Identifying salient topics for personalized place similarity

Benjamin Adams
Centre for eResearch
The University of Auckland, New Zealand
b.adams@auckland.ac.nz

Martin Raubal
Institute for Cartography and Geoinformation
ETH Zürich, Switzerland
mraubal@ethz.ch

Abstract

The ability to find similar places is an important component to geographic information retrieval applications as varied as travel recommendation services, marketing analysis tools, and socio-ecological research. Using generative topic modelling on a large collection of place descriptions, we can represent places as distributions over thematic topics, and quantitatively measure similarity for places modelled with these topic signatures. However, existing similarity measures are context independent; in cognitive science research there exists evidence that when people perform similarity judgments they will weigh properties differently depending on personal context. In this paper we present a novel method to re-weight the topics that are broadly associated with a place, based on users' interests inferred from sample place similarity rankings. We evaluate the method by training topics associated with texts about places, and perform a user study that compares user-provided similar places to those provided by automatically personalised place rankings. The results demonstrate improved correspondence between user rankings and automated rankings when personalised weights are applied.

1 Introduction

People commonly use known places as referents for communicating information about other parts of the world. For example, when we describe a place as being *like Canberra, Australia*, that description comes laden with semantics that allow us to infer attributes about that place that are otherwise left unspoken. These implicit attributes of places often differ from explicitly represented properties of places, such as census data, econometric data, and spatial footprints, that are commonly found in geographic information databases. Instead, these attributes are emergent from people's experiences and are revealed through their communication about those places. However, the implicit attributes that a person associates with a given place will also be highly contextual, and thus personalisation is required in order to make useful geographic information retrieval applications that utilise these kinds of attributes.

Operationalising background knowledge about places is potentially useful for several kinds of information applications, including geographic information retrieval and exploration systems (Larson, 1996; Jones and Purves, 2008), mobile applications, tools for social science research and education, large scale proposals such as the Digital Earth system (Grossner et al., 2008), and applications in the nascent field of spatial humanities (Bodenhamer, 2010). Additionally, in the technology sector, place knowledge is used to create location-based advertising services (Ranganathan and Campbell, 2002); with an understanding of the locational and temporal context in which a mobile user is placed, advertising content can be better targeted (Banerjee and Dholakia, 2008). A core functionality that we require in these different applications is the ability to find semantically similar places based on an operationalisation of this implicit knowledge.

In this paper we consider the problem of personalising the search for similar places, which are represented as vectors of probability values associated with a set of features or attributes. For demonstration we focus on characterising *places*

In: S. Winter and C. Rizos (Eds.): Research@Locate'14, Canberra, Australia, 07-09 April 2014, published at http://ceur-ws.org

as mixtures of topics derived from probabilistic topic modelling on a corpus of place descriptions. These descriptions can take many forms, including travel entries, encyclopaedia articles, newspaper articles, and social media postings. These data sources are plentiful and all provide insight into the experiences of people in various places and, in particular, attributes that are commonly described for those places. Topics are trained using latent Dirichlet allocation (LDA) topic model inferencing on a set of these crowdsourced documents (Blei et al., 2003). Topic model training applies Bayesian inference to arrive at a representation of each place as a probability vector over a finite set of topics. Once we have a topic distribution for each place, we can compare their similarities in terms of the relative entropy of these topic distributions. This measure captures the similarity of the places as derived from an aggregation of descriptions from many people. In other words, the similarity of two places is defined as similarity of the prototypical or "average" representations of the places. There is psychological evidence that for many kinds of categories a prototypical instance of the category can be represented as an average from a set of exemplars in this manner (Rosch, 1978).

However, there is also evidence that when performing similarity (or dissimilarity) judgements on a set of stimuli, people will assign higher salience to some properties than others (Medin et al., 1993). In conceptual space theory, it is proposed that these salience differences can be modelled as weights on the quality dimensions on which these similarity measurements are based (Gärdenfors, 2000; Raubal, 2004). When people judge the similarity of two places, they choose a set of salient properties on which to compare the two places. This set of properties will be smaller than the total set of possible properties on which one could possibly compare the two places (Tversky, 1977). In addition, the set of properties that are salient will vary from person to person and depend on the places being compared. The challenge, therefore, is to identify which weighting to use in a particular context. This feature selection problem becomes one of further reducing the dimensionality of the topic space to a subset of topics that are relevant to a user in a specific context.

In this paper we present a methodology to 1) generate a prototypical mixture of topics associated with places based on crowdsourced natural language descriptions, 2) identify the topics of interest to an individual user, and 3) provide a personalised ranking of similar places to a source place based on the inferred interests of the user. We propose a method to automatically re-weight the topics using a sample similarity ranking for a small subset of the data.

In the following section, we provide background on computing place, personalised search and ranking, and topic models. Section 3 describes our method to identify salient topics. Following that, Section 4 presents the results of a user evaluation of the method. Finally, we conclude and point to future work.

2 Background

This section provides background material on computing place, personalised search and ranking, place recommendation, and topic modelling algorithms used in this paper.

2.1 Place

Several researchers have noted that place is a subjective, experiential phenomenon that is socially constructed (Relph, 1976; Tuan, 1977). Simply put several of the characteristics of places that are salient to people cannot be captured by a spatial footprint and structured data, such as population or median income. The difficulty in developing place representations that take into account common-sense understandings has meant that, to a great degree, place has been represented in a limited manner in geographic information systems (GISs). However, there has been a renewed interest in recent years in modelling place more holistically as it becomes clear that traditional models are insufficient for the needs of many technologies (Jordan et al., 1998; Winter et al., 2009). An increasingly vast amount of data is available online that describe places and types of places in natural language. They come in a variety of forms: e.g., travelogues, encyclopaedia entries, microblogs, literary works, social media sites, etc. These documents provide us with an extraordinary opportunity to understand places not in terms of quantitative, tabular data, but rather in terms of how people write about them.

2.2 Personalised search and ranking

Personalised search and ranking on the World Wide Web has been an ongoing research area for several years (Pitkow et al., 2002; Teevan et al., 2005; Micarelli et al., 2007). Pitkow et al. (Pitkow et al., 2002) have identified two main strategies for personalising search: 1) query augmentation, which uses additional contextual information to refine the query; and 2) result processing, which filters the search results based on a user model. In addition to algorithmic approaches to personalisation, user interface design can play a large role in how well search results can be personalised (Teevan et al., 2005).

Much of the personalisation research relies on using previous search behaviour to develop a model of the user's interests and using this model to personalise search results on the web (Teevan et al., 2005). Developing a user profile based on implicit information has several advantages over asking the user for explicit feedback, because not only does gathering

explicit feedback impose an added burden on the user but also users can provide inaccurate information (Speretta and Gauch, 2005). In mobile search, user-based personalisation can be augmented by location-based personalisation and these factors can interrelate to model how a user's interests change based on location (Bouidghaghen et al., 2011). Formal ontologies can be used to align user interests that are inferred from search history with well-defined concepts described in a semantic hierarchy (Daoud et al., 2007).

Once there is a model that associates user categories with specific interests, it is possible to use different methods of characterising user similarity to provide personalised results for categories of users (McKenzie et al., 2013). Guy et al. (Guy et al., 2010) have explored comparing user similarity in social media along 3 different axes of people, things, and places.

2.3 Semantic similarity in GIScience

The importance of context in semantic similarity measurement is a well-studied research topic in geographic information science (Raubal, 2004; Rodríguez and Egenhofer, 2004; Janowicz et al., 2011). Commonly, context is modelled by applying weights to the attributes that used to describe geospatial features; however, finding the appropriate weights is a challenge (Keßler, 2012). Semantic referencing is an algorithm schema that uses the idea of control similarities provided by a user to semi-automatically calibrate factor weights for similarity measurement of geographic entities (Janowicz et al., 2010). The methods presented here are a form of semantic referencing.

The problem addressed in this paper (personalised similarity of places) is distinct from previous work on personalised web ranking, since we focus on the specific problem of finding similar *places*. Although individuals have different impressions of places, due to the nature of places being manifested in the physical world, there are broad regularities in the types of activities and features that people write about a given place. Because of these regularities, a crowdsourced representation of a place will have broad applicability. Thus, the search for places is not an information retrieval task on an individual document or artefact but rather a similarity search for places, represented as aggregations derived from multiple documents. Personalised place similarity remains a relatively unexplored research area outside of smaller-scale investigations in tourism and sense-of-place research (Moore and Graefe, 1994; Kaltenborn, 1998; Jorgensen and Stedman, 2006) and sociological investigations of place attachment (Stedman, 2002).

As with prior user behaviour for web personalisation, examples of similar places provide a mechanism to infer the interests of the user without requiring him or her to explicitly enter interests. Furthermore, the sample ranking of similar places – and which is provided directly by the user in this evaluation – can also be gathered from social data.

2.4 Topic models

Probabilistic topic models, such as latent Dirichlet allocation (LDA), provide methods to characterise the documents in a natural language corpus as multinomial mixtures of topics (Blei et al., 2003). Each topic is further modelled as a discrete probability distribution over all the words in a corpus. The top most probable words for a topic are usually semantically interpretable by humans. E.g., a topic with the top terms "wine, red, drink, cheese" can be interpreted as being about the activity of wine tasting (Adams and McKenzie, 2013).

LDA models the generation of each document in the corpus as an iterative process; first, a topic is chosen based on the topic distribution for the document; second, a word is randomly selected from that topic. This is repeated for each term in the document. Since each word is drawn randomly and is not based on the previous word, word order makes no difference in the LDA model. Assuming this generative process for how the documents were created, the words of an existing set of documents become the observed variables in the model and Bayesian inference is used to infer the most-likely topics to have generated the data. This inference can be performed programmatically in an approximate manner using a markov chain Monte Carlo Gibbs sampling algorithm (Griffiths and Steyvers, 2004).

Apart from discovering the latent topics in a corpus in an unsupervised manner, this dimensionality reduction allows one to compare the thematic similarity of two documents without requiring that the documents share exact terms. A natural way to measure the similarity of documents is the calculate the Kullback-Leibler (KL) divergence between the two topic distributions, P and Q (see Equation 1) (Steyvers and Griffiths, 2007).

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i). \tag{1}$$

Since KL divergence is asymmetric, the Jensen Shannon (JS) divergence can be used if a symmetric measure is desired (see Equation 2).

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M), \tag{2}$$

where $M = (P + Q)/2$. These methods of measuring similarity ignore that the importance of individual topics will vary from user to user.

Several extensions to LDA have been developed to characterise the mixture of topics associated with a location or place (Wang et al., 2007; Eisenstein et al., 2010; Hao et al., 2010; Sizov, 2010; Yin et al., 2011; Hong et al., 2012). In these models the generative process for creating a document is extended and some additional form of geographic evidence (either location or place name) is used to condition topics based on that evidence. Thus, a distribution of topics can be associated not only with a document but also a place or location. The technique employed here for evaluation is to represent the topic mixture for a place as a weighted average of the topic probability distributions of all the descriptions for the same place (Adams and Janowicz, 2012). Since the methods described in this paper for weighting topics are post-hoc operations (i.e. assigned after the topic model inferencing), they are broadly applicable to these other flavours of topic models.

3 Identifying salient topics

In this section, we present a method for personalising the weights on the LDA topics, given a source place s and a set of N target places $\{t_1, t_2, \ldots, t_N\}$, and a user specified ordering for those target places based on their similarity to the source place. For each place there is a discrete probability distribution for topics, and for any given individual topic the strengths of that topic for both source and target places can be compared. For example, let the following be a sample user ranking of three cities in terms of similarity to *New York City*:

1. Chicago
2. Los Angeles
3. Houston

Table 1 shows sample topic strengths for each of these cities (source and targets).

City	topic 1	topic 2	topic 3
New York City	0.2	0.6	0.2
Chicago	0.2	0.2	0.6
Los Angeles	0.42	0.38	0.2
Houston	0.8	0.1	0.1

Table 1: Sample topic strengths for source city New York and three target cities.

The set of salient topics Z_S is defined as a subset of all topics, such that $z \in Z_S$ if and only if the Kendall's τ rank correlation coefficient between the user-provided ordering and the topic strengths for topic z is positive (Kendall, 1938). The Kendall's τ measures rank correlation as a function of the number of concordant and discordant pairs in a set of joint observations. A pair of observations (i, j) of two variables (x, y) is concordant if $x_i < y_i \wedge x_j < y_j$ or $x_i > y_i \wedge x_j > y_j$. They are discordant if $x_i < y_i \wedge x_j > y_j$ or $x_i > y_i \wedge x_j < y_j$. Letting C be the number of concordant pairs and D be the number of discordant pairs in a sample and n be the number of observations, Kendall's τ is defined in Equation 3.

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}. \tag{3}$$

Thus, the relative weight (π_i) for each topic is zero for negative correlations. For positive correlations the weight is the correlation normalised such that all positive correlations sum to 1. See Table 2 for an example.

Kendall's τ-B is an extension to Kendall's τ to deal with situations where there are tied rankings. This will occur with topics when one or more documents have a value of 0.0 for the topic strength. When a model is trained using a large

user ordering	topic 1	topic 2	topic 3
Chicago	Chicago (0.0)	Los Angeles (0.22)	Los Angeles (0.0)
Los Angeles	Los Angeles (0.22)	Chicago (0.4)	Houston (0.1)
Houston	Houston (0.6)	Houston (0.5)	Chicago (0.4)
Kendall's τ	1.0	0.33	-0.67
Relative weight	0.75	0.25	0.0

Table 2: Rank ordering of target cities based on topic strength similarity. The difference between topic strengths is shown in parentheses. For each topic the Kendall's τ rank correlation is shown between the user similarity judgement and topic strength difference.

number of topics this will occur often. Kendall's τ-B (Equation 4) handles tied values by changing the denominator of the equation.

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \tag{4}$$

where $n_0 = n(n-1)/2$; $n_1 = \sum_i t_i(t_i - 1)/2$; $n_2 = \sum_j u_j(u_j - 1)/2$; n_c = number of concordant pairs; n_d = number of discordant pairs; t_i = number of tied values in the i^{th} group of ties for the first quantity; and u_j = number of tied values in the j^{th} group of ties for the second quantity.

3.1 Using Kendall's τ as product

For each salient topic, z_i, an associated salience weight, w_i, is assigned to the topic equal to the Kendall's τ correlation. Given a set of salience weights $w_1, w_2, \ldots w_n$ we can define relative weights for each topic $\pi_1, \pi_2, \ldots \pi_n$, where $\pi_i = \frac{w_i}{\sum_1^n w}$. Let π be this vector of relative weights, a *weighted Kullback-Liebler divergence A* is now defined in Equation 5.

$$D_{\mathrm{KL}}(\pi; P \parallel Q) = \sum_i \pi_i P(i) \ln \frac{P(i)}{Q(i)}. \tag{5}$$

The divergence is weighted not only on the probabilities of the topic variables but also with the salience weight based on the ordering, and it is thus a weighted average of the logarithm difference between the probabilities. From this new divergence function, the *weighted Jensen Shannon divergence A* is defined in Equation 6.

$$JSD(\pi; P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(\pi; P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(\pi; Q \parallel M), \tag{6}$$

where M is defined as in the normal JS divergence.

This weighted measure can be interpreted as a form of weighted sum model in a multi-criteria decision analysis.The salience weights drawn from a sample ranking can be viewed as a personalised general measure of topic saliency for comparing all places. Table 3 shows a comparison of Jensen Shannon divergence and weighted Jensen Shannon divergence A calculations given the user ranking specified above. Note, in this example the weighted ordering matches the user specified ordering; however, it is not guaranteed to match in all cases, since the salience weights are based on a comparison of rankings, not actual similarity values of the probabilities.

3.2 Using Kendall's τ to alter topic weights

An alternative method re-weights the topic probabilities associated with each place and uses the traditional JS divergence measure to obtain a context-dependent similarity. As with the previous method only positive τ correlations are used. Letting $\mathbf{z}$ be the topic vector for a place, Algorithm 1 describes the steps for re-weighting topics. Each topic probability, z_i, is multiplied by the weight for that topic, π_i, and then the new topic vector is normalized to sum to 1 (see Algorithm 2).

Algorithm 1 Pseudocode for re-weighting topics from user provided ranking.

```
sum ← 0.0
for all z_i do
  if w_i > 0 then
    z'_i ← w_i * z_i
    sum ← sum + z'_i
  else
    z'_i ← 0.0
  end if
end for
for all z'_i do
  z'_i ← z'_i / sum
end for
```

This new topic distribution for a place has a very intuitive interpretation. The re-weighted probability of topic i, z'_i, is a function of the prototypical probability of the topic, z_i, for the place, conditioned on the probability that this specific user is concordant with the average user for this topic. This re-weighted distribution has the advantage that it maintains the desired mathematical properties when comparing similarities using KL and JS divergences (i.e., a minimum divergence of

0 and the square root of the JS divergence is a metric). In Table 3 the Jensen Shannon result using this method is called the *weighted Jensen Shannon B* measure.

city	JS	weighted JS B
Chicago	0.151	0.049
Los Angeles	0.047	0.057
Houston	0.294	0.220

Table 3: Comparison of JS divergence and weighted JS divergence values from New York City given a user ranking of *1. Chicago, 2. Los Angeles, 3. Houston* (lower values indicate more similar). Note, that JS and weighted JS values are not comparable, only ranks.

The method described above works well in cases where the user rankings provide information to the system for all the topics. However, when using a system trained on a large number of topics (e.g., several hundred) there can be individual topics where all of the target places have a weight of 0.0. This means that all the places being ranked are equally similar to the source place in terms of that single topic.

In this case, we only want the topics that are informed by the ranking to affect each other while keeping the probability of all other topics fixed. Let Z be the set of all topics and $Z_{NI} \subseteq Z$ be the set of all topics for which the ranking does not provide any information. The final version of the algorithm is shown in Algorithm 2.

4 Evaluation

In this section, we present the results of a human participants test to evaluate the efficacy of using sample rankings to prime salience weights on individual topic dimensions. LDA topic models were trained on two corpora of georeferenced place descriptions: 1) a set of 200,000 georeferenced Wikipedia articles and 2) a set of 275,000 travel blog entries.[1] Prior to performing topic modelling, standard stemming and cleaning of the documents was performed to remove html tags and other noise in the text. The travel blog entries were matched to places by the authors according to a fixed geographic hierarchy, e.g., *Orlando, Florida, United States*. Each georeferenced Wikipedia article was matched to a named place by intersecting the location associated with the article (based on coordinates template in the article) with the spatial footprint of the named places. Thus, each article is linked to one place. It is possible that the association between text and a place can be refined further using more sophisticated techniques; however, this still remains a open research problem and not the focus of the current work (Vasardani et al., 2013).

The topic signature for a place was calculated by taking all the documents with a relation to the place and averaging over their topic vectors. Thus, topics that have a relatively high probability in a large number of documents associated with a place will be associated with the place (e.g., in travel blog entries a *theme park* topic will have high value for Orlando, FL, but a *skiing* topic will not).

4.1 Design

The user study was designed using the Amazon Mechanical Turk system to gather several place similarity rankings.[2] A qualification test was presented to the Mechanical Turk users to help ensure high-quality (non-spam) results (Kittur et al., 2008). This test was a simple filter that gathers user information and asks a set of simple questions about geographic knowledge. The primary goal of the survey was to determine the degree to which a user-provided ranking on a set of cities can be used to inform the system on the "interests" of the user more generally, and thus tailor search results.

The study focused on comparing 30 U.S. cities rather than other types of places in order to maximise participant familiarity. Although initial variants of the test were done using cities from around the world, it was difficult to find users who had broad familiarity with multiple cities from around the world. Therefore, the tests were limited to residents of the U.S., and a good sample of participants with familiarity of target cities could be gathered. Table 4 shows all the cities that were compared.

Mechanical Turk participants were asked to rank order 7 cities in terms of similarity to another city. Users were also asked to enter how the places are similar as well as describe their familiarity with each of the 8 cities. Table 5 shows two sample responses for cities that are ranked in similarity to Los Angeles.

In total 96 participants provided 5 rankings each, leading to 480 rankings total. Out of the 480 rankings, in 135 cases (28.1%) the participant was 'not' familiar with the main city being compared, in 205 cases (42.7%) the participant was

[1] downloaded from `http://www.travelblog.org`

[2] `http://www.mturk.com`

Algorithm 2 Pseudocode for re-weighting topics from user provided ranking when not all topics are informed by the ranking.

```
sum ← 0.0
sumOthers ← 0.0
for all z_i do
  if z_i ∈ Z_NI then
    z'_i ← z_i
    sumOthers ← sumOthers + z_i
  else if w_i > 0 then
    z'_i ← w_i * z_i
    sum ← sum + z'_i
  else
    z'_i ← 0.0
  end if
end for
for all z'_i do
  if z'_i ∉ Z_NI then
    z'_i ← z'_i * (1− sumOthers )/ sum
  end if
end for
```

Atlanta	Chicago	Houston	Minn.-St. Paul	Phoenix	San Francisco
Austin	Cleveland	Kansas City	Nashville	Pittsburgh	San Jose
Baltimore	Dallas	Las Vegas	New Orleans	Portland, OR	Seattle
Boston	Denver	Los Angeles	New York City	Salt Lake City	St. Louis
Charlotte	Detroit	Miami	Philadelphia	San Diego	Washington D.C.

Table 4: United States cities that were ranked based on similarity. For each comparison 7 of these cities were ranked based on similarity to 1 other city.

'somewhat' familiar, and in 140 cases (29.2%) the participant was 'very' familiar. In 50 cases (10.4%) the participant had lived in (or near) the city and in 290 cases (60.4%) the participant had visited the city.

4.2 Spearman's footrule distance

Spearman's footrule distance is a measure of correspondence (or disarray) between two rankings, similar to Spearman's ρ and Kendall's τ. Let S_n be the set of all permutations of the set of n integers $\{1, \ldots, n\}$. From (Diaconis and Graham, 1977), the Spearman's footrule distance, D_n is defined as in Equation 7.

$$D_n(\pi_n, \sigma_n) = \sum_{i=1}^{n} |\sigma_n(i) - \pi_n(i)|\,, \tag{7}$$

where π_n and σ_n are elements of S_n, i.e., different permutations. Spearman's footrule is used here rather than Spearman's rho, because we want to be able to consider the average footrule distance between several automated rankings and user provided rankings, which is not valid for correlations.

4.3 Average Spearman's footrule results

In Figure 1 a chart is shown of the average Spearman's footrule distance between the human participant rankings and automated similarity calculations. Looking over all responses, the clearest outcome was that for source places with which the participant is not familiar; generally, a participant's assessment is primarily made based on distance rather than other properties. When the participant is somewhat or very familiar with the source place there is an increase in the average Spearman's footrule distance between distance-based rankings.

Looking at Figure 1 it is abundantly clear that similarity ranking based on all travel blog topics or all Wikipedia articles does not correspond in any definitive way with user-provided rankings. In part this is because there is no "average person" – there is very little correspondence between different participants about the appropriate ranking of similar cities. In

Similar to Los Angeles ranking 1	
Miami	active nightlife
Minneapolis-St. Paul	skyscrapers
Dallas	home to corporations
Austin	hot weather
Salt Lake City	big houses
Cleveland	busy
Portland, OR	tourist towns

Similar to Los Angeles ranking 2	
Miami	beautiful people; emphasis on appearance
Dallas	very spread out
Austin	music and performing arts
Portland, OR	music and performing arts
Minneapolis-St. Paul	music and performing arts
Salt Lake City	western US
Cleveland	both are cities with large areas in economic decline

Table 5: Sample rankings of similarity to Los Angeles from two different participants.

addition, individuals have a limited set of properties that they use to compare places as opposed to the large number of different topics found in large text corpora. This is not a problem with the methodology, however, because the task that the system is to perform is not necessarily to match individual human reasoning but rather to allow exposure of patterns and similarities in place description topics that are otherwise opaque to an individual user (or human participant).

4.4 Automatic topic weighting evaluation

Despite this lack of correspondence, we can use the participant studies to evaluate the degree to which it is valid to assume a user has a set of 'interests' (i.e., a salient subset of topics) that are common across different place comparisons and whether we can identify those interests in terms of topics automatically from one or more sample rankings. For example, looking back at Table 5, these two participants consider very different properties when comparing the cities for similarity. Note, however that these participant-provided properties do not necessarily reflect all the properties used by the participant to discriminate between places – e.g., it does not explain the particular ordering of Austin, Portland, and Minneapolis-St. Paul by participant 2, because the same property ("music and performing arts") is given for all of them.

In order to do this evaluation we test whether on average automatic topic re-weighting using the techniques in the previous section result in smaller Spearman's footrule distances for the other rankings by the same participant. In other words, we evaluate the degree to which a single ranking can be extrapolated more generally to other rankings of cities. One sample ranking from the participant is used to calculate weights on the topics and new city rankings are calculated using the weighted JS divergence for each of the other city comparisons done by the same participant. For each ranking the Spearman's footrule distance is calculated between the new weighted ranking and the participant-provided ranking. This process is then repeated for each sample ranking by exchanging which one sets the weights. From this cross-validation an average footrule distance can be calculated for the weighted ranking for a participant, and this can be further averaged over all participants. Figure 1 shows this average footrule distance over all participants when using the weighted JS divergence B technique. The weighted JS divergence A has a similar result, though the overall average footrule distance tends to be slightly higher.

The average Spearman's footrule distance shows a marked decrease over the unweighted results for Wikipedia and travel blogs that were shown in Figure 1, which indicates that this method does help to identify general topics of interest for a person. One very interesting result is that the matching to the participant rankings is distinctly *better* when the participant is *not* very familiar with the source city being ranked against. One explanation for this result is that people who are very familiar with a place will have very specific impressions of the place and thus properties that are salient for that place do not transfer over from other places. Whereas when people are not familiar with a place there is a stock set of "background interests" that they use to help compare places. This indicates that automatic topic weighting will be particularly useful when done as part of a system designed to enable users to learn and explore knowledge about geographic places with which they are not already very familiar.

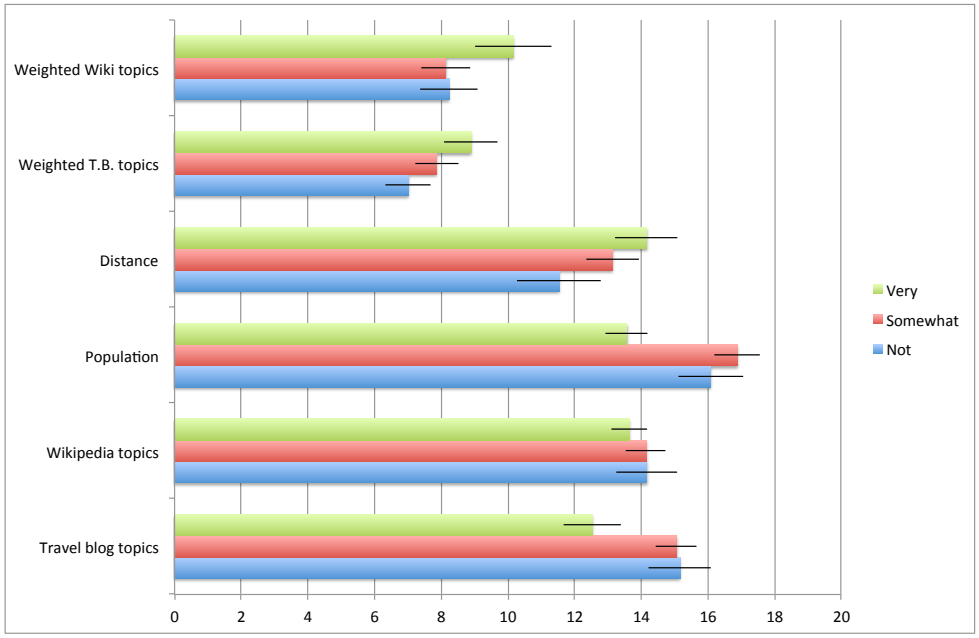

Figure 1: This chart shows the average Spearman's footrule distance between participants' rankings of U.S. city similarities and rankings based on geographic distance, population, Wikipedia topic mixtures, travel blog topic mixtures, and weighted Wikipedia and travel blog topics using weighted JS divergence B. The x-axis is the average Spearman's foot rule distance, indicating total displacement in the orderings of places being compared (Kumar and Vassilvitskii, 2010). Shorter bars indicate overall higher concordance between the participants' similar place rankings and the automated similar place rankings. Each type of automated ranking is split into three categories based on whether the participant is "very" (green), "somewhat" (red), or "not" (blue) familiar with the source place being compared with other places (based on self-report). The statistical significance bar shows $\alpha = 0.05$.

4.5 Discussion

From these results, the automatic weighting of topics based on a sample ranking can legitimately be generalised to other rankings. An interesting alternative would be to extract interests of a user implicitly from social network or other data and apply those to topic weights, but that remains beyond the scope of the current research (Kelly and Teevan, 2003; Ricci et al., 2011). On a case-by-case basis we also examined whether the reasons participants give for similarity correspond to high shared values on specific LDA topics, but apart from anecdotal evidence, it was difficult to do a quantitative evaluation on these data, because the mapping of participant-provided reasons and topics is highly subjective.

Table 6 shows the topics in Wikipedia (from 1200 topics) and travel blogs (from 1500 topics) for *Los Angeles* that increase the most (by ratio) based on re-weightings from the two sample participant rankings shown in Table 5. What is remarkable about the topics shown in Table 6 is that despite the apparent difference between the automatically-determined most-salient topics and the participant-provided reasons, the automatic weighting does significantly increase the concordance between the system rankings and the participant rankings. One explanation is that the topics that are weighted with high salience represent background properties that factor in the participants' conceptualisations of city categories but which are not at the forefront of their conscious comparisons of individual cities.

5 Conclusion

People judge the similarity or difference of places based on different contextual factors, such as their personal interests. Geographic information systems that take advantage of these factors and can provide places that are similar to places known to a user have many potential applications, including travel recommendation services, marketing analysis tools, and socio-ecological research tools. In this paper we presented a new method to automatically identify the topics that are salient to a user when performing similarity judgment. Topics can be any set of features associated with a place, such that a place is represented as a probability vector of topic values. These topic values form a topic signature that is generally associated with a place, and we demonstrated how probabilistic topic modelling can be used to generate such probability vectors for places.

Similar to Los Angeles ranking 1 – Travel blog	
Topic 827	game,basebal,team,play,watch,sport,hockey,fan,stadium,player
Topic 1242	restaur,order,food,tabl,meal,menu,eat,waiter,serv,dinner
Topic 904	money,pay,cost,expens,price,onli,cheap,pound,buy,free
Similar to Los Angeles ranking 1 – Wikipedia	
Topic 761	book,work,publish,life,wrote,histori,year,mani,author,writer
Topic 834	cathol,bishop,dioces,roman_cathol,roman,john,father,priest
Topic 890	golf,cour,club,golf_club,hole,countri_club,countri,locat,golf_cours
Similar to Los Angeles ranking 2 – Travel blog	
Topic 909	mile,road,stop,highway,gas,motel,state,sign,rout,trip
Topic 793	extrem,complet,entir,exact,time,veri,actual,howev,made,ani
Topic 1228	histor,site,build,visit,histori,tour,histor_site,area,museum,mani
Similar to Los Angeles ranking 2 – Wikipedia	
Topic 417	side,east,east_side,west,north,south,locat,north_side,west_side
Topic 164	polic,offic,polic_offic,polic_station,polic_depart,depart,enforc,forc
Topic 847	turkish,ottoman,turkey,greek,byzantin,ottoman_empir,sultan

Table 6: Topics that increase most in salience for Los Angeles based on two different participants' rankings.

Similarity calculations based on probability distributions are commonly context-neutral and only measure the relative entropy or JS divergence of the distributions. We presented a novel approach to use a small, user-provided sample ranking of similar places to automatically re-weight topic weights. This new re-weighting can be used to serve up personalised similarity values between places based on the topics that are of most interest to a user. Since this re-weighting of topic values is done after the initial training used to discover topics associated with places, it is fully compatible with a variety of different methods to create semantic signatures associated with a place, not solely topic modelling as used here.

We evaluated our method with a Mechanical Turk user study and showed that using a small sample ranking of similar places (i.e., a set of control similarities) results in a larger correspondence between automated place similarity rankings and personal users' rankings. Personalised place recommendation and similarity search remains a relatively unexplored research area. A follow-up study to better understand user motivation in performing place similarity will be a valuable next step. Future work will also involve exploring social media data and other sources such as search history to automatically provide sample similar places for a given user.

References

Adams, B. and K. Janowicz (2012). On the geo-indicativeness of non-georeferenced text. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci (Eds.), *ICWSM*, pp. 375–378. The AAAI Press.

Adams, B. and G. McKenzie (2013). Inferring thematic places from spatially referenced natural language descriptions. In D. Sui, S. Elwood, and M. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge*, pp. 201–221. Springer Netherlands.

Banerjee, S. S. and R. R. Dholakia (2008). Mobile advertising: does location-based advertising work? *International Journal of Mobile Marketing 3*(2), 68–75.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bodenhamer, D. J. (2010). The potential of spatial humanities. In D. J. Bodenhamer, J. Corrigan, and T. M. Harris (Eds.), *The Spatial Humanities: GIS and the Future of Humanities Scholarship*, pp. 14–30. Indiana University Press.

Bouidghaghen, O., L. Tamine, and M. Boughanem (2011). Personalizing mobile web search for location sensitive queries. In A. B. Zaslavsky, P. K. Chrysanthis, D. L. Lee, D. Chakraborty, V. Kalogeraki, M. F. Mokbel, and C.-Y. Chow (Eds.), *Mobile Data Management (1)*, pp. 110–118. IEEE.

Daoud, M., L. Tamine, M. Boughanem, and B. Chebaro (2007). Learning implicit user interests using ontology and search history for personalization. In *Proceedings of the 2007 international conference on Web information systems engineering*, WISE'07, Berlin, Heidelberg, pp. 325–336. Springer-Verlag.

Diaconis, P. and R. L. Graham (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(2), 262–268.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, Stroudsburg, PA, USA, pp. 1277–1287. Association for Computational Linguistics.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. A Bradford Book. MIT Press.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*(Suppl. 1), 5228–5235.

Grossner, K. E., M. F. Goodchild, and K. C. Clarke (2008). Defining a digital earth system. *Transactions in GIS 12*(1), 145–160.

Guy, I., M. Jacovi, A. Perer, I. Ronen, and E. Uziel (2010). Same places, same things, same people?: mining user similarity on social media. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, New York, NY, USA, pp. 41–50. ACM.

Hao, Q., R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Z. 0001 (2010). Equip tourists with knowledge mined from travelogues. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti (Eds.), *WWW*, pp. 401–410. ACM.

Hong, L., A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis (2012). Discovering geographical topics in the twitter stream. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab (Eds.), *WWW*, pp. 769–778. ACM.

Janowicz, K., B. Adams, and M. Raubal (2010). Semantic referencing - determining context weights for similarity measurement. In S. I. Fabrikant, T. Reichenbacher, M. J. van Kreveld, and C. Schlieder (Eds.), *GIScience*, Volume 6292 of *Lecture Notes in Computer Science*, pp. 70–84. Springer.

Janowicz, K., M. Raubal, and W. Kuhn (2011). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science* (2), 29–57.

Jones, C. B. and R. S. Purves (2008). Geographical information retrieval. *International Journal of Geographical Information Science 22*(3), 219–228.

Jordan, T., M. Raubal, B. Gartrell, and M. J. Egenhofer (1998). An affordance-based model of place in GIS. In T. Poiker and N. Chrisman (Eds.), *Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98)*, Vancouver, Canada, pp. 98–109.

Jorgensen, B. S. and R. C. Stedman (2006). A comparative analysis of predictors of sense of place dimensions: Attachment to, dependence on, and identification with lakeshore properties. *Journal of Environmental Management 79*, 316–327.

Kaltenborn, B. P. (1998). Effects of sense of place on responses to environmental impacts: A study among residents in svalbard in the Norwegian high Arctic. *Applied Geography 2*(18), 169–189.

Kelly, D. and J. Teevan (2003). Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum 37*, 18–28.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*(1/2), pp. 81–93.

Keßler, C. (2012). What is the difference? a cognitive dissimilarity measure for information retrieval result sets. *Knowl. Inf. Syst. 30*(2), 319–340.

Kittur, A., E. H. Chi, and B. Suh (2008). Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, New York, NY, USA, pp. 453–456. ACM.

Kumar, R. and S. Vassilvitskii (2010). Generalized distances between rankings. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti (Eds.), *WWW*, pp. 571–580. ACM.

Larson, R. (1996). Geographic information retrieval and spatial browsing. In L. C. Smith and M. Gluck (Eds.), *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, pp. 81–124.

McKenzie, G., B. Adams, and K. Janowicz (2013). A thematic approach to user similarity built on geosocial check-ins. In D. Vandenbroucke, B. Bucher, and J. Crompvoets (Eds.), *Geographic Information Science at the Heart of Europe*, Lecture Notes in Geoinformation and Cartography, pp. 39–53. Springer International Publishing.

Medin, D. L., R. L. Goldstone, and D. Gentner (1993). Respects for similarity. *Psychological Review 100*(2), 254–278.

Micarelli, A., F. Gasparetti, F. Sciarrone, and S. Gauch (2007). The adaptive web. Chapter Personalized search on the world wide web, pp. 195–230. Berlin, Heidelberg: Springer-Verlag.

Moore, R. L. and A. R. Graefe (1994). Attachments to recreation settings: The case of rail-trail users. *Leisure Sciences 16*, 17–31.

Pitkow, J., H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel (2002). Personalized search. *Commun. ACM 45*(9), 50–55.

Ranganathan, A. and R. H. Campbell (2002). Advertising in a pervasive computing environment. In M. S. Viveros, H. Lei, and O. Wolfson (Eds.), *Workshop Mobile Commerce*, pp. 10–14. ACM.

Raubal, M. (2004). Formalizing conceptual spaces. In A. C. Varzi and L. Vieu (Eds.), *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, pp. 153–164. IOS Press.

Relph, E. (1976). *Place and Placelessness*. Pion.

Ricci, F., L. Rokach, and B. Shapira (2011). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Eds.), *Recommender Systems Handbook*, pp. 1–35. Springer.

Rodríguez, M. A. and M. J. Egenhofer (2004). Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science 18*(3), 229–256.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and Categorization*, pp. 27–48. Hillsdale (NJ), USA: Lawrence Erlbaum Associates.

Sizov, S. (2010). Geofolk: latent spatial semantics in web 2.0 social media. In B. D. Davison, T. Suel, N. Craswell, and B. Liu (Eds.), *WSDM*, pp. 281–290. ACM.

Speretta, M. and S. Gauch (2005). Personalized search based on user search histories. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pp. 622–628.

Stedman, R. C. (2002). Toward a social psychology of place : Predicting behavior from place-based cognitions, attitude, and identity. *Environment and Behavior 34*(5), 561–581.

Steyvers, M. and T. Griffiths (2007). Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

Teevan, J., S. T. Dumais, and E. Horvitz (2005). Personalizing search via automated analysis of interests and activities. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp. 449–456. ACM.

Tuan, Y.-F. (1977). *Space and Place: the Perspective of Experience*. The Regents of the University of Minnesota.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*(4), 327–352.

Vasardani, M., S. Winter, and K.-F. Richter (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science 27*(12), 2509–2532.

Wang, C., J. Wang, X. Xie, and W.-Y. Ma (2007). Mining geographic knowledge using location aware topic model. In R. Purves and C. Jones (Eds.), *GIR*, pp. 65–70. ACM.

Winter, S., W. Kuhn, and A. Krüger (2009). Guest editorial: Does place have a place in geographic information science? *Spatial Cognition and Computation 9*, 171–173.

Yin, Z., L. Cao, J. Han, C. Zhai, and T. S. Huang (2011). Geographical topic discovery and comparison. In S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar (Eds.), *WWW*, pp. 247–256. ACM.