Social media knows what road it is: quantifying road characteristics with geo-tagged posts

Takuya Nishimura¹ · Kyosuke Nishida¹ · Hiroyuki Toda¹ · Hiroshi Sawada¹

Received: 26 December 2016 / Revised: 6 September 2017 / Accepted: 7 October 2017 / Published online: 9 November 2017
© The Author(s) 2017. This article is an open access publication

Abstract Determining traveling routes that provide opportunities to satisfy the various requirements of users in urban areas is still an open problem. This is because it is virtually impossible to manually quantify the characteristics of each street or road and there are few web-based or semantic resources of subjective requirements that describe streets and roads directly. Thus, it is difficult to satisfy all the needs and desires of users that may arise, such as for finding boulevards that are “fashionable” or “beautiful.” The goal of this study is to automatically quantify the characteristics of streets and roads in relation to requirements that can be described using keywords such as “fashionable.” To achieve this goal, we propose a two-stage method that analyzes social media and road networks. First, in estimating the topic distribution (i.e., the characteristics) of each point of interest (POI), our method uses the latent Dirichlet allocation model to analyze geo-tagged texts while considering which users posted useful information for estimating road characteristics. Next, it uses a Markov random field model to estimate the characteristics of each street or road on the basis of those of the POIs and the road networks associated with the POIs. Experiments on real datasets demonstrate that our method achieves statistically significant improvements over baseline methods in terms of ranking quality in information retrieval for 300 roads in three urban areas in Tokyo when given 25 keywords.

Keywords GIS · LDA · MRF · Social Media · Road Network · Route Navigation

1 Introduction

Lynch (1960) noted that paths are often the most important elements in people’s image of their city. Their images of paths consist of different points of view as regards geographical characteristics. Recent route navigation services provide path-finding functions that consider just a few basic characteristics, such as ease of walking, in addition to path length. For example, users can find “No traffic jam routes” by using Google Map Navigation.¹ However, user desires with regard to traveling routes are much more varied, and the existing services fail to consider most of them. The results of our private survey on user requirements show that there are about 900 distinct types (road characteristics) associated with traveling routes. For example, some users want to walk along fashionable streets or roads that offer currently popular styles of clothing. Other users want quiet and beautiful paths, while others want to walk through streets with lively atmospheres.

Our ultimate goal is to identify routes that provide opportunities to satisfy various user requirements. It is difficult to make a route navigation service that can satisfy each such requirement one by one because the number of possible requirements is so large. Our solution is to build a system that can find routes having characteristics indicated by natural language keywords, such as “fashionable.” Each keyword is assumed to represent a user’s requirements, and any of several input methods may be used, such as direct input via a search box and automatic extraction of words

✉ Takuya Nishimura
nishimura.takuya@lab.ntt.co.jp

¹ NTT Service Evolution Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa 239-0847, Japan

¹ <https://www.google.co.jp/maps/>.

Fig. 1 We quantify *any* characteristic of *any* road. For example, when we input “fashionable” as the query keyword, a score indicating how fashionable the road is will be calculated for each road, shown here as the brightness of the green tint (color figure online)

spoken during interactions with a dialog system (Jameson 2009). For existing multi-criteria path-finding algorithms (e.g., Clímaco and Pascoal 2012) to take into account a length restriction that the user sets on the traveling route and the characteristics of their requirement, we need to compute a score of each road for a given keyword.

The originality of this study lies in quantifying all the characteristics of all roads, which permits calculation of scores for any combination of a road and a keyword (Fig. 1). We set ourselves the following requirements.

1. We need to collect information from online resources and at the same time avoid having to make expensive manual annotations or surveys, as they would make the navigation service impractical.
2. We need to score all roads for any requirement in order to make a truly comprehensive route navigation system.

An attractive approach to understanding geographical characteristics, and thus satisfying Requirement 1, is to analyze geo-tagged texts posted on social media for which users post various information, including characteristics difficult to quantify, such as “fashionable.” Many successful studies have used data of geo-tagged social media such as Twitter and Flickr to understand the characteristics of geographical areas (Yin et al. 2011; Hong et al. 2012) or points of interest (POIs) (Zhuang et al. 2015). A POI is a specific location that someone thinks is useful or interesting, such as a restaurant, sightseeing location, or shop.

However, satisfying Requirement 2 is a new challenge. The difficulty of quantifying road characteristics is that most roads have never been tagged with expressive quantitative data. Recently, several characteristics have been quantified by activities such as exploration for urban planning by governments (e.g., UK provides road safety data).² Some recent studies have tried to quantify road

characteristics from statistical data such as statistical data or sensor data (Galbrun et al. 2014; Yang et al. 2014), human annotations (Quercia et al. 2014), and social media data (Quercia et al. 2015; Skoumas et al. 2014), but these approaches heavily focus on particular characteristics. Unfortunately, extending these approaches (including the ones using social media data) so that they can meet a wide variety of user requirements requires collection of information which comes at a high cost in terms of human effort and time.

Our key idea is based on the assumption that road characteristics tend to mirror neighboring POIs. For example, SoHo³ is an area with many historical and artistic POIs, and several streets in SoHo are considered to be historical or artistic. Lynch (1960) describes relationships between elements of cities including streets and POIs. Quercia et al. (2014) show that the appearance of buildings on a street affects a viewer’s impression of it. We have confirmed that there is relationship between the characteristics of roads and those of POIs by analyzing the distribution of POI categories neighboring characteristics roads (See Sect. 5). On the basis of this finding, we estimate the road characteristics by utilizing the characteristics of the POIs mentioned by a sufficient number of people in social media. Moreover, we consider the behavior of posting users when estimating the characteristics of streets, especially roads having no neighboring POIs.

We propose a new two-stage method for quantifying road characteristics from geo-tagged social media text. The first stage quantifies the POI characteristics by applying latent Dirichlet allocation (LDA) (Blei et al. 2003) to texts posted by social media users. Our previous study Nishimura et al. (2016) treated all posts as having equal importance at this stage. However, we still should determine whether the posted text does actually refer to an experience related to geographical characteristics or not because many texts are posted at the users’ house and describe daily activities (Hu et al. 2014). Thus, our method improves estimation quality by increasing the weight assigned to users who frequently post their impressions/experiences of various places. The second stage estimates road characteristics from those of the POIs and the road network associated with the POIs by using the Markov random field method (MRF) (Besag 1986). After these two stages, our method calculates a relevance score for any road/keyword combination by using the word distributions and topic distributions estimated automatically from geo-tagged social media texts; thus, it satisfies Requirements 1 and 2.

The contributions of this study and our original conference paper (Nishimura et al. 2016) are as follows.

² <https://data.gov.uk/dataset/road-accidents-safety-data>.

³ https://en.wikipedia.org/wiki/SoHo,_Manhattan.

- We conduct a survey of user requirements for traveling routes and show there are various requirements related to subjective or qualitative information not addressed by existing studies. This shows the importance of attempting to identify routes that can satisfy user requirements. To extend the scope of our previous study Nishimura et al. (2016), we classify requirements into five categories and analyze each category.
- We propose a two-stage method that quantifies POI characteristics by using LDA and estimate road characteristics by using MRF. This method can estimate road characteristics accurately without requiring direct descriptions of the roads. To extend our previous study Nishimura et al. (2016), we emphasize (weight) words posted by users who post from various POIs because posts by such users may include useful information for obtaining geographical characteristics.
- An experiment conducted with real social media data and road network data obtained from OpenStreetMap shows that our framework can quantify road characteristics. The method improved the nDCG score by 0.5 and precision@5 by 0.2 over the naive baseline (BM25) for 25 keywords and 300 roads. The results indicate that we can quantify road characteristics from geo-tagged social media texts, even though they do not describe the roads directly.
- To extend the scope of our previous study Nishimura et al. (2016), we show there is a relationship between the characteristics of roads and those of POIs by analyzing the distribution of POI categories neighboring characteristics roads.
- To extend the scope of our previous study Nishimura et al. (2016), we clarify which areas and keywords are difficult for our method to estimate road characteristics and describe why our method does not work well for these cases.

The rest of the paper is as follows. After discussing related work in Sect. 2, we present the results of a survey of user requirements for traveling route selection in Sect. 3. We introduce our method for quantifying road characteristics in Sect. 4 and present an experiment that evaluated our method in Sect. 5. Before concluding the paper, we discuss the limitations of our method and outline future work in Sect. 6.

2 Related work

Early studies on understanding geographical characteristics from documents fall into the category of geographical information retrieval (Purves and Jones 2011; Andrade and Silva 2006; Toda et al. 2009). Geographical information

retrieval uses location names in a document to assign it corresponding geographical focuses. However, this kind of method mainly treats location names that have geographical extent, such as San Francisco and Palo Alto; geographical granularities are not fine enough to quantify road characteristics.

One interesting development is the current popularity of geo-tags, which allow documents to be associated with specific locations. Several recent studies have tried to understand geographical characteristics from geo-tagged texts such as social media posts by applying topic model techniques, such as LDA (Blei et al. 2003) and probabilistic latent semantic analysis (pLSA) (Hofmann 1999), or signal processing techniques, such as principle component analysis (PCA) (Jolliffe 2002). Several studies have proposed extended topic models with geo-tagged texts for understanding geographical characteristics (Yin et al. 2011; Hong et al. 2012; Mei et al. 2008) with the assumption that documents geographically close to each other have similar topic distributions. Some studies apply PCA to geo-tagged texts to discover characteristics of geographical areas (Sengstock and Gertz 2012).

Moreover, other studies estimate geographical characteristics from information other than texts in social media. Le Falher et al. (2015) and Çelikten et al. (2016) capture urban places that have the same meaning in different cities. Anagnostopoulos et al. (2016) find which urban area is related to which characteristic on the basis of geo-tagged twitter posts and the interests of the poster. However, these studies do not focus on roads and their methods cannot be adapted to roads.

On the other hand, several studies have tried to understand road characteristics. Quercia et al. (2014, 2014) tried to quantify road characteristics, e.g., happy, quiet, or beautiful routes, from various resources such as crowdsourcing and/or geo-tagged images, scent information from geo-tagged social media (Quercia et al. 2015), and road network data indicating walkability, and opendata (Quercia et al. 2015). Their studies are useful for quantifying specific characteristics of roads. However, the emergence of a new characteristic would incur additional costs such as another round of crowdsourcing or development of a vocabulary list relevant to the new characteristic. Skoumas et al. (2014) proposed a method that adds a popularity score to each road by analyzing blog articles about travel experiences. Corcoran et al. (2015) estimate the road type such as a motorway or residential street from the road network data. Galbrun et al. (2014) proposed a method that adds a safety score to roads by using kernel density estimation on open data about crime statistics. Yang et al. (2014) estimate greenhouse gas emissions on roads from the emission data of a few users. These methods have

restrictions on the characteristics to be quantified and thus utilized.

As mentioned above, none of the existing studies can satisfy the requirements set out in Sect. 1. In what follows, we describe how we quantify road characteristics comprehensively by analyzing geo-tagged social media texts together with road network data at a low cost.

3 Requirements survey

We surveyed user requirements for traveling routes by contracting 1000 workers on a Japanese crowdsourcing platform site and gathering their responses to the following questions.

- Q1–Q3: Age of worker, gender of worker, and most common mode of transportation.
- $Q_{\text{pedestrian}}$: “Please imagine you are using a route navigation service and are moving on foot. If you could input a query keyword to indicate the route characteristic desired, what word would you input? Please provide four keywords.”
- $Q_{\text{bike}}, Q_{\text{car}}$: Same question as $Q_{\text{pedestrian}}$ for bicycle and car.

Table 1 shows the responses to Q1–Q3. We can see that the workers exhibited widely distributed demographics.

We describe below the overall results for $Q_{\text{pedestrian}}-Q_{\text{car}}$. The unique responses numbered 938 for $Q_{\text{pedestrian}}$, 798 for Q_{bike} , and 952 for Q_{car} . Table 2 shows the top ten answers for $Q_{\text{pedestrian}}-Q_{\text{car}}$. We manually grouped answers that had similar meanings into one answer. Most of the responses to questions such as those regarding the presence of a convenience store, numbers of traffic lights, and road width were similar. Figure 2 plots the number of responses for $Q_{\text{pedestrian}}$, Q_{bike} , and Q_{car} . The plot exhibits long tail characteristics. This means it is too time-consuming or too expensive to collect the required information by using any of the existing approaches.

Table 1 Distribution of demographics of respondents

Age	10–19	8	40–49	395
	20–29	113	50–59	124
	30–39	341	60–	19
Gender	Male	542	Female	458
Transport he/she usually uses	Walking	573	Train	68
	Bicycle	376	Bus	27
	Motorcycle	76	Others	84
	Car	585		

We found that requirements can be classified into five categories, and these categories can be grouped into two super categories. Below, we describe the meaning of each category and super category and the results of the analysis. Table 3 shows an example of requirements classified into each category for the case of moving on foot (responses of $Q_{\text{pedestrian}}$).

- *Super category 1: Quantitative requirements* The categories belonging to this super category are quantitative information on each road that does not depend on the user’s impression.

Category 1: Required physical characteristics This category includes requirements such as the width or slope of roads. Information related to this category is quantified and qualified by various organizations. For example, OpenStreetMap contributors qualify roads by type by assigning tags such as “highway.”⁴ Other characteristics are quantified by referring to physical characteristics collected in previous studies such as government surveys.

Category 2: Required statistics This category includes information such as volume of traffic or safety of roads. A number of route navigation services take into account several requirements. For example, Google Map Navigation provides users with “No traffic jam routes.” As an other example, Galbrun et al. (2014) tried to quantify the safety of roads by using open data provided by governments.

Category 3: Required devices and facilities This category includes characteristics such as the presence of traffic lights or convenience stores. Traffic lights and sidewalks are considered to be road network data such as OpenStreetMap. Information on facilities is collected and provided by several organizations, such as the Google Places API.⁵

- *Super category 2: Subjective requirements* This super category contains subjective information about each road that depends on the impressions of individual users.

Category 4: Required experience near roads This category is for routes meeting desires such as “Users can view cherry blossoms” or “Users can enjoy sightseeing.” Current route navigation services cannot provide routes related to these requirements because what event users experienced on the side of each road is not quantified.

Category 5: Required impression: This category includes desires such as “beautiful roads” or “bright roads.” These requirements are not provided by

⁴ <http://wiki.openstreetmap.org/wiki/Way>.

⁵ <https://developers.google.com/places/>.

Table 2 Number of responses about requirements for traveling route

Walking		Bicycle		Car and motorcycle	
Existence of convenience stores	226	Flat road	233	Number of traffic lights is small	261
Flat road	172	Existence of bicycle paths	217	No traffic jams	209
Number of traffic lights is small	155	Traffic volume is small	204	Existence of convenience stores	186
Traffic volume is small	120	Existence of convenience stores	189	Wide road	158
Existence of sidewalk	112	No slopes	172	Existence of gas stations	146
Scenery is beautiful	106	Number of traffic lights is small	132	Scenery is beautiful	102
Safe road	106	Safe road	107	Shortcut path	84
Existence of shops	106	Wide road	106	Traffic volume is small	82
No traffic jam	92	Scenery is beautiful	94	Existence of shops	68
Wide road	91	Existence of shops	72	Safe road	64

Fig. 2 Frequency of responses for $Q_{\text{pedestrian}}$, Q_{bike} , and Q_{car} . The horizontal axis is the popularity rank of each response. The vertical axis is the number of responses**Table 3** Example of requirements and number of responses regarding each requirement in five categories when moving on foot

Required quantitative characteristics						Required subjective characteristics			
Required physical characteristics		Required statistics		Required devices and facilities		Required experiences near roads		Required impression	
Requirement	#	Requirement	#	Requirement	#	Requirement	#	Requirement	#
Flat road	172	Traffic volume is small	120	Existence of convenience stores	226	Existence of rest place	47	Beautiful scenery	106
Wide road	91	Safe road	106	# of traffic lights is small	155	Existence of sightseeing spots	18	Bright road	76
No hill	90	No traffic jam	92	Sidewalk	112	Suitable for strolling	5	Bright at night	48
Shortest route	49	Many pedestrian	47	Existence of shops	106	Suitable for running	3	Beautiful road	12

existing route navigation services either. To provide routes related these requirements, we need to quantify how users feel on the side of each road.

The categorization is ambiguous, so many requirements can be classified to multiple categories and we cannot distinguish demands category automatically. For example, “safe road” can be defined as the probability of the occurrence of crime or traffic accident; however, the safeness that people feel may differ from the statistics. In this study, we classify requirements into quantitative requirements when we can define the score of characteristics quantitatively even if the impression of people may differ.

The above analysis shows that requirements classified to quantitative requirements are provided by existing route navigation services or can be realized with information which is already quantified. Some requirements classified to categories of subjective requirements can be quantified from existing data; for example, “existence of sightseeing spots” can be quantified with POI category data. However, most requirements of subjective requirements cannot be quantified from existing data. Thus, we need a new database about roads for finding routes related to such desires. Such desires cannot be satisfied at a reasonable cost with the current methods described in Sect. 2. For example, Quercia et al. (2014) used crowdsourcing to quantify three

characteristics of roads (beautiful, quiet, happy). However, these three characteristics are present in only 111 of the 4000 responses for $Q_{\text{pedestrian}}$. In order for this method to satisfy other requirements, we need more crowdsourced respondents. This would incur long delays and/or enormous costs; thus, it cannot satisfy requirements set out in Sect. 1.

Fortunately, information about these requirements is often found in social media posts. Users post many texts detailing their experiences on social media (Hu et al. 2014). Quantifying that information into road characteristics such as “suitability of roads for cherry blossom viewing” or “suitability of roads for shopping” will be helpful for users to get the desired experience near roads. Moreover, road characteristics related to desired impression can be quantified from social media text. Impressions of POIs are posted as text information such as “This is fashionable caf”; thus, we can estimate the characteristics of POIs from the posted impressions. With that *POI* information, we can estimate the road characteristics because they tend to mirror those of neighboring POIs (Lynch 1960).

The super category classification is also related to the relationship between the transportation method and the requirements. Requirements classified as categories of super category 1, and quantitative requirements differ depending on the transportation method because the value of these requirements depends on method. For example, 217 respondents answered “cycling roads” to Q_{bike} and 146 respondents answered “road with gas station” to Q_{car} . These answers are classified as requirements with physical characteristics or requirements related to devices and facilities, and they are dependent on the user’s mode of travel. On the other hand, responses classified into super category 2, and subjective requirements were similar among transportation modes. For example, the desire for “Beautiful Scenery” which was classified as a desired impression was the most common reply to $Q_{\text{pedestrian}}-Q_{\text{car}}$.

We aim at a route navigation system for various modes of movement and propose a method that can solve these problems and so create a route navigation system that can satisfy various user requirements, especially those of super category 2 which existing services and resources have difficulty satisfying.

4 Proposed method

The simplest way of calculating the relevancy between a query keyword and each road is keyword-based matching, i.e., counting how many times each descriptive word occurs near the road. However, the capability of keyword-

based matching is poor for the relevancy calculation. This method cannot treat synonyms and context and often fails to estimate the relevancy. Several studies have tried to solve this problem by enriching the vocabulary related to the characteristics (Quercia et al. 2014, 2015), but this method requires a lot of vocabulary information to be collected, and this could be costly. To overcome this difficulty, we propose a topic-based matching method. Recent studies use topic models to find abstract topics from documents automatically (Blei et al. 2003; Hofmann 1999). Our method represents a given keyword and a road as a vector, whose dimensions correspond to topics and calculate the relevancy between a keyword and each road by using the similarity between vectors.

However, obtaining a vector for each road is difficult because most geo-tagged social media texts do not describe roads directly. We solve this problem by utilizing the characteristics of POIs. POIs are individually mentioned in sufficient numbers and in detail, and it is known that road characteristics tend to mirror the neighboring POIs (Lynch 1960).

To realize the above idea, we propose a two-stage method that obtains the characteristics of each road in the form of a vector from geo-tagged social media texts. Figure 3 overviews our framework.

In the first stage (see Fig. 3), we quantify POI characteristics as feature vectors by applying latent Dirichlet allocation (LDA) (Blei et al. 2003); a topic model reflects typical characteristics of documents and words. Vector representations of POIs can be regarded as topic distributions of geographical characteristics. When estimating vector representations, we take user behavior into consideration when calculating the weights of words so as to extract geographical information more effectively.

The first stage outputs topic distributions of each POI (θ_p) and word distributions of each topic (ϕ_k). The second stage (see Fig. 3) estimates road characteristics from the results of the first stage and the road network data by using the Markov random field method (MRF) (Besag 1986). It outputs a topic distribution of each road (η_r). After that, the relevance score of a road for a given keyword is calculated by using the word distributions of each topic and the topic distributions of each road. Table 4 shows the notation used in this paper.

4.1 POI characteristic estimation by using user-weight-oriented LDA

We obtained the characteristics of POIs by regarding texts posted about one POI as one document. Various information retrieval studies have attempted to obtain vector representations of the characteristics of documents. Latent

Fig. 3 Framework of our two-stage method **Table 4** Notation used in this paper

Symbol	Description
K	Number of topics
P	Number of POIs
D_p	Number of posts related to POI p
D_{pu}	Number of posts related to POI p by user u
P_u	Number of POIs which user u posts
P_r	Number of POIs which nearest road is r
N_d	Number of words in post d
N_p	Number of words in all posts related to POI p
V	Size of vocabulary
ϕ_k	Word distribution of topic k
W_{ud}	Weight of user who write post d
θ_p	Topic distribution of POI p
η_r	Topic distribution of road r
$\mathbf{R}$	Set of all roads
$\mathbf{G}$	Weighted undirected graph which indicates road network Nodes indicate road r , edges indicate relation between roads
$\mathbf{A}_{G_r}$	Set of roads connected to road r
ψ_r	Observed topic distribution of road r
γ_r	Confidence of observed topic distribution of road r
ρ	Prior topic distribution of road
δ	Weight of prior topic distribution of road

semantic indexing (LSI) (Deerwester et al. 1990) decomposes the term-document matrix and obtains a vector representation of each document. LDA (Blei et al. 2003) is the most famous generative statistical model for analyzing

document characteristics, and it also can obtain vector representations. Recently, word and document embedding methods using deep neural networks (DNN) (Mikolov et al. 2013; Chen 2017) have been attracting attention. We used LDA to obtain the characteristics of POIs because the topic-word distribution obtained by it is more suitable for qualitative analyses of characteristics obtained from social media than of word embeddings obtained by DNN-based methods. It is important to analyze what information can be obtained from social media because this study is first one to estimate characteristics of roads automatically from social media.

It is known that applying LDA to short texts such as social media posts does not work well (Zhao et al. 2011). Thus, we must gather multiple posts describing similar content into one document. The previous studies regard multiple tweets related to a user (Hong and Davison 2010; Steyvers et al. 2004) or a hashtag (Mehrotra et al. 2013) as one document. We regard posts related to the same POI as one document for creating documents of adequate length. Each POI is a unit of geographical activities such as sightseeing and shopping. Thus, posts related to the same POI describe similar topics, especially from the viewpoint of geographical characteristics such as fashionable or suitable for shopping. The text of each post is assumed to be generated according to the process shown in Fig. 4. The graphical model of this model is shown in Fig. 5.

However, social media texts include information other than geographical activities, such as cooking at home Hu et al. (2014) or the commercial appeal of shops or

Fig. 4 Generative process of words in each post

```

for Topic  $k = 1$  to  $K$  do
  Draw word distribution  $\phi_k \sim \text{Dirichlet}(\beta)$ 
end for
for POI  $p = 1$  to  $P$  do
  Draw topic distribution  $\theta_p \sim \text{Dirichlet}(\alpha)$ 
  for Post  $d = 1$  to  $D_p$  do
    for Word  $n = 1$  to  $N_d$  do
      Draw topic  $z_{pdn} \sim \text{Categorical}(\theta_p)$ 
      Draw word  $w_{pdn} \sim \text{Categorical}(\phi_{z_{pdn}})$ 
    end for
  end for
end for

```

Fig. 5 Graphical model of LDA for POI characteristics quantification. A shaded circle and unshaded circles represent observed and latent variables, respectively

restaurants. Thus, we need to ensure that only geographical information is extracted from the various posts. User behavior on social media has the following characteristics.

- Active users who tend to post geographical information tend to post texts about many POIs rather than texts about activities in one particular place (usually their home).
- Posts that do not include geographical information are usually posted about POIs by users wanting to draw attention to their activities at home or wanting some commercial appeal.

Considering the above, we weight the texts of each user on the basis of their behavior on social media so as to extract geographical characteristics. The likelihood function of our user-weight-oriented LDA is

$$\prod_{p=1}^P \prod_{d=1}^{D_p} W_{u_d} \prod_{n=1}^{N_d} \sum_{z=1}^K P(w_{pdn}|z)P(z|\theta_p) \quad (1)$$

W_{u_d} is calculated as

$$W_{u_d} = \ln P_u \times \ln \frac{D_p}{D_{p_u}}. \quad (2)$$

The first term in Eq. (2) raises the weights given to users who post on various POIs. The second term penalizes users who post excessive amounts of text about one POI and users who post about POIs that no other user describes.

We use a collapsed Gibbs sampler (Liu 1994) to identify variables ϕ_k for all topics, K , and θ_p for all POIs, P . The sampling procedure iteratively draws samples from a conditional distribution. The conditional distribution of the

topic of word w_{pdn} , the n th word in the d th post related to the p th POI, is calculated as follows.

$$p(z_{w_{pdn}} = k) \propto (N_{pk} + \alpha) \frac{N_{kv} + \beta}{N_k + \beta V} \quad (3)$$

v indicates the index of the term type of w_{pdn} in a vocabulary V . N_{pk} is the sum of weights of words which belong to topic k in POI p . N_{kv} is the count of the times term type v appears in topic k except for w_{pdn} . α and β are the hyperparameters of LDA. We set them to 0.05. After sampling, we calculate the conditional posterior distributions of ϕ_k and θ_p as follows. The v th dimension of ϕ_k is calculated as $\frac{N_{kv} + \beta}{N_k + \beta V}$. The k th dimension of θ_p is calculated as $\frac{N_{pk} + \alpha}{N_p + \alpha K}$.

4.2 Estimating road characteristics by using MRF

Model concept The simplest method for obtaining the topic distribution of each road is allocating the average of those of neighboring POIs to each road. It is known that road characteristics tend to mirror the neighboring POIs (Lynch 1960), so this method will work. However, the characteristics of roads indicated by the topic distribution are not completely the same as those calculated from neighboring POIs, especially when the amount of information posted by users on POIs is small. Moreover, there are often no POIs near some roads, so this simple method is unreliable at obtaining topic distributions. Thus, we should estimate the road characteristics from the resources available.

We propose a method that estimates the road characteristics by using MRF while considering the amount of information on POIs and the road network structure. We use a weighted undirected graph G in which each node indicates a road and each edge indicates the connection between two roads. Considering the amount of information means that topic distributions obtained from a few social media posts are assumed to be more likely to differ from the actual road characteristics than topic distributions obtained from a lot of information; this is realized by determining the confidence of each observation based on

the amount of text posted. Considering that the road network structure means making use of the knowledge that connected roads are more likely to share characteristics and that their connective relationship such as difference in road size determines how they affect each other, this can be realized by determining the similarity of each edge from the road network structure. This method can, through its consideration of similarities, estimate the characteristics of all roads, even those that have no neighboring POIs.

Using MRF for estimating road characteristics MRF is a popular denoising algorithm in image processing (Besag 1986). It handles data consisting of a set of random variables having Markov properties that are described by an undirected graph. The true value of each node can be estimated from noisy observed values by optimizing a cost function.

We estimate the topic distribution of each road by regarding each topic distribution as the true value of each node. We regard topic distributions calculated from POIs as noisy observed values. A graph is obtained from the road network data by regarding roads as nodes and connecting nodes when the roads are connected. Moreover, for making an accurate estimation, we determine the parameters of the cost function by taking into account the amount of information and the road network's structure.

Let us show how MRF estimates road characteristics by taking into account the confidence and similarity of the topic distributions in Fig. 6. The left side shows the relationship of the roads and POIs in a city. The pairs of (Road 1, Road 2) and (Road 2, Road 3) are connected. The neighboring POIs of Road 1 are POI-A and POI-B, while Road 3 has only POI-C. The right side shows the weighted undirected graph G which indicates road networks, the observed topic distribution, and the true topic distribution of each node. Each node r of graph G indicates a road and has an observed topic distribution ψ_r and a true topic distribution η_r , which is a vector representation of the road characteristics. Each edge indicates the relationship of two roads. For example, nodes indicating Road 1 and Road 2 are connected with an edge whose weight is $\text{weight}(\text{Road1}, \text{Road2})$. This weight indicates the similarity of the geographical characteristics of the two roads. We take the example of estimating η_{Road1} , the topic distribution of Road 1. POI-A and POI-B are neighboring POIs of Road 1, and Road 2 is connected to Road 1. By applying MRF, we obtain the estimated value of η_{Road1} , which will be close to ψ_{Road1} and η_{Road2} . ψ_{Road1} is the observed value of the topic distribution of Road 1 and is calculated as the average of $\theta_{\text{POI-A}}$ and $\theta_{\text{POI-B}}$. The closeness of η_{Road1} to ψ_{Road1} and to η_{Road2} depends on the amount of information of POI-A and POI-B and the road network structure. In contrast to the simple allocation method, η_{Road1} is estimated much

Fig. 6 How MRF is applied to the road characteristics estimation task

more accurately because of η_{Road2} . Moreover, Road 2 has no neighboring POI, so η_{Road2} cannot be calculated using the simple allocation method. MRF enables us to estimate η_{Road2} through consideration of η_{Road1} and η_{Road3} .

Cost function To define the cost function of each road r 's topic distribution η_r , we first describe three terms of the cost function.

The first term is the difference in cost between the true topic distribution of road r (η_r) and the observed topic distribution of road r (ψ_r).

$$\gamma_r \|\eta_r - \psi_r\|^2 \quad (4)$$

γ_r indicates the confidence in the observed topic distribution calculated from the POI topic distribution.

$$\gamma_r = \sum_{p=1}^{P_r} \sum_{d=1}^{D_p} W_{ud} * N_d \quad (5)$$

P_r indicates POIs neighboring road r and D_p indicates posts for POI p . W_{ud} is the weight of user who write post d shown in Eq. (2). Confidence is high when the amount of data of each POI is large. Of particular note, the confidence level of POIs that include references to geographical activities should be high. Thus, the POI weight reflects the sum of words in posts related to the POI.

The second term is the difference in cost between the topic distributions of connected roads (road r and road r').

$$\sum_{r' \in \mathbf{A}_{G_r}} \text{weight}(r, r') \|\eta_r - \eta_{r'}\|^2 \quad (6)$$

$\mathbf{A}_{G_r}$ is a set of roads connected to road r in graph G . $\text{weight}(r, r')$, indicates the similarity of the connected roads, which depends on the road network structure.

$$\text{weight}(r, r') = (1 - \text{sizeDiff}(r, r') \times 0.2) \quad (7)$$

We determine $\text{weight}(r, r')$ by considering the difference in road size because connected roads that are similar in type have similar characteristics (Lynch 1960). Our formulation is given below. $\text{sizeDiff}(r, r')$ indicates the difference in road size. For roads of the same size, it returns 0. When the sizes are quite different, it returns 4. In this study, the definition of road size is based on OpenStreetMap (<http://wiki.openstreetmap.org/wiki/Way>).

The third term indicates the difference in cost from that of the prior distribution, and it works as the regularizer when there is a few posts near roads. This term works

$$\|\eta_r - \rho\|^2 \quad (8)$$

We set ρ to be a uniform distribution vector; the value of each dimension of the prior vector is $\frac{1}{K}$.

As result, the cost function is given by Eq. (9).

$$\begin{aligned} c(\eta_r, \psi_r, \mathbf{G}) &= \gamma_r \|\eta_r - \psi_r\|^2 \\ &+ \epsilon \sum_{r' \in \mathbf{A}_{G_r}} \text{weight}(r, r') \|\eta_r - \eta_{r'}\|^2 \\ &+ \delta \|\eta_r - \rho\|^2 \end{aligned} \quad (9)$$

$\mathbf{A}_{G_r}$ is the set of nodes connected to node r . ϵ and δ indicate scaling factors for three terms.

We optimize the cost function of all nodes by using Gibbs sampling (Geman and Geman 1984). The topic distribution of each node r (η_r) can be optimized by iterative sampling from a Gaussian distribution with mean of μ in Eq. (10) and variance of σ^2 in Eq. (11).

$$\mu = \frac{\gamma_r \psi_r + \epsilon \sum_{r' \in \mathbf{A}_{G_r}} \text{weight}(r, r') \eta_{r'} + \delta \rho}{\gamma_r + \epsilon \sum_{r' \in \mathbf{A}_{G_r}} \text{weight}(r, r') + \delta} \quad (10)$$

$$\sigma^2 = \frac{1}{\gamma_r + \epsilon \sum_{r' \in \mathbf{A}_{G_r}} \text{weight}(r, r') + \delta} \quad (11)$$

ϵ and δ are hyperparameters of the model and should be optimized using a method such as cross-validation.

4.3 Calculating the relevance score

To calculate the relevance score of a road r for a given keyword v , our method first represents the given keyword and the road as vectors λ_v and η_r , respectively. Each dimension of λ_v is calculated as follows,

$$\lambda_{vk} = \frac{\phi_{kv}}{\sum_{s \in K} \phi_{sv}} \quad (12)$$

where K is the number of topics and ϕ_{kv} is the occurrence probability of word v in a topic-word distribution of the topic k , ϕ_k , calculated in the first stage. η_r is calculated in the second stage, and the relevance score is the correlation between λ_v and η_r .

(c)OpenStreetMap Contributors

Fig. 7 Area in Tokyo, Japan, analyzed in our experiment. Geo-tagged social media texts and road network data from inside the green box were collected. One hundred roads from inside the red boxes were annotated for the evaluation (color figure online)

5 Experiment

5.1 Data collection

We collected Japanese texts from geo-tagged social media posts and the road network data in a 16-km square area in Tokyo, shown as the green box in Fig. 7. For summarizing the social media posts for each POI, we applied mean-shift clustering (Comaniciu and Meer 2002) to each geo-tagged post and assigned each cluster to a neighboring POI with our POI database, as in the previous study Crandall et al. (2009). Next, we obtained a bag of words for each POI by applying MeCab (Kudo et al. 2004), a Japanese morphological analyzer with mecab-ipadic-NEologd, a system dictionary (Toshinori 2015). As the road network data, we used OpenStreetMap, a user-generated map database.⁶ This yielded 1,553,711 posts by 285,787 users, 238,285 POIs, and 106,899 roads.

Next, we used human annotation to assign correlation scores to some roads for the 25 requirements shown in Fig. 8. These requirements were chosen from the survey described in Sect. 3; they cannot be satisfied by existing route navigation services. The seven requirements on the left were chosen from the quantitative requirements

⁶ <http://www.openstreetmap.org/>.

Quantitative Demand	Subjective Demand	
safe road (安全)	bright (明るい)	quiet (静か)
restaurant (レストラン)	shopping (買い物)	autumn leaves (紅葉)
shopping street (商店街)	rest spot (休憩)	walking (ウォーキング)
a lot of trees (並木)	sightseeing (観光)	strolling (散歩)
parks (公園)	cherry blossom viewing (花見)	famous spot (名所)
red-light district (風俗)	fun (楽しい)	moon viewing (月見)
cats (ネコ)	lively atmosphere (にぎやか)	jogging (ジョギング)
	fashionable (おしゃれ)	cycling (サイクリング)
	night viewing (夜景)	beautiful scenery (景観)

Fig. 8 List of annotated characteristics. The original word used in the experiment is Japanese. English words are translated. Seven characteristics shown on the left side were chosen from the quantitative requirements; the others are from subjective requirements

Table 5 Statistics of whole data set and nearby annotated roads from each area shown in Fig. 7

	# of users	# of POIs	# of posts
All data of Tokyo (green rectangle in Fig. 7)	285,787	238,285	1,553,711
100 roads from Asakusa (area 1 in Fig. 7)	12,150	2986	20,649
100 roads from Shinjuku (area 2 in Fig. 7)	19,315	4058	33,269
100 roads from Shibuya (area 3 in Fig. 7)	14,632	3653	23,405

described in Sect. 3. The other requirements are from the subjective desires.

One hundred annotated roads were selected from each of the three areas; these roads are targeted by many users for work or sightseeing, and many users may need to use a route navigation system in these areas.

1. Asakusa (area 1 in Fig. 7)
2. Shinjuku (area 2 in Fig. 7)
3. Shibuya (area 3 in Fig. 7)

These three areas have different characteristics. Asakusa has many traditional buildings such as Buddhist temples⁷ Shinjuku is a major commercial and administrative center and site of the busiest railway station in the world.⁸ Shibuya is a shopping district, especially popular with young people.⁹ Table 5 lists the statistics of each annotated one hundred road set in the three areas.

The road segments of the OpenStreetMap format are too short for annotation because they terminate at the first intersection. Thus, we connected several road segments together into one road and set an endpoint only when it intersected a road bigger than a secondary road. The definition of road size is based on OpenStreetMap (<http://wiki.openstreetmap.org/wiki/Way>).

We collected road annotations from participants who lived more than two years in the area examined. Participants evaluated how well the road matched each criterion. They were presented with a map that showed a road and four photographs taken on the road (Fig. 9). Camera directions are indicated by arrows on the map. The task

was to assign a score to each road for each criterion (i.e., they responded on a Likert scale: strongly agree, agree, undecided, disagree, and strongly disagree). The average standard deviation of the annotated scores for each road/keyword combination was 0.54.

5.2 Relation between road characteristics and POI categories

To confirm the assumption that the characteristics of roads tend to mirror those of neighboring POIs (Sect. 1), we analyzed the distribution of categories of POIs neighboring roads. We counted the categories of POIs neighboring roads having annotation scores higher than 4 for each characteristic shown in Fig. 8. The definitions of the categories are from those of Foursquare's top level categories.¹⁰ Figure 10 shows the distribution of categories of POIs neighboring roads having annotation scores higher than 4 for “fashionable,” “sightseeing,” and “quiet.” Number of road having enough score is 14 for “fashionable,” 9 for “sightseeing,” and 14 for “quiet.”

We conducted a Chi-square test between the category distribution of POIs for each characteristic shown in Fig. 8 and the overall category distribution of those POIs. We found the significant dependency ($p < .01$) of POI category distributions on road characteristics. This result indicates there is the relation between characteristics of POIs and those of roads.

We also found a qualitative relation between the characteristics of roads and those of POIs. POIs near “Quiet” roads are categorized as “College and University” and “Professional and Other Places” more frequently than

⁷ <https://en.wikipedia.org/wiki/Asakusa>.

⁸ <https://en.wikipedia.org/wiki/Shinjuku>.

⁹ <https://en.wikipedia.org/wiki/Shibuya>.

¹⁰ <https://developer.foursquare.com/categorytree>.

Fig. 9 Map image and photographs shown to annotators. The map image on the left shows where the evaluated road is. The photographs on the right are taken on the evaluated road

Fig. 10 Distribution of POI categories near roads having annotation scores higher than 4 for some characteristics

“fashionable” or “sightseeing” roads. These categories’ POIs are quieter than the POIs of other categories. Comparing “fashionable” and “sightseeing,” we see that there more “Shop and Service” POIs near “fashionable” roads. The typical “fashionable” POI is a shop for youths and young adults; thus, it is a reasonable result.

5.3 Comparing methods

We compared nine methods including ours, UWLDA–MRFAR (User-weight-oriented LDA and MRF considering amount of information and road network structure) in terms of normalized Discounted Cumulated Gain (nDCG) (Manning et al. 2008) and precision@5.

nDCG is a measure of the whole ranking quality that is used in information retrieval. In this experiment, by regarding each road as a retrieved document, we could measure the quantifying quality of each method in terms of nDCG. Before calculating nDCG, we produced a ranking of roads based on the relevance score of each road estimated by each method. The nDCG score of the estimation of roads from area a is calculated as follows.

$$\text{DCG} = \sum_{i=1}^{R_a} \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)} \quad (13)$$

$$\text{nDCG} = \frac{\text{DCG}}{\text{idealDCG}} \quad (14)$$

R_a indicates the number of roads from area a (100 in this experiment). r_i indicates the road ranked i th among the roads from area a . rel_i indicates the relevance score of r_i . In this experiment, the relevance score of roads having mean annotation scores higher than 4 was 3, and the relevance score of roads having annotation scores higher than 3 was 1. The DCG score becomes high when the road annotated with a higher score is ranked higher. The nDCG score is the normalized DCG score calculated by dividing DCG by idealDCG, which indicates the highest possible DCG score and it is produced when the rank of the estimation result is perfect. A high nDCG score indicates that the document annotated as relevant to the keyword query is given a high ranking. In this study, a good nDCG result means that the method could output a high score for roads relevant to the requirement or desire.

Table 6 Two-stage methods, i.e., POI characteristics quantification and road characteristics estimation, tested for comparison

Method	Stage 1 Quantifying POI characteristics	Stage 2 Estimating road characteristics
LDA	LDA	Average topic distribution of neighbor POIs
UWLDA	UWLDA	Average topic distribution of neighbor POIs
LDA-MRF	LDA	MRF
UWLDA-MRF	UWLDA	MRF
LDA-MRFA	LDA	MRFA
UWLDA-MRFA	UWLDA	MRFA
LDA-MRFAR	LDA	MRFAR
UWLDA-MRFAR	UWLDA	MRFAR

Precision@5 evaluates the accuracy of finding roads strongly related to the keyword. It is calculated by dividing the number of roads ranked higher than or equal to 5 by 5. It can evaluate the accuracy of finding roads strongly related to the keyword.

Both scores were calculated for each combination of each query and method, and scores for each method were calculated by averaging the scores for all queries.

We compared our method with BM25 (Manning et al. 2008) and eight two-stage methods consisting of POI characteristics quantification and road characteristics estimation. The components used in each stage depended on the method. Table 6 shows the list of methods and those components. All models were learned with the data from inside the green rectangle of Fig. 7.

- **BM25** This standard method of information retrieval for calculating the relevance score is based on word count (Robertson and Walker 1994). It considers document length (average length of all documents) in addition to the number of times a keyword occurs. Below is the formula of BM25 score the combination of road r and query keyword q .

$$\text{BM25}(r, q) = \text{IDF}(q) \times \frac{\text{TF}(q, r) \times (k_1 + 1)}{\text{TF}(q, r) + k_1 \times (1 - b + b \times \frac{|R|}{\text{avgnw}})} \quad (15)$$

$$\text{IDF}(q) = \log \frac{N - n(q) + 0.5}{n(q) + 0.5} \quad (16)$$

TF is the term frequency, the number of times a query keyword q occurs near road r . avgnw is the average number of words near each roads. IDF is the inverse document frequency, and it indicates the how important word q is. $n(q)$ is the number of roads where the keyword q occurs and N is the number of roads. k_1 and b are the hyper parameters, and we set k_1 to 2 and b to 0.75 as in previous studies (Robertson and Walker 1994).

- **LDA** The topic distribution of each road is calculated as the average of those of neighboring POIs. The topic

distributions of POIs are obtained by using LDA. Its score is calculated as the correlation between the keyword vector and the road vector, as detailed in Sect. 4.3. We set the number of topics to 100. This setting is also used in the following methods.

- **UWLDA (User-weight-oriented LDA)** Topic distributions of POIs are obtained from user-weight-oriented LDA described in 4.1. The way of calculating the score is the same as that of LDA.
- **LDA-MRF** The topic distribution of each road is estimated using MRF based on the topic distributions of the POIs obtained from LDA. The confidences of the obtained topic distributions from the POIs are not considered. All edges in the graph used in MRF have the same weight. We set ϵ to 4500 and δ to 1 on the basis of a grid search; this setting is also used in the following methods.
- **UWLDA-MRF** The topic distributions of POIs are obtained from user-weight-oriented LDA. The remaining part is the same as in LDA-MRF.
- **LDA-MRFA (MRF considering amount of information)** The topic distribution of each road is estimated using MRF by considering the amount of information near each road as the confidence of the topic distributions obtained using LDA.
- **UWLDA-MRFA** The topic distributions of POIs are obtained using user-weight-oriented LDA. The remaining part is the same as in LDA-MRFA.
- **LDA-MRFAR (MRF considering amount of information and road network structure)** This is the method proposed in Nishimura et al. (2016). The topic distribution of each road is estimated using MRF by considering the road network structure as a similarity based on the road size in addition to the amount of information.
- **UWLDA-MRFAR (User-weight-oriented LDA and MRF considering amount of information and road network structure)** This is the method proposed here. The topic distribution of each road is estimated using MRF by considering the road network structure and the

amount of information based on topic the distributions of POIs obtained by user-weight-oriented LDA.

Figures 11 and 12 show the nDCG and precision@5 of each method. Each score is the average of the scores for each combination of a keyword indicating a characteristic in Fig. 8 and an area in Fig. 7. The results of these figures show a similar tendency. BM25 is not accurate; its precision@5 score is worse than random selection. Few keywords are posted near annotated roads, so even the standard method cannot calculate the relevance score accurately. LDA offers improved accuracy in terms of both nDCG and precision@5 ($p < .01$; paired t test). This shows that applying topic-based matching is effective for quantifying the characteristics of roads. The score of UWLDA is worse than that of LDA. User-weight-oriented LDA focuses on obtaining an accurate topic distribution from POIs containing geographical information; thus, the characteristics of roads neighboring POIs that do not contain such information are not estimated effectively by UWLDA. LDA-MRFA scores higher, especially in nDCG, than LDA. UWLDA-MRFA is an improvement on UWLDA because UWLDA (and LDA) cannot calculate the scores of many roads due to the lack of POIs. On the other hand, LDA-MRFA and UWLDA-MRFA can calculate the scores of all roads and consequently exhibit higher nDCG scores, and thus higher overall ranking quality. By contrast, LDA-MRF and UWLDA-MRF yield worse scores even though they calculate scores for all roads. This indicates that a

confidence based on the amount of information is necessary for improving estimation quality. Moreover, UWLDA-MRFA scores higher than LDA-MRFA, especially in terms of nDCG. The confidences of the POIs are based on user weights; thus, UWLDA-MRFA can propagate geographical information selectively and estimate road characteristics effectively. There is no significant difference between UWLDA-MRFAR and UWLDA-MRFA with regard to nDCG or precision@5. We assume that the characteristics of two roads that are connected and are of the same size are similar for UWLDA-MRFAR; however, whether this assumption is correct or not depends on the characteristics of the area. UWLDA-MRFAR scores significantly better than UWLDA-MRFA in the Asakusa area and Shinjuku area ($p < .1$; paired t test). On the other hand, UWLDA-MRFA is better in the Shibuya area ($p < .1$; paired t test). In Asakusa and Shinjuku, there are many places where the characteristics greatly change when entering a narrow road from a big street. In contrast, Shibuya has several areas whose characteristics (e.g., “commercial district”) do not depend on the road size.

However, the standard deviations shown in Figs. 11 and 12 are large. This result indicates the accuracy of each method depends on the combination of the query keyword and the area. We analyze the dependency in the following part.

The above results indicate that our method, UWLDA-MRFA, achieves statistically significant improvements over LDA-MRFAR, the method proposed in our previous study Nishimura et al. (2016). However, UWLDA-MRFA showed no significant difference.

5.4 Analysis of the results

Figure 13 shows the nDCG score of each keyword estimated by UWLDA-MRFA. The scores of “jogging” and “strolling” are high, because information on these activities is frequently posted on social media and they are activities done on roads. Road characteristics related to such keywords are often found in social media posts. On the other hand, the scores of “moon viewing” and “shopping street” are low. There are no roads suitable for these keywords in the areas examined in this study. As a result, there is not enough information related to the keyword on roads in the target areas; thus, our method cannot estimate road characteristics accurately.

The average nDCG scores for quantitative requirements (0.665) and that of subjective requirements (0.665) are similar, and there are no significant differences between them. As mentioned in Sect. 3, some characteristics have both quantitative and subjective aspects. Moreover, characteristics having a subjective aspect tend to be mentioned

Fig. 11 nDCG scores of each method. Error bars indicate the standard deviation of the scores for each combination of a road and a keyword (**: $p < .05$, ***: $p < .01$; paired t test)

Fig. 12 Precision@5 scores for each method. Error bars indicate the standard deviation of the scores for each combination of a road and a keyword (**: $p < .05$, ***: $p < .01$; paired t test)

frequently in social media and tend to be estimated accurately in our proposed method. For example, “safe road” and “a lot of trees” are quantitative requirements; however, they have a subjective aspect, so their nDCG scores are high. This is why there was no significant difference between the quantitative requirements and subjective requirements.

Figure 14 shows the nDCG scores for each combination of a keyword and area as estimated by UWLDA–MRFAR. The scores depend on the area even for the same keyword. For example, the score of the keyword “sightseeing” is high for Asakusa, but not for Shinjuku. This is because Shinjuku is a business district, so there are few streets suitable for sightseeing. The scores of “jogging” and “strolling” are high, because information on these activities is frequently posted on social media and they are activities done on roads. Road characteristics related to such keywords are often found in social media posts. On the other hand, the scores of “moon viewing” and

“shopping street” are low. There are no roads suitable for these keywords in the areas examined in this study. As result, there is not enough information related to the keyword on the roads in the target areas; thus, our method cannot estimate the road characteristics accurately.

Figures 15 and 16 are visualizations of the human annotated scores and the scores of our method (UWLDA–MRFAR) and two baselines (LDA and LDA–MRF) for the keyword “Fashionable.” Brighter green indicates a higher score. Figure 15 also shows social media post density near roads. In that map, brighter green means there are more posts near roads instead.

Figure 15 shows that our method can calculate accurate scores for all roads in the Shinjuku area, unlike the existing methods described in Sect. 2. The LDA method cannot calculate the score of some roads for lack of information, while the results of LDA–MRF are too uniform because all the observed topic distributions for each road are treated equally. The results of our method are overall similar to the manually annotated scores. Actually, the nDCG score of “Fashionable” is high in the Shinjuku area, 0.8084. This is because our method can consider the confidence of information of each road based on the post density. The post density map shows that there are many posts near a station; thus, the accuracy of estimation is high for roads near a station.

In the Asakusa area, the results of human annotation and our method look different; the nDCG score is 0.5633. Asakusa is an area with many traditional shops and buildings. The social media information is mainly about such traditional attributes. Accordingly, there is little information about “Fashionable,” and our method failed to estimate road characteristics for lack of information. On the other hand, the nDCG score of the keyword “Sightseeing” is 0.8648 in the Asakusa area. This result indicates

Fig. 13 nDCG score of each keyword calculated by UWLDA–MRFAR. The left seven queries (red bars) are about requirements classified into the quantitative requirements described in Sect. 3. The others are classified into subjective requirements

Fig. 14 nDCG scores for each combination of keyword and area calculated by UWLDA–MRFAR. The top row shows the queries of quantitative requirements. The other rows show subjective requirements

Fig. 15 Visualization of human annotated scores and scores calculated by our method (UWLDA-MRFAR) and two baselines (LDA and LDA-MRF) for the keyword “Fashionable” in Shinjuku (area 2),

and social media post density near roads. Brighter green indicates a higher score or more posts. The color of each road under the highlight indicates the size of road

that our method can effectively estimate the characteristics that represent the target area.

We tried to determine whether our similarity settings in Eq. (8) were correct or not by annotating the results qualitatively. Figure 16 shows a visualization of the annotation for the Shibuya area for the characteristics of “Fashionable roads.” Pairs of roads that are connected directly and are of the same kind tend to have similar annotations. These results indicate that the settings shown in Sect. 4.2 reflect the relationships of empirical road characteristics.

Although the estimates of our method, UWLDA-MRFAR, of the Shibuya area are on the whole similar to the human annotations, several instances were not accurate. The nDCG score is 0.7817. The red boxes on the Fig. 16 indicate the area of Omotesando,¹¹ which is a popular “Fashionable” street. In this case, our method could not estimate the scores of these roads. There are not so many posts near the roads, and residential roads are connected to this street. The estimated road characteristics are influenced by the information obtained in the housing estate. MRF can be used to estimate the characteristics of roads for which there are few posts. Unfortunately, if the

connected roads have different characteristics, the results are not useful.

Figure 17 shows the histogram of weights distributions of users and POIs. The user weights follow a Gaussian distribution, and users who frequently post geographical information are assigned weights that are more than five times those of users who do not post such information. These weights indicate the influence of each user as regards the estimation of road characteristics. The weights of POIs are widely distributed, ranging from less than 10 to more than 1000. These weights impact the MRF as confidence parameter settings. This enables us to estimate road characteristics much more accurately, as shown by the significant difference between LDA-MRFAR and UWLDA-MRFAR in Fig. 11.

Table 7 shows how many topics obtained by our user-weight-oriented LDA are related to each category shown in Sect. 3. The numbers of topics related to the Experience and Impression categories are bigger than those of the other categories. As shown in Sect. 3, the information contained in these categories is currently not enough. However, social media does provide us with useful information for estimating road characteristics that is not provided by existing route navigation services.

¹¹ <https://en.wikipedia.org/wiki/Omotesand%C5%8D>.

Fig. 16 Visualization of human annotated scores and scores calculated by our method (UWLDA-MRF) for the keyword "Fashionable" in Asakusa (area 1) and Shibuya (area 3). Brighter green

indicates a higher score. The color of each road under the highlight indicates the size of road. Red boxes on the right indicate the roads for which our method could not calculate scores accurately

Fig. 17 Histogram of weight distributions of users and POIs. The horizontal axis is bins of weights, and the vertical axis is number of data belonging to each bin

Our method successfully discovered geographical characteristics of POIs useful for identifying routes that can satisfy the users' requirements. Table 8 lists topics related to Requirements of devices and facilities, Desired experience near roads and Desired impression. As shown, topics

of Devices and facilities are related to Requirements of devices and facilities. These topics describe what shops, sightseeing spots, or restaurants are near roads. Topic sample 1 of the Devices and facilities category includes Japanese shrines and the Imperial Palace. Other words

Table 7 Number of topics related to each category shown in Sect. 3

Physical characteristics	Statistics	Devices and facilities	Experience	Impression	Others
0	0	17	62	46	16

Table 8 Sample of topics obtained by using user-weight-oriented LDA. Topics are manually associated with requirements categories. This table is translated from Japanese

Category	Topic sample 1		Topic sample 2	
	Word	Probability	Word	Probability
Devices and facilities	Shrine	0.03882	Pasta	0.045102
	Imperial palace	0.024531	Delicious	0.0319
	New year's visit to a shrine	0.013914	Italian	0.027321
	Run	0.013574	Lunch	0.023558
	Tokyo	0.012644	Tomato	0.01752
Experience	Cherry blossom viewing	0.042023	Rest	0.014721
	Full bloom	0.024263	Business	0.014545
	Cherry blossom	0.023635	Tully's	0.011824
	Night cherry viewing	0.022148	Finish	0.010528
	Park	0.018645	Doutor	0.008462
Impression	Buy	0.021032	Come	0.017253
	Come	0.009888	Event	0.012594
	Enchanting	0.009555	Today	0.012067
	Make	0.00929	Exciting	0.009319
	Cute	0.008928	Come	0.008993
	Desire	0.008385	Participate	0.00788

indicate specific experiences near them. For example, a “New Year’s visit to a shrine” is when people go to a Japanese shrine and pray for happiness on New Year’s day. The Japanese Imperial Palace is a popular spot for runners and joggers. The topics of the Experience category are related to Desired experience near roads. These topics describe what activities users can experience near roads (e.g., sightseeing). Topic sample 1 of the Experience category includes cherry blossom viewing, which is a specific activity in some areas. Topic sample 2 of the Experience category includes rest spots such as coffee shops or restaurants. The topics of the Impression category are related to Desired impression. These topics refer to impressions such as beautiful or fashionable that users may experience on each road. Topic 1 of the Impression category covers the impressions of young people when they go to buy items and clothes. Topic 2 is the impressions people feel when they participate in various events such as concerts. Estimating which roads have these topics is useful for understanding road characteristics and providing specific routes that can satisfy users’ requirements and desires. For example, many users want to find rest places, as shown in Table 3, and we can find roads that have rest places by referring to topic 2 of the Experience category.

6 Discussion

Here, we discuss the limitations of our method as revealed by an error analysis and outline future work.

As shown in Sect. 5.4, our method performed poorly for keywords that SNS users rarely posted on directly. We had thought that textual information in social media would cover a wide range of user demands for route navigation; unfortunately, social media texts are incomplete sources for quantifying user requirements and desires for all areas. For the topics not popular with SNS users, we plan to explore the use of other resources, such as geo-tagged photographs for quantifying such non-textual information. For context-dependent topics, we must identify contexts by considering users’ interests (Teevan et al. 2005) or by interacting with the user Dumais et al. (2003).

We tested a model which consider the intersection angle between roads when estimating road characteristics on the assumption that roads with small intersection angles have higher similarity than those with large intersection angles. However, the accuracy did not improve as a result. This indicates that the difference in angle does not affect the similarity of road characteristics. This result is not intuitive. In the future, we will analyze this problem and improve the model so that it considers intersection angles adequately. In our previous study Nishimura et al. (2016),

we used the length of roads when calculating confidence levels; however, we found that this did not improve the estimation quality compared with that of user-weight-oriented LDA. This is because user-weight-oriented LDA can calculate the confidence of each POI and road; thus, information related to geographical characteristics can be propagated through roads accurately without considering their length.

We used LDA to obtain the characteristics of the POIs; however, various alternative methods exist, including ones based on DNNs (Chen 2017). We should compare these methods and choose the most appropriate one for obtaining the necessary characteristics.

This study assumed that the similarity of each road pair is always the same regardless of the characteristics. However, this assumption is not strictly correct; for example, “Fashionable” characteristics may be similar for connected road pairs in many cases, but the ‘Existence of Restaurants’ is not. In the future, we should consider the relationship between characteristics and similarities.

The surveyed urban areas in the experiment of this study have relatively many social media posts. However, there are many areas that have too few posts to permit MRF to estimate road characteristics. Estimations may, however, be possible by considering other factors such as the relationship between road type or size and geographical characteristics.

As shown in Sect. 5.4, the accuracy of our method depends on the target road and keyword query. However, the correlation coefficient between the nDCG scores and the occurrence frequency of the query keyword neighboring each of the one hundred annotated roads was 0.13. To calculate accurate confidence scores, we need to analyze the results of the road characteristics estimation and various statistics about keywords and target roads.

We believe that combining our method with a multi-criteria path-finding algorithm (Clímaco and Pascoal 2012) will enable us to satisfy compound requirements such as beautiful and quiet routes. This is because our method can calculate a score for each query keyword and road pair, and the algorithm can find the paths that yield the highest scores possible for all query keywords. However, it cannot handle negation expressions such as “not dirty.” We will solve this problem by semantically transforming the negation expressions into positive ones, for example, transforming “not dirty” into “clean.”

7 Conclusion

For realizing a novel new route navigation service that can satisfy various user demands, we tackled the problem of quantifying road characteristics, which was shown to be an

important task by our user survey. We tried to quantify all characteristics indicated by a keyword from geo-tagged social media without having to resort to expensive methods such as manual annotation or surveys. The difficulty of quantifying road characteristics is the lack of sufficient quantitative data associated with the road characteristics that correspond to all user possible requirements and desires. Our key idea is based on the knowledge that road characteristics tend to mirror neighboring POIs and POIs tend to be mentioned in sufficient number and detail in social media texts.

We proposed a two-stage method for estimating road characteristics. The first stage quantifies POI characteristics by using LDA for assigning user weights according to user behavior. The second stage estimates road characteristics from those of neighboring POIs and the road network associated with the POIs by using MRF. After these two stages, we can calculate the score of each road for any given query keyword by using the word distributions and topic distributions obtained in the above two stages.

We evaluated our method in an experiment on real datasets of social media texts and road networks. Our method improved nDCG by 0.5 and precision@5 by 0.2 compared with the conventional baseline for 300 roads and 25 keywords that indicated geographical characteristics. We also found that LDA settings that reflect user weights improve nDCG performance. The results indicate that we can process geo-tagged social media texts to quantify road characteristics.

The differences from the related conference paper (Nishimura et al. 2016) are listed below.

- In the survey of user requirements, we classified the requirements into five categories and analyzed each category.
- Our method emphasizes (weights) words posted by users who post from various POIs because posts by such users may include useful information for obtaining geographical characteristics.
- An experiment that analyzed the distribution of categories of POIs neighboring roads showed there is a relationship between characteristics of roads and those of POIs. After that, we clarified which areas and keywords are difficult for our method to estimate road characteristics and tried to determine why our method did not work well for these cases.

As discussed in Sect. 6, we will examine the use of other resources such as geo-tagged images and human mobility data. We will also try to create a new route navigation service based on the results of this study. It will help users to find the route they really want.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Anagnostopoulos A, Petroni F, Sorella M (2016) Targeted interest-driven advertising in cities using twitter. In: ICWSM, pp 527–530
- Andrade L, Silva MJ (2006) Relevance ranking for geographic ir. In: GIR
- Besag J (1986) On the statistical analysis of dirty pictures. *J R Stat Soc Ser B Methodol* 48(3):259–302
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Çelikten E, Falher GL, Mathioudakis M (2016) Extracting patterns of urban activity from geotagged social data. ArXiv preprint [arXiv:1604.04649](https://arxiv.org/abs/1604.04649)
- Chen M (2017) Efficient vector representation for documents through corruption. In: ICLR
- Clímaco JC, Pascoal M (2012) Multicriteria path and tree problems: discussion on exact algorithms and applications. *Int Trans Operat Res* 19(1–2):63–98
- Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. *IEEE Trans Pattern Anal Mach Intell* 24(5):603–619
- Corcoran P, Jilani M, Mooney P, Bertolotto M (2015) Inferring semantics from geometry-the case of street networks. In: SIGSPATIAL
- Crandall DJ, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos. In: Proceedings of WWW, pp 761–770
- Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. *J Am Soc Inf Sci* 41(6):391–407
- Dumais S, Cutrell E, Cadiz JJ, Jancke G, Sarin R, Robbins DC (2003) Stuff i've seen: a system for personal information retrieval and re-use. In: SIGIR
- Galbrun E, Pelechris K, Terzi E (2014) Safe navigation in urban environments. In: SIGKDD workshop on urban computing, p. 9
- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6:721–741
- Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc, pp 289–296
- Hong L, Ahmed A, Gurumurthy S, Smola AJ, Tsioutsoulis K (2012) Discovering geographical topics in the twitter stream. In: WWW, pp 769–778
- Hong L, Davison BD (2010) Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics, pp 80–88
- Hu Y, Manikonda L, Kambhampati S et al (2014) What we instagram: a first analysis of instagram photo content and user types. In: ICWSM, pp 595–598
- Jameson A (2009) Adaptive interfaces and agents. *Human-Comput Interact Des Issues Solut Appl* 105:105–130
- Jolliffe I (2002) Principal component analysis. Wiley Online Library, Hoboken
- Kudo T, Yamamoto K, Matsumoto Y (2004) Applying conditional random fields to japanese morphological analysis. *EMNLP* 4:230–237
- Le Falher G, Gionis A, Mathioudakis M (2015) Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. In: ICWSM, pp 228–237
- Liu JS (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J Am Stat Assoc* 89(427):958–966
- Lynch K (1960) The image of the city, vol 11. MIT Press, Cambridge
- Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
- Mehrotra R, Sanner S, Buntine W, Xie L (2013) Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: SIGIR, pp 889–892
- Mei Q, Cai D, Zhang D, Zhai C (2008) Topic modeling with network regularization. In: WWW, pp 101–110
- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: NIPS, pp 3111–3119
- Nishimura T, Nishida K, Toda H, Sawada H (2016) How fashionable is each street?: Quantifying road characteristics using social media. In: ASONAM, pp 429–436
- Purves R, Jones C (2011) Geographic information retrieval. *SIGSPATIAL Special* 3(2):2–4
- Quercia D, Aiello LM, Schifanella R, Davies A (2015) The digital life of walkable streets. In: WWW, pp 875–884
- Quercia D, O'Hare NK, Cramer H (2014) Aesthetic capital: what makes London look beautiful, quiet, and happy?. In: CSCW, pp 945–955
- Quercia D, Schifanella R, Aiello LM (2014) The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In: Hypertext, pp 116–125
- Quercia D, Schifanella R, Aiello LM, McLean K (2015) Smelly maps: The digital life of urban smellscape. In: ICWSM
- Robertson SE, Walker S (1994) Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR, pp 232–241
- Sengstock C, Gertz M (2012) Latent geographic feature extraction from social media. In: SIGSPATIAL, pp 149–158
- Skoumas G, Schmid KA, Jossé G, Züfle A, Nascimento MA, Renz M, Pfoser D (2014) Towards knowledge-enriched path computation. In: SIGSPATIAL, pp 485–488
- Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T (2004) Probabilistic author-topic models for information discovery. In: KDD, pp 306–315
- Teevan J, Dumais ST, Horvitz E (2005) Personalizing search via automated analysis of interests and activities. In: SIGIR
- Toda H, Yasuda N, Matsuura Y, Kataoka R (2009) Geographic information retrieval to suit immediate surroundings. In: GIS, pp 452–455
- Toshinori S (2015) Neologism dictionary based on the language resources on the web for mecab. <https://github.com/neologd/mecab-ipadic-neologd>
- Yang B, Kaul M, Jensen CS (2014) Using incomplete information for complete weight annotation of road networks. *IEEE Trans Knowl Data Eng* 26(5):1267–1279
- Yin Z, Cao L, Han J, Zhai C, Huang T (2011) Geographical topic discovery and comparison. In: WWW, pp 247–256
- Zhao WX, Jiang J, Weng J, He J, Lim E-P, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: Advances in information retrieval. Springer, pp 338–349
- Zhuang C, Ma Q, Liang X, Yoshikawa M (2015) Discovering obscure sightseeing spots by analysis of geo-tagged social images. In: ASONAM, pp 590–595