報がある。本発表では、宿泊施設自身が提供する情報と、一般的 Web ユーザの個人的なブログ情報のマッチングシステムの可能性について考察する。

# Matching System of Hotel Information and Blogs

Xiaobin Wu,[†2] Jun Zeng,[†2] Chengjiu Yin,[†1]
Tetsuya Nakatoh[†1] and Sachio Hirokawa [†1]

Travel portals provide many information on hotels and inns. They are useful information for tourists who are considering to visit the location and the area. However, the external blog sites contain much more individual opinions and experiences of tourists, which can not be found in the official portal sites. This paper considers the possibility of the matching the information offered by portals and by blog sites.

## 1. はじめに

インターネット上の情報は日々増え続けている。またそれを検索する技術も進歩し続けており、今日では一般個人が非常に大量で多種な情報を収集することが可能となった。観光に関する情報に関しても例外ではなく、実際に多くの人が観光に出かける前に、目的地や宿泊に関する情報をインターネットで収集している。特に宿泊施設に関する情報は宿泊サービスを提供する側から、あるいはそれを紹介するサービスによって、オフィシャルな情報が大量に発信されており、ほぼ全ての宿泊施設に関する情報をインターネット上で発見することが可能である。

一方、オフィシャルな情報とは別に、一般ユーザの個人的な体験に基づく感想などを集めた口コミサイトや、個人的な旅行記や感想などを記したブログ記事なども存在する。近年では、客観的な情報の収集を目的として口コミサイトが人気を博しているが、施設などの対象に直結している分、辛口のコメントが書き辛い状況がある他、サイトによっては恣意的なコメント操作が行われている場合もあり、中立性が少しばかり憂慮される。

それに対して個人によるブログ記事は、対象の施設などとは無関係に個人的な経験や感想を表現しやすく、中立性の問題は少ない。また、対象の施設を含んだ観光の記録の側面があり、周囲の観光情報や飲食店に関する情報など、対象施設に関連した幅広い情報が収集可能であると思われる。しかしながら、ブログ記事は対象としたい施設を明記しておらず、記事の内容から施設やサービスを明確にすることは容易ではない。

我々は、インターネット上の観光情報を収集・分析し、ユーザに分かりやすく提供するシステムの研究を行っている[3)–5)]。本発表では、観光に関する情報の一つとしてホテル・旅館などの宿泊施設（以下、まとめてホテルと記す）を取り上げ、ホテル情報とブログ記事とのマッチングを試みる。

## 2. 実験データの収集と基本的分析

### 2.1 実験データの収集

ホテルに関する情報としては、ホテル自身によって発信される公式ホームページに記載された情報がある。また、それとは別に、旅行社の予約サイト等に登録されている情報もある。前者に関しては、現時点においても公式ホームページを持たないホテル、あるいは公式ホームページに充分な情報の無いホテルなどが多く存在する。後者では殆どのホテルに関しての情報が揃っている反面、テンプレートに従った情報のみで、ホテルの特色が殆ど現れない。本研究では、両方の情報を用いるために、ホテル名をクエリとして一般検索エンジンから得られた情報を、ホテルに関する情報とした。

ブログから得られる情報（以下、ブログ情報）に関しては、非常に多くのブログサイトがあり、また個人で運営しているブログもあるので、それらの情報を漏れなく収集することは

†1 九州大学 情報基盤研究開発センター
Research Institute for Information Technology, Kyushu University
†2 九州大学大学院 システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University

**表 1** ホテル情報の高頻度単語 上位 50 単語
Table 1 Higher Rank 50 Words with High Frequency at Hotel Information

| | | | | |
|---|---|---|---|---|
| ホテル (3519) | 福岡 (3518) | 情報 (3299) | 予約 (3247) | 宿泊 (3185) |
| 利用 (3002) | 駅 (2917) | ビジネス (2709) | 施設 (2575) | ページ (2549) |
| サイト (2543) | プラン (2502) | 温泉 (2422) | ください (2389) | 詳細 (2349) |
| 室 (2343) | 料金 (2305) | 場 (2233) | アクセス (2229) | 観光 (2208) |
| 無料 (2094) | 旅館 (2091) | 周辺 (2082) | 券 (2080) | 旅行 (2038) |
| 地図 (2034) | 宿 (2009) | 一覧 (1989) | 徒歩 (1976) | サービス (1964) |
| 博多 (1937) | 検索 (1934) | 部屋 (1903) | 駐車 (1889) | 航空 (1877) |
| バス (1874) | 区 (1849) | 九州 (1825) | チェック (1815) | 朝食 (1768) |
| ポリシー (1747) | 国内 (1730) | おり (1717) | 確認 (1696) | 住所 (1693) |
| でき (1689) | 問い合わせ (1683) | 場合 (1677) | アウト (1675) | マップ (1669) |

容易ではない。収集自体が研究対象となっている。本研究は、ホテル情報とのマッチングを目的としているので、新たなブログ記事の収集は行わず、これまでの研究で収集済みのブログ記事を用いた。

マッチングの実験データとして、先ずは対象データを福岡県のホテルに限定した。旅行の口コミサイト フォートラベル[⋆1]に掲載されている情報から、福岡県内のホテル名 401 件を収集した。更に、収集したホテル名 401 件をクエリとして google 検索エンジンからそれぞれ 10 ページの関連ページを収集し、計 4010 件をホテル情報とした。また、九州観光推進機構[⋆2]が選定し推薦している、九州地区の観光に関するブログ記事[⋆3]、1303 件をブログ情報とした。

### 2.2 実験データの基本的分析

マッチングの実験に先立って、用いるデータの基本的分析を行った。先ずはホテル情報全体で高頻度に出現している単語を求めた。出現頻度が上位 50 位までの単語リストを**表 1** に示す。() 内の数値は出現頻度である。これらの単語はホテルの基本的な情報に関するものであり、それ故、特徴を示すものではない事が見て取れる。

次に、ホテル情報における特徴語を調べた。クエリキーワードを与えて GETA[⋆4]で検索した結果得られたホテル情報について特徴語を求めた。**表 2** には実例の一つとして、クエリキーワードに「麺」を用い、得られたホテル 88 件のうちの検索順位上位 5 件のホテルについて、特徴語の上位 5 単語を示したものである。

**表 2** ホテル情報をクエリ「麺」で検索
Table 2 Search of "noodle" from Hotel Information

| ホテル名 | 特徴語 上位 5 単語 |
|---|---|
| 福岡 山の上ホテル | 福岡, 情報, 山の上, バス, レストラン |
| 美奈宜の杜温泉　杜の湯 | 美奈宜の杜, 杜, 処, 湯, 風呂 |
| ウィークリーイン二日市 | ウィークリー, 筑紫野, イン, 予約, 情報 |
| 博多温泉　旅館　富士の苑 | 苑, 三宅, 富士, 掛け, 流し |
| ホテルハミングバード中央イン | ホテルハミングバード, 清川, 中央, レジャー, 快適 |

**表 3** ブログ記事をクエリ「麺」で検索
Table 3 Search of "noodle" from Blog Information

| ブログ文書番号 | 特徴語 上位 5 単語 |
|---|---|
| 603 | 酸, のぞい, しゅう, 担, 辣湯 |
| 223 | ピン, 趙, スーラー, ソラリアステージ, 辣 |
| 212 | 酔, モツラーメン, プルプル, トンコツスープ, モツ |
| 184 | ふくま, イカゲソ, 竹中, ほしかっ, 血圧 |
| 23 | 夜中, 替え玉, 薬院, 担, べつに |

これらも、検索キーワードに関係なく、各ホテル情報の特徴語が得られている状態が見て取れる。しかしながら、少なくとも 88 件のホテルに関しては、ホテル情報にクエリキーワードが含まれている。特徴語として、更に幅広い単語を用いることで、食べ物などに関する情報を抽出できる可能性がある。

ブログ情報に関しても特徴語を調べた。同様にクエリキーワードを与えて GETA で検索した結果得られたブログ情報について、特徴語を求めた。**表 3** に、クエリキーワード「麺」を用いて検索した結果得られたブログ情報 121 件のうち、検索順位上位 5 件のブログ文書について、その文書に含まれる特徴語の上位 5 単語を示す。

この分析から、ブログの情報には、クエリに関連する幅広い情報が含まれていると考えることができる。

以上の分析から、ホテルの情報は若干固定的で、一方ブログの情報は幅広いことが想定される。この両者の情報をキーワードで繋ぐために、ホテル情報からキーワードを求め、キーワードからブログ情報を求める 2 段階の検索を行い、共通するキーワードを軸にして関連性を図示するマッチングシステムを設計した。

⋆1 http://4travel.jp/
⋆2 http://www.welcomekyushu.jp/
⋆3 http://www.welcomekyushu.jp/demon-kakka/
⋆4 汎用連想計算エンジン GETA　http://geta.ex.nii.ac.jp/geta.html

**表 4** クエリ単語のリスト
Table 4 Query Word List

| | | | | |
|---|---|---|---|---|
| 麺 (488) | スープ (347) | ラーメン (331) | 料理 (282) | ランチ (165) |
| チャーシュー (157) | 野菜 (142) | 肉 (141) | 豚 (128) | ネギ (120) |
| 酒 (112) | うどん (112) | グルメ (107) | 鶏 (102) | 魚 (101) |
| 醤油 (100) | 卵 (98) | レストラン (97) | 食事 (91) | 牛 (87) |
| 米 (86) | とんこつ (84) | | | |

## 3. マッチングシステムの概要

ユーザにより与えられたクエリを用い、収集済みのホテル情報を検索する。得られたホテル文書を $h_i$ とする。但し、$i$ は検索結果のランキングである。検索件数を $N$ とするとき、検索結果全体のホテル情報を $H$ とし、また、$H$ のうち、検索結果の上位 $n$ 件までのものを $H'$ とする。

$$H = \{h_i | 1 \leq i \leq N\}$$
$$H' = \{h_i | 1 \leq i \leq n\}$$

ホテル情報 $H$ に含まれる文書から GETA を用いて特徴語を求め、$w_x$ とする。但し、$x$ は検索結果のランキングである。得られた特徴語の上位 $k$ 件を $W$ とする。

$$W = \{w_x | 1 \leq x \leq k\}$$

次に、特徴語集合 $W$ を用いてブログ文書を検索し、得られたブログ文書を $b_j$ とする。但し、$j$ は検索結果のランキングである。得られたブログ文書の上位 $m$ 件の集合を $B$ とする。

$$B = \{b_j | 1 \leq j \leq m\}$$

ホテル情報 $H'$ と、ブログ情報 $B$ の関連について、特徴語 $W$ による共起関係をラベルとした 3 部グラフを生成して、ユーザに提示する。その際、ラベルを得られなかったノードは描画しないものとする。

## 4. マッチングシステムによる検索例

設計したマッチングシステムを実装し、実際に検索を行った。ホテル情報の描画数 $n = 20$、特徴語の個数 $k = 20$、ブログの個数 $m = 10$ とした。検索用には、ブログ中に出現している食べ物関係の単語を抽出し、そのうち高頻度語 22 個をクエリとして選定した。選定されたクエリを**表 4** に示す。

これら 22 個のクエリの各々について 3 部グラフを描画したところ、2 個のクエリではラベルが生成されなかった。これは、制限されたホテル情報と共起しなかったものと思われる。その他のクエリでは、3 部グラフが生成された。グラフに描画されたノード数は、特徴語が平均 5.7、ホテルのノード数が 4.3、ブログのノード数が平均で 8.4 であった。

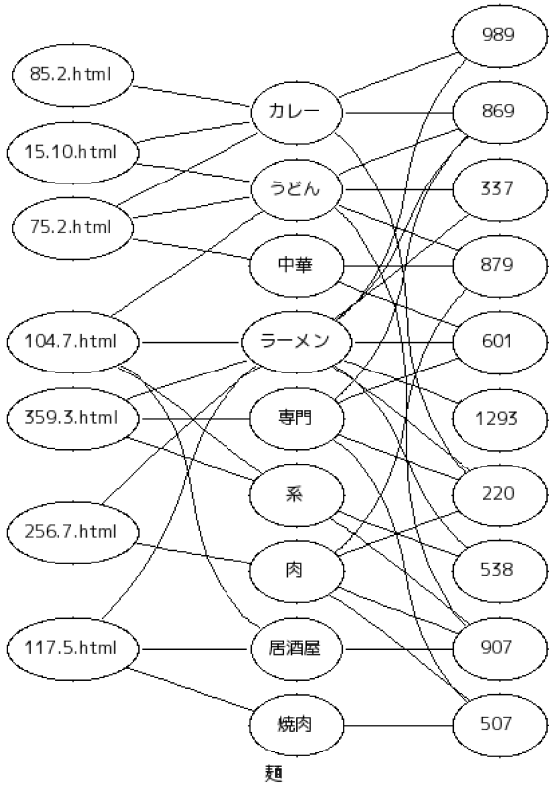


**図 1** 「麺」によるマッチングの例
Fig. 1 Example of Matching by "noodle"

実例として、「麺」をクエリとした検索結果を**図 1** に示す。左の楕円がホテル文書 ID、中央の楕円が共起するキーワード、右の楕円がブログ記事 ID を示している。

グラフからは、ホテル情報とブログ情報が共通キーワードを軸に関連している様子が実際に

見て取れる。特に左上に表示されたホテル文書（ID が 85.2.html, 15.10.html, 75.2.html のもの）は、軽食系のキーワードで多くのブログと繋がっている。また、左下の ID が 117.5.html のホテル文書は、ラーメンと言うキーワードで多くのブログと関連している一方、居酒屋や焼肉と言ったキーワードでは、ただ１つのブログと関係しており、キーワードによっては特定のブログと強い結びつきを持つ事が読み取れる。

## 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、6) 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。7) 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、8) 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。10) 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。9) 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。

ブログ情報からのマッチングに関する研究としては、1), 2) がある。何れもブログの記事と Wikipedia エントリをマッチングすることを目的としており、本研究の目的であるコンテンツ同士の相互関係を想定したマッチングとは質が異なる。

## 6. まとめと今後の課題

本発表では、インターネット上の観光情報を収集・分析し、ユーザに分かりやすく提供するシステムの一部として、ホテル情報とブログ記事とのマッチングを試みた。ブログ記事の内容とホテル情報が、共通のキーワードで有機的に結合される事を確認した。一方で、全てのクエリに対して、マッチングが得られるわけではない事が明らかになった。これは、ブログ記事の総数が不充分であるためと思われるため、今後は多量のブログ記事を用いた検証を行う予定である。また、本発表ではブログ記事の内容とホテル情報のマッチングを試みたに留まっているので、今後は有用で新たな知識が獲得できるようなシステムの構築を目指す予定である。

## 参 考 文 献

1) 佐藤由紀, 横本大輔, 中崎寛之, 宇津呂武仁, 吉岡真治, 福原知宏, 神門典子, 中川裕志, 清田陽司, Wikipedia を介した関連ニュース・ブログの対応付け - Wikipedia エントリの分析 - , 情報処理学会, Vol.2009-NL-194 No.10 研究報告 - 自然言語処理（NL）(2009)
2) 川場真理子, 中崎寛之, 宇津呂武仁, 福原知宏, Wikipedia エントリとブログサイトの対応付けによる日本語ブログ空間のトピック分布推定, 情報処理学会, Vol.2008 No.90 研究報告 - 自然言語処理（NL）(2008)
3) 殷成久, 呉小斌, 廣川佐千男, 中藤哲也, 観光イベントについての「といえば検索」の提案, 電子情報通信学会技術研究報告 110(301), pp.43-47, 2010
4) X. Wu, S. Hirokawa, C. Yin, T. Nakatoh, Y. Tabata, Extraction and Comparison of Tourism Information on the Web, Proc. AROB2011, 2011
5) C. Yin, T. Nakatoh, S. Hirokawa, X. Wu, J. Zeng, A proposal of search engine “ XYZ ”for tourism events, Proc. JCAI2010, 2010
6) 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
7) 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
8) 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
9) 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.
10) 石野亜耶, 難波英嗣, 田熊遥, 尾崎貴紘, 小林大祐, 竹澤寿幸, 旅行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 WI2 研究会, pp.19-23, 2009.