

観光記事における概要文章の個人適応に向けた 文章分類手法の提案

三笠 弘貴[†] 奥野 拓[‡]

公立はこだて未来大学大学院[†] 公立はこだて未来大学[‡]

1. はじめに

観光情報サイトや個人旅行ブログには、各記事を端的に説明する概要文章が存在する。閲覧者は、それらの概要文章やサムネイル、記事のジャンルなどを、本文を閲覧する際の判断基準として用いる。しかし概要文は性質上、記事中の観光トピック(食事、宿泊施設、交通手段、等)について網羅的かつ詳細に記述することは難しい。また、閲覧者ごとに検索目的や求める情報は異なるため、概要文章を提示する場合、すべての閲覧者の要求を同時に満たすことは困難であると考えられる。

本研究では、概要文章を閲覧者の興味に応じて動的に生成することで、情報検索を支援する事を目指す。本稿では特に、旅行ブログを対象に概要文章と本文中に存在するトピック数を比較調査し、概要文章の現状について考察を行う。

2. 概要文章の記述における制約と問題

一般的なサイトの記事一覧ページ内には、各記事の概要文章やタイトル、サムネイル画像などが用意されており、各要素が果たす役割は異なる。図1は観光情報サイトにおける概要文章と詳細文章の一例である。

Yesilada ら[1]は、アイトラッキングによりページ内の要素の役割について明らかにしている。渡辺らは、記事タイトルが閲覧者の興味を引く役割を持つ一方で、スニペット(概要文章)は最終的に詳細文章を閲覧するかどうかを判断する役割を果たすと考察している[2]。

上記の役割を果たすために、多様な閲覧者の検索目的を満たすよう本文の情報を網羅的に盛り込むことが望ましい。しかし、本文を網羅的に解説することと文章量の抑制はトレードオフの関係にあり、両方のバランスがとれた概要文章の作成には高度な文章力が必要である。

3. 検索目的に基づくスニペット生成

ウェブ検索の分野において、ユーザの検索目的から検索結果ページにおけるスニペットを構築し提示する研究として、高見ら[3]の研究がある。スニペットの種類を2つの軸で分類し、各検索目的に合わせたスニペットを動的に生成する手法を提案している。



図1: 函館市公式観光情報サイトはこぶらにおける概要文章と詳細文章

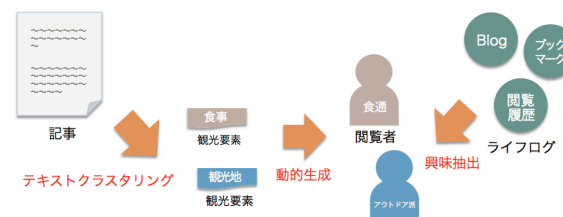


図2: 閲覧者の興味に基づく概要文章の動的生成手法

4. 概要文章の動的生成と個人適応

2節で述べたように簡潔さを維持したまま閲覧者の多様な検索目的を満たす概要文章を作成することは非常に困難である。

そこで、図2のように閲覧者ごとの興味に応じて記事から動的に概要文章を生成することで、この問題の解決を目指す。本提案手法は以下の3つの段階からなる。

- 観光トピックに基づくテキストクラスタリング
- 閲覧者の観光興味抽出
- 観光興味による概要文章の動的生成

本稿では、観光トピックに基づき観光記事をクラスタリングすることで、本手法の必要性について考察する。

5. 観光トピックに基づくテキストクラスタリング

各地方の観光サイトを対象に、共通するカテゴリーを簡易調査したところ、観光トピックとして「食事」「名産・土産」「観光スポット」「イベント」「宿泊」「交通」「歴史」「芸術・文化」が挙げられた。

次の2つの例文は、函館旅行の記事の一部分である。

- 幾つか分類方法を検討夕方最終目的地の函館に到

表 1: 観光記事における観光トピックの記載率

	食事	お土産・名産	観光地	イベント	宿泊	交通	歴史	芸術・文化
概要文章における記載率	20%	33%	10%	49%	38%	42%	14%	14%
詳細文章における記載率	84%	82%	67%	86%	93%	86%	60%	64%

表 2: 概要文章と詳細文章における平均トピック数

	概要文章	詳細文章
平均トピック数	2.22	6.23

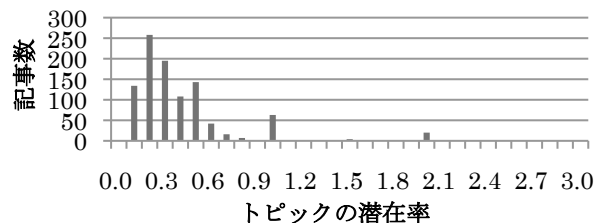


図3: 概要文と詳細文におけるトピック数の比率と、記事の分布

着しました。

- まずは空港から函館駅までバスで1時間ほど。

1文目は観光スポット、2文目は交通手段に関して述べられている文である。他の文についても、それぞれ1つから2つ程度のトピックが存在することが分かった。

本研究では、前述した観光トピックをクラスターとし分類を行うこととする。

6. 実データの収集

6.1 調査概要

旅行ブログポータルであるフォートラベル (<http://4travel.jp>)を対象とし、函館旅行に関する記事1209件を収集した。記事は、2008年10月24日から2013年1月1日に投稿されたものであり、それぞれ書き手や旅行時期なども異なっている。また、記事ごとに概要文章と詳細文章が存在し、概要文章は10字から250字程度、詳細文章は10字から3万字程度である。

記事の中には、概要文章に旅行のあらすじを記し、詳細文章には写真のみを記載するアルバムスタイルの記事が存在したため、本調査ではそれらを除外した記事994件(約46000文)についての分析を述べる。

記事中の文の分類については、条件付き確率モデルを利用したナイーブベイズ分類器を用いた。教師データとして、函館市公式観光情報サイトはこぶら (<http://hakobura.co.jp>)の記事、約250件(約1200文)を用いた。各記事はカテゴリーごとに記載されており、あらかじめ各文が存在する記事のカテゴリー情報を元に学習させた。

6.2 調査結果と考察

表1は、観光記事の概要文章と詳細文章それぞれに存在する観光トピックの割合について示したものである。詳細文章における記載率に対して、イベントや交通に関しては概要文中でも触れられている事が多いが、歴史、芸術・文化については、概要文章で触れられていない場合が多い事がわかる。

表2は、概要文章と詳細文章のそれぞれにおける平均トピック数を示したものである。概要文章では平均種類以上の2トピックについて書かれているが、詳細文章には6種類以上存在する。この結果から、10から250文字

程度の概要文章では、本文の3割程度のトピックしか説明できない。

図3はトピックの潜在率を表したものである。91%の記事は潜在率が1未満の値であり、約6%の記事は潜在率が1の値となった。つまりこれら約90%の記事については、詳細文章中に含まれるトピックより概要文章に含まれるトピックが少ないため、概要文章から本文のすべてのトピックを推察することは難しいと考えられる。また7割の記事については潜在率が0.5以下であり、本文中の観光トピックの半分は埋もれてしまっているため、詳細文章に目を通して確認する必要がある。

これらの結果から、現状のブログ記事における概要文章からは詳細文章に存在するトピックを知ることは難しく、詳細文章を閲覧するかどうかの判断材料として十分なものであるとは言えない。

7. 概要文章の個人適応に向けて

概要文章の生成手法を提案した上で、観光記事の収集し記事中の文章を分類することで、観光ブログにおける概要文章の現状を調査し明らかにした。

今後は概要文章の個人適応に向け、閲覧者の観光興味抽出や概要文章の動的生成の手法について検討を行なう。

参考文献

- [1] Yesilada, Y., Jay, C., Stevens, R. et al.: Validating the use and role of visual elements of web pages in navigation with an eye-tracking study, Proc. WWW'08, ACM, pp.11-20 (2008)
- [2] 渡辺奈夕子ら, Web 検索結果の推薦における提示項目が印象に与える影響, 情報処理学会, 2009.3.13, P61-68
- [3] 高見真也ら, 検索目的に基づくスニペットの動的再生成によるウェブ検索結果の個人適応化, 日本データベース学会, 2007.9 P33-36