

テキストマイニングを応用した観光言説分析の提案

守屋 豊[†] 井出 明[‡]近畿大学[†] 首都大学東京[‡]

1. 研究の背景

現在、世界観光機関（WTO）は観光統計として TSA（ツーリズム部門会計：Tourism Satellite Account）を提唱している[1]。これは各種の統計情報を元に観光がどれだけ GDP に貢献しているかについて測定するもので、主として経済、経営面についての分析である[2]。そのためこの分析からでは、観光地の「魅力」や「個性」などを明らかにすることはできない[3]。例えばパリという街について考えてみた場合、ビジターにとってどのようなイメージで捉えられているのかといった全体像を計量的に把握することが不可能である。

筆者らはこれまで、観光情報の検索時における検索漏れを防ぐためのツールとして、観光に特化した言語シソーラスの構築を提案し、実験を行って来た[4]。その中で、ある地域についてビジターから寄せられた口コミを元に構築したシソーラスは、その地域の特徴的な情報を示していることが明らかとなった。

そこで本稿では、国際的な観光地であるパリについて、日本と欧米圏の観光関連ブログからコメントを抽出し、テキストマイニングを行い、それぞれについての言語シソーラスの構築を行った。その上で、文化の違いによってシソーラスに現れる語に変化が生じるか調査し、最終的にはテキストマイニングが観光の言説分析手法の一つとなり得るのかについて検討した。

2. 先行研究

実験に先立ち、国内外におけるテキストマイニングを用いたコンテキスト分析の先行研究を調べた。特徴的な 3 件について概要をまとめた。

まず、野田らの研究では、ユーザの入力したキーワードを手がかりとしてインターネット上の情報からそれに関連する話題を自動抽出する手法について述べられている。実装としては、検索サイトに表示されるページサマリに対して形態素解析を行い話題語を求め、それを元にキーワードと適合する Web ページを探し出す実験が行われている[5]。

Chang Choi らは観光情報サイトで、入力された利用者のプロフィールを元にユーザの好みを分析し、推論エンジンを用いてデータベースからユーザごとに興味のあるような情報を選択し提示する技術について提案している。推論の過程で語の関連性を明確に定義するためにオントロジーが使用されている[6]。

Lobna Karoui らの研究では、インターネットのウェブサイトの HTML 構造を解析し、単語の共起関係や重み付けを利用して文脈を読み取る手法が用いられている[7]。

以上のような先行研究調査の結果、これまでのところ、観光についての口コミレビューに対してテキストマイニングなどの図書館情報学の手法を導入するといった試み

は行われておらず、観光情報学の新たな分析ツールとして、テキストマイニングの手法を用いることは十分な合理性と新規性があると考えられる。

3. パリを対象とした観光言説分析

実験対象には世界最大の観光地であるパリを選んだ。なぜなら 2006 年におけるフランスの外国人訪問者数（暫定値）は約 7,910 万人（その内、日本人が約 65 万人）を記録しており、世界で最も人気を博している観光地であると言えるからである[8]。

3.1. 実験方法

実験は、対象について言及されたコメントを抜き出しテキストマイニングを行って、そこから抽出された語を用いてシソーラスを構築するという手順を踏んだ。以下では日本語、英語それぞれについて分析手法を述べる。

3.2. 日本語レビューの分析手法

日本語におけるテキストマイニングの言語データには、「フォートラベル」を用いた[9]。これは日本人の海外旅行者が、現地での体験情報をブログ形式で書き込めるものである。現在、会員数は 4 万人を超しており、活発に利用されている。まず、同サイトに存在するパリに関するレビューで 2006 年に投稿されたもの全 391 件から 300 件を取り出した。それらに対して形態素解析システム“ChaSen”を用いて品詞分解を行い、助詞、助動詞、接頭詞、数詞、連体詞などの今回の分析上意味を持たない品詞を選別し省き、名詞、形容詞といった分析上有用な語については、その出現頻度を測定した[10][11]。その後、同義語はひとまとめに、表記上の揺れについては統一し、上位に昇った語句からカテゴリごとに整理した。

3.3. 英語レビューの分析手法

次に英語で書かれたレビューについてであるが、情報源は“travelblog”を用いることとした[12]。これは欧米圏で主に用いられているもので「フォートラベル」と同様にフリーの登録をすれば国、地域ごとに、レビューの投稿、閲覧が可能なものとなっている。

日本語との正確な比較が可能なように同様に 2006 年に書き込まれた 410 件から 300 件を抜き出した。“ChaSen”は現時点では英語の解析に対応していないため、こちらの解析については“KWIC Concordance for Windows”を用いた[13]。解析後、有用語を取り出し、日本語と同様のカテゴリ分類を行った。同義語、表記揺れもここで統一した。以下にその一部を図で表現した。

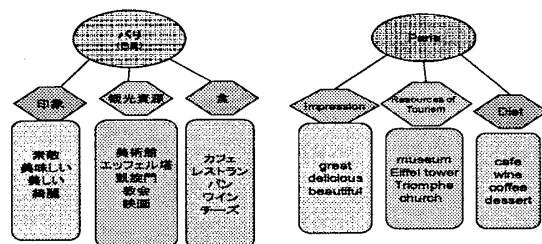


図. パリを対象とした観光情報ソースの比較 (一部分)

4. 考察

3で構築した2つのシソーラスの比較を行った。各カテゴリの語を比較したところ、「印象」、「食」、「交通」などのカテゴリにおいては日本語、英語共に出現頻度の多少の差はあるものの全体的な語構造自体に大きな変化は見られなかった。しかし日本語側の「観光資源」のカテゴリに分類されたものに注目したところ、映画に関する記述、特に『ダ・ヴィンチ・コード』に関連するものが多く挙げられることが分かった。『ダ・ヴィンチ・コード』はアメリカの作家ダン・ブラウンの著作で、日本では2004年に出版されている。フランス、イギリス、ローマを舞台とした推理小説であり、2006年には映画も世界同時公開された。日本語側において『ダ・ヴィンチ・コード』に関する語は表記揺れを合わせると41件見られた。それに対して、英語側の記事においては10件しか見られなかった。ルーブル美術館についての言及は日本語では228件、英語では241件と大差がなかったが、その美術館前にあり、小説内で重要な役割を果たすピラミッドのモニュメントに関する言及は日本語側で93件と非常に多かった。また「映画」という語そのものに関する記述にも日本語側では111件に上るのに対し、英語側では43件に留まっていた。ここから、この時期における日本のビジターの観光行動には映画のロケ地を見て回ったというのが、欧米圏のビジターのそれと比較して多かった、ということが推定できる。

表「観光資源」カテゴリ語の計量比較 (一部分)

ダ・ヴィンチ・コード,ダビンチコード...	41
Code	10
ルーブル美術館	228
Louver	241
ピラミッド	93
Pyramids	14
映画	111
movie,movies,film...	57

(「フォートラベル」, “travelblog”から2006年1月1日を起点として、パリに関するレビュー300件を時間軸に沿って抽出)

5. まとめと今後の展望

今回、日本語と英語で書かれた記事について、それぞれにテキストマイニングを行ってシソーラスを構築し、

計量的な分析を試みた。その結果、パリにおいてある時期の観光資源の捉え方並びに観光行動が、日本人と欧米人では異なっていたことが読み取れた。こういった文化の違いによる観光行動の違いはツーリズムリサーチの観点からは非常に重要とされるが、これまでは調査票の配布やインタビューといった限定的な方法によってのみ調査されてきた^[14]。情報化社会の進展に伴って、ネットワーク上に多くの観光レビューが蓄積されるようになった結果、このような新しい分析の提案が可能となった。もちろんシソーラスという形を用いるため語単位での分析になってしまい、調査の内容によっては語から再び文脈を辿る必要が生じることもありうる。しかし、このような計量化を行うことで、数百件の記事をただ漫然と眺めていただけでは気づき難い事象を発見することが可能となり、サプライヤーにとって有用なマーケティングデータを掘り起こすツールとして利用できる可能性を秘めていると考えられる。

また、今回はコメントの収集から形態素解析を行うまでの手順を手作業で行ったが、この部分を計算機による自動化を行うことによって、さらに大規模なテキストの解析を図ることが可能になると考えている。また、時間軸や季節などのパラメータの付与、他の観光地との対照実験についても計画している。

なお実験結果については今後、観光系学会においても報告することを予定しており、観光の専門家からの意見も取り入れ学際的な観光情報学の研究として発展させたいと考えている。

参考文献

- [1] WTO: Statics & Economic Measurement of Tourism (<http://www.unwto.org/statistics/>) (2007年12月1日確認)
- [2] 佐竹真一「観光資源評価の基礎概念」『日本観光学会誌』第48号 No.48 June 2007 pp.68-80 日本観光学会
- [3] 塩谷英生「観光消費の経済効果の推計：観光統計の現状と TSA の登場」『オペレーションズ・リサーチ：経営の科学』Vol.50 No.1 pp.17-22 2005年 日本オペレーションズ・リサーチ学会
- [4] 守屋豊・井出明「『伊勢・志摩』を対象とした観光情報シソーラスの構築実験」『マルチメディア、分散、協調とモバイル (DICOM2007)シンポジウム論文集』2007年 pp.1464-1470 情報処理学会
- [5] 野田武史 他「主題語からの話題語自動抽出とこれに基づく Web 情報検索」『情報処理学会研究報告 データベース・システム研究会報告』Vol.2006, No.78 pp. 305-311 情報処理学会
- [6] Chang,Choi. et al. “Travel Ontology for Recommendation System based on Semantic WebAdvanced Communication Technology.” ICAC2 2006. The 8th International Conference Volume 1, (2006) pp.624-627
- [7] Lobna,Karoui. Et al. “Context-based Hierarchical Clustering for the Ontology Learning” Web Intelligence archive Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (2006) pp.420-427
- [8] 国際観光振興機構: 世界の国際観光の動向 (http://www.jnto.go.jp/jpn/tourism_data/global_tourism_trends.html) (2007年12月1日確認)
- [9] フォートラベル (<http://4travel.jp/>) (2007年12月20日確認)
- [10] “ChaSen”(http://chasen.naist.jp/hiki/ChaSen/) (2007年12月10日確認)
- [11] 林俊克『Excelで学ぶテキストマイニング入門』オーム社2002年 pp.51-71
- [12] travelblog (<http://www.travelblog.org/>) (2007年9月20日確認)
- [13] “KWIC Concordance for Windows” (http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/) (2007年9月10日確認)
- [14] ボニータ・M・コルプ著 近藤勝直監訳『観光都市のマーケティング』多賀出版2007年 pp.119-138