一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

旅行ブログエントリ閲覧システムのためのテキスト要約モジュール

飯沼 俊平　難波 英嗣　竹澤 寿幸

広島市立大学大学院 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東三丁目 4 番 1 号
E-mail: {iinuma,nanba,takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし ブログなどのソーシャルメディアでは, 実際に観光地を訪れた旅行者の経験に関する情報を得ることができ，それらは訪問場所や宿泊施設を選択する際に，大いに役立つ情報である．本研究では，旅行ブログエントリ閲覧システムを構築し，ユーザが指定した地理的範囲のエントリを要約して提示する機能を実装した．要約にはグラフベースの手法を用いて，文と同様の仕組みで画像の重要度を算出し，代表画像付きの要約を出力する手法を提案した．評価実験の結果，ベースラインよりも高い性能で要約を作成できることが確認できた．

キーワード 自動要約，観光情報処理，旅行ブログ

Text Summarization Module for the Travel Blog Browsing System

Shumpei IINUMA　Hidetsugu Nanba　and　Toshiyuki Takezawa

Graduate School of Information Sciences, Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
E-mail: {iinuma,nanba,takezawa}@ls.info.hiroshima-cu.ac.jp

Abstract Travelers can obtain information about their destinations from other travelers through social media, such as blogs. This kind of information is useful for choosing destinations and accommodation. In this study, we built a travel blog browsing system, which can generate a summary of user-specified location. We proposed a method for generating a summary with images, using a graph-based method to compute importance scores of images as well as sentences. Our experimental result revealed that our method outperforms a baseline method.

Keywords Automatic Summarization, Travel Information Processing, Travel Blog

1. はじめに

旅行者が旅先の観光情報を収集するために利用する情報源の一つとして，旅行ガイドブックが挙げられる．一般的な旅行ガイドブックには，有名な観光名所，土産物，宿泊施設，飲食店など，観光に関連する基本的な情報が掲載されている．一方，ブログなどのソーシャルメディアでは，実際に観光地を訪れた旅行者の経験に関する情報を得ることができ，それらは訪問場所や宿泊施設を選択する際に，大いに役立つ情報である．我々は，過去の旅行者が投稿した情報へのアクセスを容易にするために，旅行ブログエントリを自動検出し，タイプ分類を行うシステムを構築している[1, 2]．このシステムは広島 P2 ウォーカーで公開されている“ぶらり広島電停 MAP”[1]で使用されている．このシステムにより，目的の観光地に関する情報を地図上で探すことができる．

本研究では，上述のシステムの利便性を向上させるため，ユーザが指定した地理的範囲の旅行ブログエントリを要約して提示する機能を実装する．そのために，

図 1 旅行ブログエントリ閲覧システム

テキストだけでなく画像も対象とした要約手法を提案する．画像付きの要約を提示することで，旅行者は目的地の特徴を容易に把握することができる．本論文の構成は以下の通りである．2 節ではシステムの概要および動作例，3 節では関連研究，4 節では旅行ブログエントリの自動要約手法，5 節では評価実験について述べ，6 節で本稿をまとめる．

[1] http://p2walker.jp/peace/ja/blog/

図 2 サマリービュー
（平和記念公園周辺，タイプ“見る”の要約例）

2. システムの概要および動作例

本節では，構築したシステム[2]の概要および動作例について説明する．図 1 に本研究で開発した旅行ブログエントリ閲覧システムを示す．地図上にエントリ集合を表示しており，画面下のボタンで，“見る”，“体験する”，“買う”，“食べる”，“泊まる”などの，旅行者の目的に沿ったエントリタイプ[2]を選択することができる．また，マーカーをクリックするとポップアップでエントリへのリンク付きのタイトルを表示する．“ぶらり広島電停 MAP”（2015 年 5 月 27 日時点）からの拡張点は以下の通りである．

- リストビュー：表示範囲のエントリ一覧
- サマリービュー（図 2）：表示範囲のエントリ集合の要約

画面左下の緑のボタンをクリックするとリストビューおよびサマリービューが表示され，上部のタブでビューを切り替えることができる．リストビューには単純に表示範囲のエントリの一覧を表示する．サマリービューでは，ユーザが指定したタイプ，表示範囲のエントリ集合をトピックごとに要約した結果と，関連するエントリへのリンクを表示する．これにより，ユーザの指定した地域の見所が発見しやすくなり，さらに，個々のエントリへのリンクを示すことで，より詳細な情報へのアクセスが容易になる．

3. 関連研究

Web 上のリソースを利用して旅先の情報をユーザに提示するための研究が行われている．Wu らは，観光情報を要約するシステムを提案しており，クエリのカテゴリごとに，テキストや画像，動画などの異なるメディアを情報源として選択する手法を提案している[3]．Hao らは，旅先の特徴を表すタグやスニペットを要約として出力するモジュールなど，地域特有の情報をブログから発見する手法を提案している[4]．安田らは，“歴史”や“食べ物”といったトピックと地理的範囲を入力として受け取り，対象範囲の情報を簡潔にまとめた文書を生成する要約手法を提案しており，ブログを要約対象として実験を行っている[5]．エントリに含まれる画像はユーザに視覚的理解を促す有用な情報源であるため，本研究では画像も要約の対象とする．また，ブログ中の文章はくだけた表現が多く，重要文抽出による要約では，意味が不明瞭な文が抽出されてしまう可能性がある．我々は，要約に画像を含めることで，文の意味の不明瞭さを補えると考えた．

文の重要度計算には，文が持つ特徴から回帰モデルで重要度を計算する手法やグラフベースの手法などが用いられてきた．特に，グラフベースの手法はテキストと画像をリンクさせることができれば，テキストと同様の枠組みで画像の重要度を計算できると考え，本研究ではグラフベースの手法を採用した．その代表的な手法としては，LexRank が挙げられる[6]．LexRank は文のグラフ表現における固有ベクトル中心性の概念に基づいて文の重要度を計算する．前田らの研究[7]では LexRank を画像の類似性グラフに適用していたが，本研究では画像と文を一つのグラフで扱う点が異なる．

4. 旅行ブログエントリの自動要約

本研究で扱う旅行ブログエントリは，藤井らの手法により 5 種類のタイプ（“その他”は除く）に分類されていると仮定する．システムは，エントリタイプと地理的範囲を入力として，該当するエントリ集合の要約を出力する．なお，閲覧システムの表示領域を考慮して，1 トピックにつき画像 3 枚，3 文から 5 文程度の短い要約を目標とする．要約は大まかに次の手順で作成する．

1. エントリ集合をクラスタリング
2. クラスタごとに LexRank を適用し，文と画像の重要度を計算
3. 重要度が高い順に文と画像を選択

エントリにはすでに“見る”や“食べる”といったタイプが付与されているが，ユーザが指定した範囲には，さらに複数のトピックが混在する可能性があるため，これらをグループ化する必要がある．本研究では，最遠隣法を用いた階層的クラスタリングを行い，クラスタ間の距離が閾値以下の時にクラスタを統合する方法をとる．なお，エントリは tfidf 値を要素とする文書ベクトルとして扱い，距離関数は $f(c_i, c_j) = 1 - \cos(c_i, c_j)$ を用いる．ここで，$\cos(c_i, c_j)$ はコサイン類似度を表す．

4.1. LexRank による文と画像の重要度計算

LexRank では，まず文間の類似度が閾値以上であれば 1，

[2] http://www.ls.info.hiroshima-cu.ac.jp/blogMap/

それ以外は 0 を要素とする隣接行列を用意する. 作成したグラフから, ノード(文)uの重要度は式(1)で求められる. これは, PageRank と同様, 隣接行列に対してべき乗法を用いて, 主固有ベクトルを計算することで得られる.

$$p(u) = \frac{1-d}{N} + d \sum_{v \in adj[u]} \frac{p(v)}{\deg(v)} \qquad (1)$$

ここで, Nはノードの数(文の数), dはダンピングファクター[8] (Brin らを参考に 0.85 に設定), $adj[u]$はノードuに隣接するノード集合, $\deg(v)$はノードvの次数を表す. 計算される重要度は, 他の多くの文と類似する文ほど高く, さらに, 重要度の高い文と類似する文の重要度も高くなる.

同様に, エントリ集合に含まれる画像間の類似度を計算し, 隣接行列を用意すれば, 多くのエントリに出現する物体が写った代表画像を得ることができると考えられる. 本研究では, 文および画像をノードとしてグラフを作成し, 文と画像の重要度を同時に計算する. なお, 画像の前後に出現する文は被写体の説明をしている可能性が高いと仮定し, 画像とその前後に出現した文の関係を表す隣接行列の成分は 1 に設定する. これにより, 重要な文に隣接する画像の重要度が高くなり, 文, 画像ともに重要かつ関連性の高いものが選ばれやすくなると考えた. まとめると, 隣接行列の各成分は次のように決定する.

$$a_{i,j} = \begin{cases} 1 & (\mathrm{type}(s_i) = \mathrm{type}(s_j) \text{ and } \mathrm{sim}(s_i, s_j) > \mathrm{threshold}) \\ 1 & (\mathrm{type}(s_i) \neq \mathrm{type}(s_j) \text{ and } |i-j| = 1) \\ 0 & (\text{otherwise}) \end{cases} \qquad (2)$$

ここで, sはエントリを構成する要素(文または画像)のシーケンスを表し, s_iはそのi番目の要素を示す. $\mathrm{type}(s_i)$は要素s_iの型(文または画像), simは類似度関数を表している. なお, simおよびthresholdは要素の型ごとに用意する. 要素間の類似度計算に関しては次節で説明する.

4.2. 文間および画像間の類似度

文は tfidf 値を要素とするベクトルで表し, 類似性尺度にはコサイン類似度を用いる. 画像は 2 種類のベクトルで表現し, それぞれのコサイン類似度を計算し, その平均値を画像間の類似度とする. 利用するベクトル表現は, 色ヒストグラムと Bag of Visual Words ベクトルである.

色ヒストグラムは, HSV 色空間を用いて H, S, V の値域をそれぞれ 10, 4, 4 分割することで 160 色に減色させ, ヒストグラムを計算する. Bag of Visual Words は画像から得られる複数の局所特徴をベクトル量子化してヒストグラムを化したものである[9]. 本研究では, まず, 画像集合から SIFT 特徴量を, スケールを固定して格子状に抽出し, 得られた特徴量を k-means 法によりクラスタリングする. 次に, 個々の画像から抽出した SIFT 特徴量を, クラスタリングにより得たセントロイドを用いてベクトル量子化ヒストグラムを作成する.

4.3. 冗長性の削減

LexRank で計算したスコアの高い順に文を抽出すると, 冗長性のある要約が作成される可能性がある. Radev らは, 文中の情報の包含関係(CSIS)[10]に基づき, 文をリランキングすることでこれを解決している. 本研究では, 文を重要度順に選択する際, 要素間の類似度に閾値を設定しておき, 類似度が閾値以上の要素がすでに選ばれているときは対象要素を要約に追加しないという処理を最後に行う.

5. 評価実験

5.1. 実験設定

Nanba らの手法で収集し, 藤井らの分類手法でエントリにタイプを自動付与したエントリを用いる[1, 2]. 日本全国 22 地点に関して, 地点ごとに平均 10 件ずつエントリを選択し, タイプごとに画像 3 枚, 5 文程度の正解要約を合計 48 件人手で作成した. 同様に, 提案手法を用いて 48 件の要約を自動作成し, ROUGE [11]による評価と, 被験者 2 人による主観評価を行い, 提案手法の有効性を検証した. 主観評価では, 原文と正解要約を提示したうえで, 3 つの手法による要約を良い順に並べるよう被験者に指示した. この評価方法は NTCIR-3 のテキスト要約タクスでも用いられている[12]. なお, 要約がほぼ同じ場合は同一順位にすることも許可している. 比較手法は以下の通りである. いずれの手法でも, 抽出する文と画像の数は正解と同じになるように調整した.

- LexRank+IMG : 4 節で述べた, 文と画像を一つのグラフで扱い LexRank を適用する手法.
- LexRank : 文と画像それぞれに LexRank を適用する手法.
- Lead 法(baseline) : 各エントリの先頭から順に文と画像を抜き出す手法.

なお, 各種パラメータは次のように設定した. エントリを階層的クラスタリングする際のクラスタの距離の閾値を 0.9, Bag of Visual Words ベクトルを計算する際は特徴量のスケールを 16 ピクセルに固定して 8 ピクセルごとにサンプリング行い, k-means 法はクラスタ数を 1,000 に設定した. LexRank における隣接行列作成の際の類似度の閾値は文の場合 0.1, 画像の場合 0.5 と設定した. また, 冗長性を削減するためのコサイン類似度の閾値は文の場合 0.5, 画像の場合 0.9 と設定した.

5.2. 実験結果

表 1 に ROUGE の平均値を示す. 有意水準 5%で両側検定の t 検定を行ったところ, LexRank+IMG と LexRank の ROUGE-1,2 の平均値に有意差は見られなかったが, LexRank+IMG と Lead 法, LexRank と Lead 法

では有意差が確認できた．表 2 に主観評価による順位付けの結果を示す．同様に，有意水準 5%で t 検定を行ったところ，LexRank+IMG と LexRank の平均順位に有意差は見られなかったが，LexRank+IMG と Lead 法，LexRank と Lead 法では有意差が確認できた．まとめると，LexRank の有用性は確認することができたが，文と画像を一つのグラフで扱っても，要約に大きな変化は生じなかった．

また，実際に閲覧システム（図 1, 2）を使用して気がついた提案手法の問題点について述べる．まず，一つのエントリに複数のトピックが混在する場合に，タイプに沿った要約が作成出来ないことがある．たとえば，宮島では牡蠣祭りが毎年行われているが，参加者の多くは，宮島に行くためにフェリーに乗ることも記述している．そのため，宮島周辺でタイプ“食べる”のブログエントリを要約すると，船に関する文や画像が出現する．この問題に関しては，他のタイプが付与されたエントリを参照し，情報利得などからタイプ特有の単語に重みを設定し，文の重要度に反映させることで解決が可能であると考えている．

旅行ブログエントリには観光情報として有用な情報以外にも，個人的な事柄に関する記述も多く含まれるため，それらを要約から除外することも重要な課題である．エントリ数が多いクラスタに関しては，内容の共通部分を重要箇所として検出できるが，エントリが少ないクラスタに関してはその検出が難しくなるため，観光には関係のない情報が要約に含まれる可能性も高くなる．

6. おわりに

本研究では，旅行ブログエントリ閲覧システムを構築し，ユーザが指定した地理的範囲のエントリを要約して提示する機能を実装した．グラフベースの，文の重要度計算手法である LexRank を文と画像に適用することで，画像付きの要約を作成する手法を提案した．今後の課題としては，より有用な観光情報を抽出するために，旅行ブログエントリのコンテンツを詳細に分析する必要があると考えている．

表 1 ROUGE による評価結果（平均値）

	ROUGE-1	ROUGE-2
LexRank+IMG	0.310	0.192
LexRank	0.315	0.194
Lead 法(baseline)	0.253	0.155

表 2 主観評価の結果

	順位の平均値
LexRank+IMG	1.625
LexRank	1.646
Lead 法(baseline)	2.573

文　献

[1] H. Nanba, H. Taguma, T. Ozaki, D. Kobayashi, A. Ishino, and T. Takezawa, “Automatic Compilation of Travel Information from Automatically Identified Travel Blogs,” Proc. of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, Short Paper, pp. 205-208, 2009.

[2] 藤井一輝，石野亜耶，藤原泰士，前田剛，難波英嗣，竹澤寿幸，“多言語旅行ブログエントリを用いた観光情報提示システム，”第 6 回データ工学と情報マネジメントに関するフォーラム，2014.

[3] X. Wu, J. Li, and S.-Y. Neo, “Personalized Multimedia Web Summarizer for Tourist,” Proc. of World Wide Web Conference, 2008.

[4] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang, "Equip Tourists with Knowledge Mined from Travelogues," Proc. of World Wide Web Conference, 2010.

[5] 安田宜仁，西野正彬，片岡良治，“地理範囲とトピックに応じた動的要約生成，” 第 26 回人工知能学会全国大会，2012.

[6] G. Erkan and D. R. Radev, “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization,” Journal of Artificial Intelligence Research, pp. 457-479, 2004.

[7] 前田剛，難波英嗣，竹澤寿幸，“場所に焦点を当てた複数旅行ブログの自動要約，”第 7 回データ工学と情報マネジメントに関するフォーラム，2015.

[8] S. Brin and L. Page, “The Anatony of a Large-scale Hypertextual Web Search Engine,” Computer Networks and ISDN Systems, pp. 107-117, 1998.

[9] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual Categorization with Bags of Keypoints,” Proc. of ECCV International Workshop on Statistical Learning in Computer Vision, pp. 1-22, 2004.

[10] D. R. Radev, H. Jing, and M. Budzikowska, “Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies,” Proc. of the NAACL-ANLP Workshop on Automatic summarization, Vol. 4, pp. 21-30, 2000.

[11] C. Y. Lin, “ROUGE: a Package for Automatic Evaluation of Summaries,” Proc. of Workshop on Text Summarization Branches Out, pp. 74-81, 2004.

[12] T. Fukushima, M. Okumura, and H. Nanba, “Text Summarization Challenge 2 / Text Summarization Evaluation at NTCIR Workshop 3,” in Working Notes of the 3rd NTCIR Workshop Meeting, PART V, pp. 1-7, 2002.