Classification and Visualization of Travel Blog Entries Based on Types of Tourism

Naoki Shibata[1],
Hiroto Shinoda[1],
Hidetsugu Nanba[2],
Aya Ishino[3],
Toshiyuki Takezawa[1]

[1] Graduate School of Information Sciences, Hiroshima City University, Hiroshima, Japan
[2] Faculty of Science and Engineering, Chuo University, Tokyo, Japan
[3] Faculty of Media Business, Hiroshima University of Economics, Hiroshima, Japan

[1]{shibata, shinoda, takezawa}@ls.info.hiroshima-cu.ac.jp
[2]nanba@kc.chuo-u.ac.jp
[3]ay-ishino@hue.ac.jp

Abstract. We propose a method for classifying travel blog entries into one or more tourism types among six predetermined types by using textual and image information in each entry. Together with this information, we use Wikipedia entries, which are automatically linked from each travel blog entry by entity-linking technology, because information beneficial for classifying blog entries is often mentioned in Wikipedia entries, and we combine this information by using a deep-learning-based method. We conducted an experiment with a neural network using three types of input data. Using the Sparse Composite Document Vector (SCDV) technique, we obtained precision, recall, and F-measure scores of 0.743, 0.217, and 0.336, respectively. We also conducted ensemble learning by using SCDV and support vector machines (SVM), and obtained precision, recall, and F-measure scores of 0.807, 0.179, and 0.293, respectively. Finally, we constructed a system that enables travelers to look for travel blog entries from a map in terms of tourism type.

Keywords: Types of Tourism, Travel Blog, Document Classification, Wikification.

1 Introduction

In recent years, tourism has expanded into various types. For example, tourism for the purpose of health recovery is called “health tourism,” and that for the purpose of experiencing sports is called “sports tourism.” If the types of tourism could be automatically classified for travel blog entries, it would enable people to determine what types are available at tourist spots around the world. It would also be possible to recommend tourist sites and travel plans on the basis of tourism types. In this study, we define six

types of tourism and propose a method for classifying a large amount of travel blog entries into these types automatically by using machine learning that considers multiple input data.

At present, many people around the world use social networking services (SNSs). If a user has information to share, they will post it to SNSs. The tourism industry is no exception. We can help the development of the tourism industry by extracting useful information from such enormous data on SNSs. In particular, many travel blog entries have detailed information such as experiences and photos taken at a tourist site. In this study, we analyze travel blog entries.

The remainder of this paper is organized as follows. In Section 2, we discuss related work. Section 3 describes our method. To investigate the effectiveness of our method, we conducted experiments, whose results are reported in Section 4. Section 5 shows the behavior of a system we developed in terms of snapshots. We present our conclusion in Section 6.

2 Related Work

In this study, we use Wikification [1, 2] methods in addition to textual and image information. Wikification is the linking of text and Wikipedia entities. We classify travel blog entries automatically on the basis of tourism types by using this information. Furthermore, we map the classification results.

To enable the distributed representation of documents, Iyyer et al. [3] proposed a model called a "deep averaging network" (DAN) that converts words contained in a document into vectors and uses the average of them for classification. Furthermore, Mekala et al. [4] proposed the Sparse Composite Document Vector (SCDV) technique as an alternative method of generating document vectors. SCDV takes word vectors into consideration with Gaussian mixture modelling (GMM) and inverse document frequency (IDF) and uses the average of the generated word vectors as a document vector. A schematic diagram of SCDV is shown in Figure 1. In this figure, classification is performed for each cluster (a–e) by using GMM, and a document vector is generated by averaging the result and the word vector in consideration of the IDF. In this study, we use this SCDV in the proposed method and a baseline method in an experiment on classifying travel blog entries on the basis of tourism types.

There have been a number of studies related to document classification regarding tourism [5, 6]. Takahashi et al. [5] used travel tweets from Twitter and proposed a method of classifying a traveler's behaviors, i.e., what the traveler is doing, into "sightseeing," "business," "eating," and "shopping." In addition to that, Fujii et al. proposed a method of classifying a traveler's behaviors, i.e. what the traveler is doing, into "buy," "eat," "experience," "stay," and "see" from travel blog entries written in English [6] and Japanese [7]. In their study, although there is some relevance to the classification focusing

on types of tourism in our study, it is basically considered to be another viewpoint. Combining Fujii et al.'s classification with our proposed one based on tourism types could potentially enable more detailed searches, such as examining information related to "eat" content with "cultural tourism."

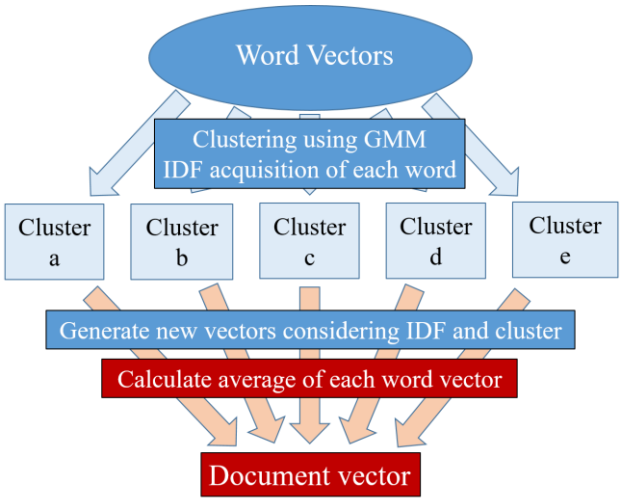

Fig. 1. Schematic of SCDV (Mekala et al. [4]) (compiled by author)

To enable travel information recommendations, Xiong et al. [8] constructed a personalized online hotel marketing recommendation system by extracting hotel characteristic factors and analyzing customers' browsing and purchasing behaviors. Also, Iinuma et al. [9] proposed a method for generating a summary of multiple travel blog entries that contain images and constructed a system. The system classifies the travel blog entries, which were collected by Nanbas' method (Nanba et al. [10]), by using Fujii et al.'s method.

3 Classification of Travel Blog Entries on the Basis of Tourism Types

3.1 Definition of Tourism Types

At present, there are many types of tourism, and most of them have no strict definitions. We selected and defined six tourism types on the basis of the ease of automatically clarifying them. Table 1 shows the different types, their definitions, and examples. We automatically classify travel blog entries written in English to clarify the tourism types of travelers on the basis of the six types.

Table 1. Definition and example of the types of tourism (own material)

types of tourism	definition	examples
infrastructure and hard tourism	Tourism for modern buildings and recreational facilities.	bridges, dams, theme parks, shopping malls, aquariums
health tourism	Tourism aimed at health recovery, health maintenance, and health improvement.	religious pilgrimages, hot springs, hiking, trekking
sports tourism	Tourism aimed at experiencing or watching sports.	MLB, Soccer World Cup, Olympics
green tourism	Tourism aimed at interacting with nature.	agricultural experience, fruit hunting, picnics
heritage tourism	Tourism for historic buildings such as world heritage sites.	World Heritage sites, national treasures, castles
cultural tourism	Tourism for life, culture, ethnicity, and tradition of areas.	festivals, interchange with local people

3.2 Classification of Travel Blog Entries Based on Types of Tourism

In this study, we classify travel blog entries automatically on the basis of the types of tourism shown in Section 3.1. In this section, we explain the policy of automatic classification and the automatic classification of travel blog entries with machine learning.

3.2.1 Automatic Classification Policy

We analyze text, images, and Wikification results from travel blog entries and automatically classify them into tourism types by using the results of the analysis. Among them, textual information is the most important. For example, if a blog entry contained the phrase "I went to Mont Saint Michel, a UNESCO World Heritage Site," this is considered to be an example of "heritage tourism" because the blog text contains the expression "UNESCO World Heritage Site."

A disadvantage of text analysis is that misinterpreted context could produce inaccurate results. For example, if a blog entry contained the phrase "I wanted to ski but I could not do it," although the word "ski" is present, the entry would not be related to "sports tourism." However, if an image related to skiing is in the blog entry, it can be assumed that the author was skiing. In this study, we used the Google Cloud Vision API[1] for detecting objects in images. The API can classify images into thousands of categories, detect objects/faces, etc. We used words that obtained the object detection results in addition to textual information when classifying blog entries.

Classification requires external knowledge of what can be read from text or images. For example, if a blog entry contained the phrase "I saw Pyramid in Egypt. It was very big,"

[1] https://cloud.google.com/vision/?hl=en

this blog entry is classified as “heritage tourism” because the Pyramid is a World Heritage Site. However, there is no information in the text. To judge this correctly as “heritage tourism,” the inclusion of external knowledge such as information from Wikipedia is required. For this, we used the Google Cloud Natural Language API[2]. When text is sent to the cloud via the API, in addition to part-of-speech tagging, parsing, and lexical expression extraction, Wikification is also performed. By using the results of Wikification, we can obtain the information that the Pyramid is a World Heritage Site from the linked page of Wikipedia. Accordingly, in this study, we aim to achieve more accurate classification by using Wikification information as external knowledge in addition to the information from text and images.

3.2.2 Automatic Classification of Travel Blog Entries Using Machine Learning

In this study, we first prepare input data from the text, images, and Wikification results of a travel blog entry. Objects are detected in images, and words included in the linked Wikipedia abstract (the first paragraph) from the Wikification results are extracted. After that, we construct classifiers in consideration of each piece of input data. Since the classifiers are binary classifications, blog entries that do not fall into the six tourism types are classified as “other.” A schematic of the classifier is shown in Figure 2, where each piece of input data is processed, and the results are combined in a hidden layer.

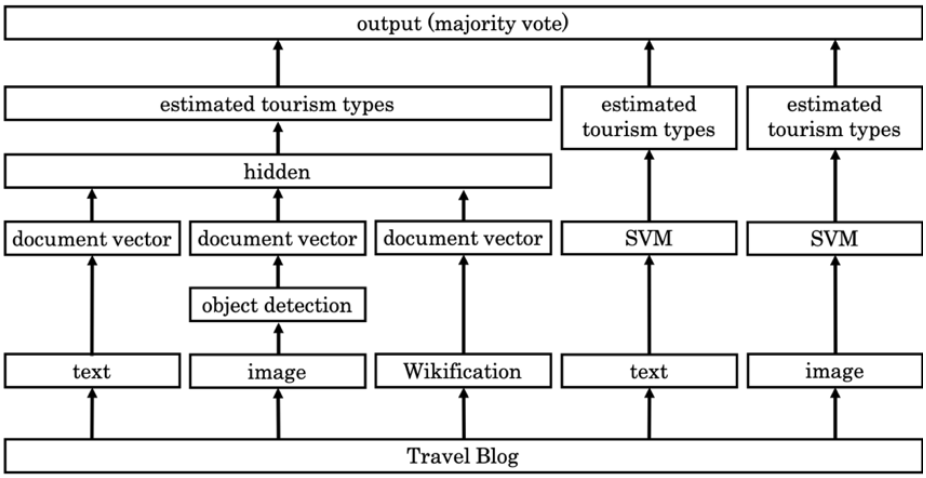

Fig. 2. Schematic of proposed classifier (authors’ own figure)

This classifier integrates the analysis results of each piece of input data in the hidden layer. This is why the number of words included in each is significantly different. For example, in a travel blog, when the text includes 1,000 words, the object detection result includes 30 words, and the abstract of Wikipedia includes 100 words, so the number of

[2] https://cloud.google.com/natural-language/?hl=en

words differs; thus, the influence of input data with a small number of words is reduced. For this reason, analysis results for each piece of input data are integrated in the hidden layer in this study.

4 Experiments

We conducted an experiment by using TravelBlog[3], one of the largest travel blog websites. It hosted over 700,000 blog entries in 2013. The aim of this experiment was to achieve a high precision accuracy as the number of blog entries is very high.

4.1 Experimental Conditions

Each blog entry was classified into one or more tourism types among six predetermined types manually and used as training data and test data for machine learning. A breakdown of the results classified manually is shown in Table 2. First, we originally defined nine tourism types, i.e. dark tourism, contents tourism, study tourism, and the six tourism types listed in Table 1. Second, we manually classified 2,017 randomly-selected travel blog entries. This data was created by a student, whose major is International Studies and is good at speaking English. Third, due to a lack of blog entries related to dark, contents, and study tourism (less than 30), we discarded those types because it was insufficient for machine learning. Finally, we obtained 2,017 travel blog entries and categorized them into the remaining tourism types as shown in Table 2. Also, it should be noted that 227 entries were classified with multiple types of tourism and 1,227 could not be classified at all.

Table 2. Breakdown of results classified manually (own material)

types of tourism	number
infrastructure and hard tourism	168
health tourism	125
sports tourism	57
green tourism	453
heritage tourism	198
cultural tourism	49
targeted for classification	2,017

For the distributed expression of words, we used a pre-trained model provided by Google, the Word2Vec model of 300-dimension vectors[4]. This well-known model learned from the Google News dataset about 100 billion words, which is much larger than the TravelBlog corpus. The object detection function used the Google Cloud Vision API described in Section 3.2.1. We conducted classification by taking into account the results of Wikification, which, as stated above, is a method of obtaining a distributed

[3] https://www.travelblog.org/

[4] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

expression from a Wikipedia abstract. Also, we experimented with ensemble voting methods with the same weights. For the ensemble (proposed) method, it used three classifiers: SCDV(txt+img+wiki) (proposed), SVM(txt), and SVM(img). For the ensemble (baseline) method, we used three classifiers: SCDV(txt), SVM(txt), and SVM(img). We applied the radial basis function (RBF) kernel to both SVM(txt) and SVM(img). We obtained the best epoch values using the optimization function "RMSpropGraves." For the activation function, we adopted ReLU in the hidden layer and softmax in the output layer.

The evaluation was performed by 5-fold cross validation to decrease problems like overfitting or selection bias, and precision, recall, and F-measure scores were used. The cross-validation process was repeated five times, with each of the five subsamples (the number of data samples are 403, 403, 403, 403, and 405) being used as the test data with the remaining four as training data. The five results can then be averaged to produce a single estimation. To calculate these, we used a micro average to take into account the bias in the number of data for each type of tourism. The precision and recall formulas are shown below.

$$\boldsymbol{precision} = \frac{\boldsymbol{Travel\ blog\ entries\ classified\ correctly}}{\boldsymbol{Travel\ blog\ entries\ classified\ as\ tourism\ types}}$$

$$\boldsymbol{recall} = \frac{\boldsymbol{Travel\ blog\ entries\ classified\ correctly}}{\boldsymbol{Travel\ blog\ entries\ labelled\ manually\ as\ tourism\ types}}$$

Generally, a trade-off between precision and recall is necessary. In our study, high precision is more important even if recall is low, and because more than 230,000 travel blog entries are available, this will resolve the low recall. In this experiment, we used the following three proposed methods (Table 3) and six baseline methods (Table 4). We conducted t-tests ($p < 0.01$), which confirmed that there were significant differences between SCDV(txt) and SCDV(txt+img) (proposed), and between SCDV(txt+img) and SCDV(txt+img+wiki).

Table 3. Proposed methods and features for use (own material)

	text	image	Wikification
Ensemble (proposed) ● SCDV(txt+img+wiki) ● SVM(txt) ● SVM(img)	○	○	○
SCDV(txt+img+wiki)	○	○	○
SCDV(txt+img)	○	○	−

Table 4. Baseline methods and features for use (own material)

	text	image	Wikification
Ensemble (baseline) ● SCDV(txt) ● SVM(txt) ● SVM(img)	○	○	−
SCDV(txt)	○	−	−
SVM(txt)	○	−	−
SCDV(img)	−	○	−
SVM(img)	−	○	−
SCDV(wiki)	−	−	○

4.2 Results and Discussion

Table 5 shows the experimental results of each baseline and the proposed methods described in Section 4.1. The highest precision of 0.807 was obtained with Ensemble (proposed). Also, the highest recall and F-measure scores of 0.272 and 0.385 were obtained with SVM(txt).

Compared with SVM(img), which had the highest precision among the baseline methods, Ensemble (proposed) produced better results for precision, recall, and F-measure, indicating that it is more effective. In terms of different input data, the proposed SCDV(txt+img) obtained a higher precision than the baseline SCDV(txt) and SCDV(img). Furthermore, the proposed SCDV(txt+img+wiki) obtained a higher precision than the proposed SCDV(txt+img), and baseline SCDV(txt), SCDV(img), and SCDV(wiki).

Table 6 shows the number of blog entries changed from ensemble (baseline) method to ensemble (proposed) method. The reason only two types are shown in this table is that there was no difference between outputs of the two methods. Focusing on green tourism, the number of misclassifications decreased, while the number of correctly classified blog entries also decreased. It appears that textual information is more important than image and Wikipedia information when classifying travel blog entries as green tourism. On the other hand, the number of correctly classified blog entries increased in heritage tourism. In this case, image and Wikipedia information are useful, and they contribute to improve the recall value. Thus, increasing the number of inputs is valid when classifying travel blog entries.

Table 5. Experimental results (micro average) (own material)

method	epoch	precision	recall	F-measure
Ensemble (proposed) ● SCDV(txt+img+wiki) ● SVM(txt) ● SVM(img)	-	**0.807**	0.179	0.293
SCDV(txt+img+wiki) (proposed)	30	0.752	0.218	0.338
SCDV(txt+img) (proposed)	30	0.729	0.227	0.347
Ensemble (baseline) ● SCDV(txt) ● SVM(txt) ● SVM(img)	-	0.747	0.216	0.335
SCDV(txt) (baseline)	20	0.639	0.169	0.268
SVM(txt) (baseline)	-	0.654	**0.272**	**0.385**
SCDV(img) (baseline)	10	0.725	0.140	0.235
SVM(img) (baseline)	-	0.788	0.170	0.279
SCDV(wiki) (baseline)	30	0.528	0.116	0.191

Table 6. The number of blog entries changed from ensemble (baseline) method to ensemble (proposed) method (own material)

	green.	heritage.
correctly classified into “a tourism type”	25	9
incorrectly classified into “a tourism type”	15	1
correctly classified into “not a tourism type”	47	1
incorrectly classified into “not a tourism type”	69	4
total number of blog entries (as show in Table 2)	453	198

5 System Behavior

In this section, we introduce our system’s behavior in terms of the travel blog entries collected and classified by our proposed method. The system intuitively reveals the features of the tourism types for each tourist site. The procedure for visualization is as follows.

(1) Collect travel blog entries, and extract text and images from entries.
(2) Analyze images by using the Google Cloud Vision API, and estimate object detection and location information.
(3) Perform Wikification on text by using the Google Cloud Natural Language API.
(4) Collect Wikipedia entity information with the results obtained by Wikification and extract abstracts of linked Wikipedia articles.

(5) Classify on the basis of tourism types automatically using the obtained text, image analysis results, and abstracts of Wikipedia.
(6) Visualize data on a Google Earth map by using the location information obtained by the image analysis results. If multiple location information references can be extracted, the first one extracted is adopted.

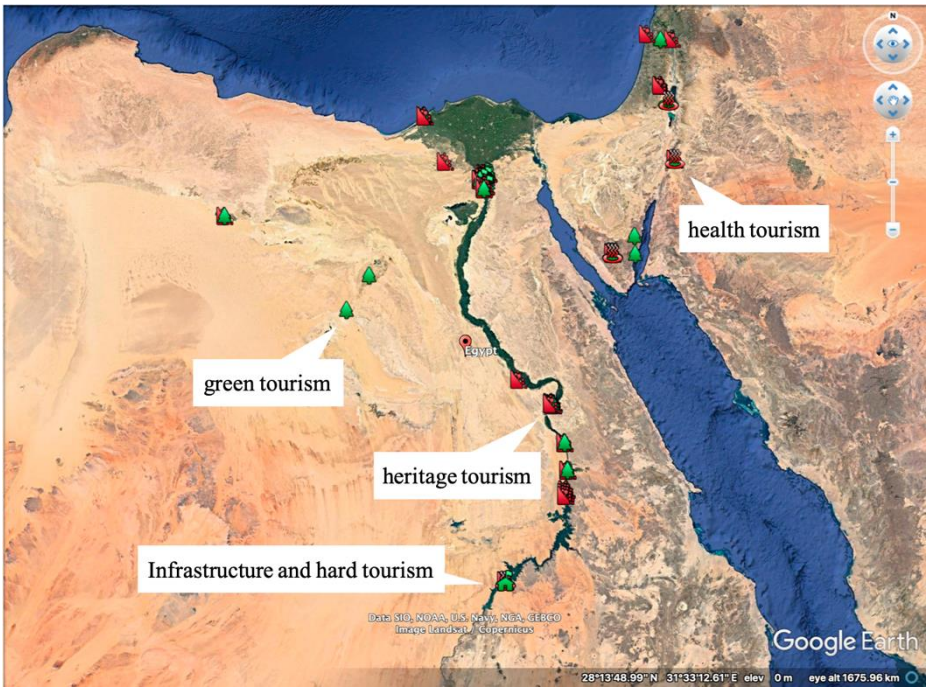

Fig. 3. Travel blog entries classified on the basis of tourism types (Egypt) (compiled by author)

We collected about 230,000 random travel blog entries from TravelBlog, and used 24,023 entries whose location information could be estimated for classification. The system is shown in Figure 3. This figure illustrates Egypt and its surrounding. We used icons to indicate type, such as a green house for "infrastructure and hard tourism," a hot spring for "health tourism," a bike for "sports tourism," a tree for "green tourism," the rocks for "heritage tourism," and a temple for "cultural tourism." If the user clicks an icon, the corresponding travel blog entry is shown on the map. From this figure, we can confirm many icons of the rocks, "heritage tourism" around the Nile River. In addition, the users of this system can find other tourism types information, such as "green tourism," "health tourism," and "infrastructure and hard tourism." Thus, this system enables to look up the information about tourism types that are useful for the users.

6 Conclusion

In this study, we proposed a method for automatically classifying travel blog entries into one or more tourism types among six predetermined types in consideration of text and images found in them and Wikipedia information. For images, we used the Google Cloud Vision API to detect objects and adopted the results as classification features. For Wikipedia information, we performed Wikification by using the Google Cloud Natural Language API, and we used the word sets included in the abstracts of linked Wikipedia articles as classification features. The experimental results show that a precision score of 0.807 was obtained for ensemble learning, which combined SCDV(txt+img+wiki), SVM(txt) and SVM(img).

For the visualization system, we classified 24,023 travel blog entries and visualized travel blob data on a map by using Google Earth. The proposed system enables analysts to investigate traveler behavior (Wenger et al. [11]) and marketing (Mack et al. [12]) via massive numbers of travel blog entries. However, currently, the system assumes travel blog entries written in English as input. Our future work is to expand to blog entries written in other languages.

References

1. Mihalcea, R. & Csomai, A.: Wikify! Linking Documents to Encyclopedic Knowledge. In: The ACM Conference on Information and Knowledge Management, pp. 233–242 (2007).
2. Yugo, M. & Sinsuke, M.: Wikification for Scriptio Continua. In: The 10th Edition of the Language Resources and Evaluation Conference, LREC, pp. 1346–1351 (2016).
3. Iyyer, M., Manjunatha, V., Boyd-Grader, J. & Daume III, H.: Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In: The Association for Computational Linguistics, ACL (2015).
4. Mekala, D., Gupta, V., Paranjape, B. & Karnick, H.: SCDV: Sparse Composite Document Vectors using Soft Clustering over Distributional Representations. In: Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 659–669 (2017).
5. Takahashi, K., Kato, D., Endo, M., Araki, T., Hirota, M. & Ishikawa, H.: Analyzing Travel Behavior using Multi-label Classification from Twitter. In: The 9th International Conference on Management of Digital EcoSystems, MEDES'17 (2017).
6. Fujii, K., Nanba, H., Takezawa, T., Ishino, A., Okumura, M. & Kurata, Y.: Travellers' Behavior Analysis Based on Automatically Identified Attributes from Travel Blog Entries. In: Workshop of Artificial Intelligence for Tourism, PRICAI (2016).
7. Fujii, K., Nanba, H., Takezawa, T. & Ishino, A.: Enriching Travel Guidebooks with Travel Blog Entries and Archives of Answered Questions. In: International eTourism Conference, ENTER 2016 (2016).
8. Xiong Y & Geng L.: Personalized Intelligent Hotel Recommendation System for Online Reservation--a perspective of product and user characteristics. In: Management and Service Science, MASS, pp. 1–5 (2010).
9. Iinuma, S., Nanba, H., & Takezawa, T.: Investigating the Effectiveness of Computer-produced Summaries Obtained from Multiple Travel Blog Entries. Information Technology & Tourism, Vol.21, No.1, pp. 83-103 (2019).

10. Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A., & Takezawa, T.: Automatic Compilation of Travel Information from Automatically Identified Travel Blogs. In: The 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).
11. Wenger, A.: Analysis of Travel Bloggers' Characteristics and their Communication about Austria as a Tourism Destination. In: Journal of Vacation Marketing, 14(2). pp. 169-176 (2008).
12. Mack, R. W., Blose, J. E. & Pan, B.: Believe it or not: Credibility of Blogs in Tourism. In: Journal of Vacation Marketing, 14(2). pp. 133-144 (2008).