Integration of folksonomies into the process of map generalization

Meysam Aliakbarian
University of Zurich
Winterthurerstrasse 190
Zurich, Switzerland
meysam.aliakbarian@geo.uzh.ch

Robert Weibel
University of Zurich
Winterthurerstrasse 190
Zurich, Switzerland
robert.weibel@geo.uzh.ch

Abstract

The growth of user-generated content in quantity and quality has changed the way people use digital services, including geo-services. The process of map generalization is not an exception to this phenomenon. Earlier research has considered user-generated content as *data* sources for the generalization process. However, little work has been accomplished to date considering the *knowledge* that may be extracted from those sources, in particular from special place-related semantics captured in user-contributed feature tags. This study considers doing so from the perspective of *folksonomies*, presenting some first steps in that direction. In particular, this short paper shows, on the example of OpenStreetMap, how different similarity measures can be used to exploit folksonomy-based semantics in map generalization. And it shows how these semantics can be used to introduce behaviour changes in generalization operators, in particular in the selection and aggregation operators, respectively.

Keywords: map generalization, generalization operators, user-generated content, folksonomy, OpenStreetMap

1 Introduction

The process of map making is traditionally based on standardized data collection, processing and visualisation of geographic data. In the traditional process data is collected by experts at official mapping agencies using field surveys and/or photogrammetry, with well-defined results and complete coverage, complying with given standards.

In the wave of the Web 2.0, ordinary passive content users of the internet have turned into active content producers for different content platforms. The term User-Generated Content (UGC) covers content produced in this way. The wave has also included production of geographic information. Different projects and platforms have started to gather implicit and explicit geographic information about objects and phenomena. Some examples are OpenStreetMap, Flickr and WikiMapia. A hidden gem of UGC datasets is the possibility of extraction of knowledge in different forms. A possible approach is to explore for taxonomies based on content production behaviour of UGC contributors. Vander Wal [8] coined the term *folksonomy* to reflect the fact that such taxonomies are based on people's perception and contributions. According to him folksonomies represent a bottom-up social classification.

The research reported here is investigating the integration of folksonomies into the process of map generalization in the form of modifications of atomic generalization operators. The concept is presented here for Points of Interest (POIs) but is thought to be extensible to the case of more complex input data. The motivation for integrating folksonomies into map generalization is to move toward maps that convey not only the *data* generated by users but also the *knowledge* that is engrained in user contributions, e.g. in the form of special semantics and tags that users attach to places captured in UGC, and which can be extracted from the UGC. The knowledge may then be used when making decisions during the map generalization.

In this short paper, we use OpenStreetMap (OSM) as an example of UGC, focusing primarily on the semantics that is captured in OSM feature tags, and how this information can be exploited in generalization. The contributions of this short paper are twofold: 1) We show how the user-contributed semantics contained in OSM feature tags can be exploited in folksonomy-based semantic similarity measures and how this can be used in the map generalization process. 2) For the selection and the aggregation operator, we show first examples of behaviour changes that can be introduced to generalization operators by folksonomy-based semantics.

2 Background

This section briefly reviews the background in four areas that are relevant to this research: folksonomies, semantic similarity, generalization, and OSM.

2.1 Folksonomies

Folksonomies are taxonomies formed by tagging behaviour of users. Folksonomy models generally consist of three components — *users*, *resources* and *tags* — where the users use the tags to describe the resources. A common notation is given in [2] as well as in Equation 1 (u, r and t represent *users*, *tags* and *resources*, respectively, where y represents the relation between the three).

$$F = f(u, r, t, y) \qquad (1)$$

Folksonomies are often seen as emergent semantics as they provide decentralized semantic structure [4]. Early examples of folksonomies were based on social bookmarking services [7]. Different research approaches have analyzed the triple combination of users, resources and tags. There have also been approaches that have taken two members of the three. A primary outcome is the ability to define concepts based on the tagging behaviour of users. Another outcome is the ability to measure the semantic similarity between pairs of the aforementioned entities (users, resources, tags).

2.2 Semantic Similarity

In order to measure the similarity (or relatedness) between concepts represented in folksonomies, semantic similarity measures are needed. Such investigation will help defining and measuring the quality of concepts being semantically close to or far from each other, respectively. Thus, similarity measures define similarity in terms of distance. Examples are the Jaccard, dice and cosine similarity measures [4] (cf. Equations 2, 3, 4). Generally, the level of commonality between two sides of similarity assessment is measured and normalized with the size of their attribute space.

$$\text{sim\_jaccard}\ (X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (2)$$

$$\text{sim\_dice}\ (X,Y) = \frac{2\ |X \cap Y|}{|X| + |Y|} \quad (3)$$

$$\text{sim\_cosine}\ (X,Y) = \frac{X\,.\,Y}{||X||\ ||Y||} \quad (4)$$

Using such measures, new relations can be shaped based on semantic similarities. These realtions will be helpful when informing the process of map generalization. An example is the case of generating a compound feature by aggregating smaller features based on their semantic similarities.

2.3 Generalization and Semantics

Considering the classic definition of generalization [3], the triggers to the procedure are scale and map purpose. While scale is thought of as a geometric trigger, map purpose is related with the information content and the user needs. Understanding the semantic structures behind geographic data helps us generating maps that are closer to user needs. While most research in map generalization to date has focused on topographic maps, where the semantics is typically highly standardized, maps relying on UGC will display a wider variety of semantics, as folksonomies are more heterogeneous than the ontologies defined for topographic data of official sources. Consequently, there is need to invest more in integrating semantics in the generalization procedure.

Efforts to control the complexity of modeling the generalization process have resulted in different approaches, of which constraint-based modeling represents the current state of the art [1]. In such a model there is need to control the procedure with appropriate contraints. In [6], a taxonomy of relations was proposed and a threefold relationship between *relations*, *contraints* and *measures* was described.

In order to properly integrate semantics into the generalization process, there is need to detect and define semantic relations based on geographic data. Here we aim to detect and utilize semantic measures derived from geographic UGC and will investigate their potential in the generalization process.

2.4 OSM

On the highest level of granularity OSM data is categorized into three element types: *nodes*, *ways* and *relations*. All three element types can be enriched with tags. Tags are textual key-value pairs that extend meanings of geographic features by attaching additional information to them (e.g. key=leisure and value=park to describe a park). OSM data comes with its own peculiarities. Besides typical accuracy concerns, consideration of feature attributes also comes with other concerns. Tags do not have a strict structure in the OSM project and generally reflect the particular view of users over mapped geographic features and could thus be seen from the perspective of folksonomies. Tagging strategies are negotiated and agreed upon in the OSM wiki. However, there is no single widely accepted regime of tagging. This results in a dataset in which entities have different sets of attributes. Considering a table as a classical way of storing data, keys could be seen as columns and features as rows. Typical data has many empty cells in such a form of data representation. Therefore, there is a need of adapting data procedures further to this phenomenon, including the development of similarity measures that can cope with missing tag values in UGC such as OSM.

3 Methodology

The present study aims to fill the above research gaps. Already mentioned measures of semantic similarity will be calculated from a folksonomy perspective and then utilized in modified generalization operators (selection and aggregation).

3.1 Semantic similarity based on folksonomy

From the triple aspects of folksonomies (users, tags, resources), here we consider tags and resources. Taking OSM as the focus of the study, each feature in OSM has a unique URL in the form of *www.osm.org/[node|way|relation]//[OSM-ID]*. Thus, such feature qualifies as a resource. As mentioned before tags are given in the form of keys and values. We will use the semantic similarity measures introduced earlier and also propose a new method, which measures feature-to-feature (or resource-to-resource) similarity.

3.1.1 Preprocessing

In order to apply the feature-to-feature similarity analysis, there is need to preprocess the data. The main step iterates over the features and removes tags that express spatial information and retain only tags that exclusively express semantic information. The logic behind this step is that the geometry of features contains enough spatial information and thus tags expressing spatial information are not needed and should be removed if the focus should be on semantic analysis. Example of such spatial tags are *postal_code*, *addr:street* and *addr:housenumber.*

3.1.2 Processing

By having appropriate tags for each feature we are ready to calculate measures. In order to calculate the Jaccard and dice measures, having tags of two sides of similarity measurement is enough and it is feasible to calculate the length of both sides, their intersection, and their union. As cosine similarity is defined in a vector space, there is a need to know the dimensions of the space. This is given by getting the union of tags of all features. Cosine similarity is also a frequency-based measure. In the case of OSM, it is impossible to measure the frequency of terms (tags) for each feature as they cannot have repetitive tags. If a feature has a certain tag, it gets assigned 1 for that dimension and if not, 0 would be assigned. An example of the described process is given below.

F1 : {amenity=parking, name= ExCel, fee= yes}
F2 : {amenity=parking, name= Fox@Connaught, fee=no, access= customer}
F3 : {amenity=fast_food, name= McDonald's, wheelchair= yes, cuisine= burger}

$$\text{sim\_jaccard (F1, F2)} = \frac{|\mathrm{F1} \cap \mathrm{F2}|}{|\mathrm{F1} \cup \mathrm{F2}|} = \frac{|\{amenity, name, fee\}|}{|\{amenity, name, fee, access\}|} = \frac{3}{4} = 0.75$$

$$\text{sim\_dice (F1, F2)} = \frac{2\,|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|} = \frac{2\,|\{amenity, name, fee\}|}{|\{amenity, name, fee\}| + |\{amenity, name, fee, access\}|} = \frac{2 * 3}{3 + 4} = \frac{6}{7} = 0.857$$

$$\text{VectorSpace} = [\text{amenity, name, fee, access, wheelchair, cuisine}]$$
$$\text{VectorF1} = [1, 1, 1, 0, 0, 0]$$
$$\text{VectorF2} = [1, 1, 1, 1, 0, 0]$$
$$\text{Sim\_cosine (F1, F2)} = 0.866$$

The above similarity measures are based on set-theory intersections or term frequencies, but if we have *key-value* pairs it is important to include *values* in the similarity measurement. Measuring similarity solely based on keys results in a misconception of feature matches. In order to overcome such problem a new measure is proposed here, *KeyValue similarity*, which can be seen as an extension of dice similarity:

$$\mathrm{Sim}_{\mathrm{KeyValue}}(\mathrm{X}, \mathrm{Y}) = \frac{\left(\frac{2|K_X \,\cap\, K_Y|}{|X| + |Y|} + \frac{|V_X \,\cap\, V_Y|}{|K_X \,\cap\, K_Y|}\right)}{2} \qquad (5)$$

where K_X represent keys of X, V_X represent values of X. This equation normalizes the features' commonalities. While shared keys are normalized with the length of feature tags, shared values of those keys are normalized with the number of shared keys. Taking the above example, the similarity of the mentioned features is as follows:

$$\mathrm{Sim}_{\mathrm{KeyValue}}(\mathrm{F1}, \mathrm{F2}) = \frac{\left(\frac{2 * 3}{3 + 4} + \frac{1}{3}\right)}{2} = 0.428$$

The proposed measure generally yields lower values than the other aforementioned measures and reports a value of 1.0 only if all of the keys and values are equal.

3.2 Semantic measures in map generalization

In the process of map generalization based on geographic datasets, there is room for better integration of semantics. This integration can happen in data modelling, process modelling and in the operators themselves (cf. [1]). Regarding operators, the model may include measures that are based on semantics derived from UGC tags, and thus folksonomy-based. Here we study modifying the input of two generalization operators: selection and aggregation.

Our initial hypothesis is to apply two thresholds to similarity measurements: a lower threshold α and an upper threshold β, where a value of less than α would be taken as *dissimilarity*, a value between α and β as *similarity*, and a value greater than β would be taken as a candidate for *equality* (i.e. two equal features). In other words:

$$s = Sim\,(X, Y)\colon \begin{cases} s < \alpha \rightarrow \text{dissimilar} \\ \alpha \leq s \leq \beta \rightarrow \text{similar} \\ s > \beta \rightarrow \text{test if X = Y} \end{cases} \qquad (6)$$

3.2.1 Selection

The motivation for applying selection is to reduce the number of features on a map, thus aiming for less data or less visual clutter. The filter criteria may be spatial (e.g. overlap or congestion) or semantic, such as an importance ranking or classification function deciding to retain or remove features.

Using Eq. 6, *similarity-based selection* as defined here will take a feature as a representative (or search feature, or exemplar) and decide about the *similar* and *dissimilar* features. Two approaches are possible: retain the *similars* and eliminate the *dissimilars*, or take the feature as the representative and remove the *similars* as they are already represented. Both options are shown schematically in Figure 1.

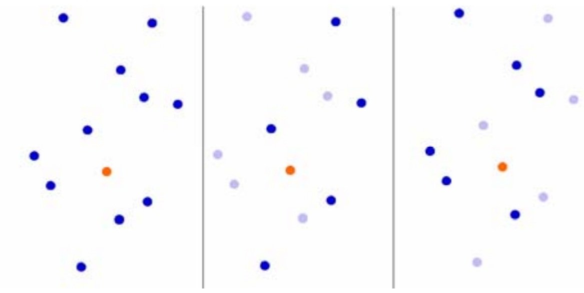

Figure 1 - Schematic illustration of *similarity-based selection*. Left: initial situation (search feature shown in orange, selected features shown in dark blue, filtered features in light blue). Middle: the selection process has selected the semantically similar features and eliminated the dissimilars. Right: the selection process has selected the semantically dissimilar features and eliminated the similars.

3.2.2 Aggregation

Aggregation is the operator that merges features in order to decrease the number of features or to decrease the detail of rendered features. A group of features might be aggregated if they are close enough to each other and have enough similarity to be taken as one feature. This is well projectable to the semantic perspective where there is the possibility of measuring similarity between the features. Considering Eq. 6, the *similarity-based aggregation* will take a feature as aggregation candidate and *similar* features will be merged to that feature while not touching the *dissimilar* features. A schematic example is given in Figure 2. The resulting feature of this aggregation can be placed on the anchor point (search feature), the centroid of the collection of features, or as a minimum convex polygon. An important constraint is to limit the process of finding candidates within a meaningful radius, as aggregation of very far features is not meaningful.

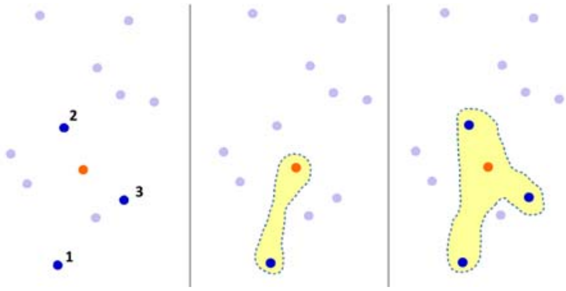

Figure 2 - Schematic illustration of *similarity-based aggregation*. Left: initial situation (with top 3 most similar features). Middle: aggregating the search feature with the most similar feature. Right: aggregating the search feature with the top 3 most similar features.

4 Examples

In a scenario of visualizing points of interest (POIs) in a location-based service (LBS), a situation like Figure 3 may happen. Taking a restaurant as the search feature (on which the visualization is centred) POIs within a certain radius *R* are first fetched. A classic selection operation would then yield a result similar to Figure 4, where only restaurants are selected. Utilizing the semantic similarity measure we could select the POIs similar to our search feature. Different measures will result in the same ranking of features but with different scores. Figure 5 and Figure 6 show the results of filtering by two different thresholds applied to different similarity measures (in this example the thresholds are set manually but theoretically thresholds could be set based on statistics, pattern recognition or other numerical approaches). The KeyValue measure tends to yield consistently lower numeric values. The list of features that have been selected by the similarity-based selection, tend to be from different classes of amenities but also tend to have shared keys and values with our search feature. In this example an important key is *cuisine*. The test of the aggregation operation results in Figure 7 and Figure 8, again showing different thresholds for the four similarity measures. As mentioned, the aggregation operator should consider a spatial distance limit when attempting to aggregate features as they might be far from each other.

5 Discussion and Conclusion

We have proposed to include measures based on semantic similarity between map features into the process of map generalization. The analysis was based on four different measures and two generalization operators. The input data was in the form of POIs but in principle other types of input data (roads, buildings, natural features and etc.) could be taken into account as well. The idea behind this proposal is to include knowledge derived from user-generated data. Including this knowledge helps us moving the generalization process one step closer to people's (i.e. users') definition and perception of geographic phenomena. In the case of OSM data, people's common perception is reflected in the feature attributes (tags). Commonly agreed definitions and attributes in the form of tags naturally include uncertainty and have lower data quality, but their nature is interesting when users' contributions are central to the research subject

A modification of the behaviour of generalization operators (selection and aggregation) has been observed, where by inclusion of *semantic similarity* measures, the similarity-based selection process has fetched different sets of features (which are closer based on their tag definitions), while the similarity-based aggregation operator has resulted in different combinations of features to be merged into one feature (it remains to be discussed which properties the new feature should inherit).

Besides evaluating the results of the proposed operators with repeated experiments and different situations, we intend to extend this initial study in different ways. The first extension is to apply the concept of semantic similarity on other map generalization operators. Where here we have taken feature-feature similarity (or resource-resource similarity in folksonomy parlance) another study could investigate tag-tag similarity, which will give similarities at a higher level of the classification hierarchy. This could be in line with relating the current study with other ontologies based on OSM (such as OSMonto). Another crucial step is to work toward an efficient strategy of combining the semantic similarities with spatial measures and constraints (Euclidean distance, network distance etc.).

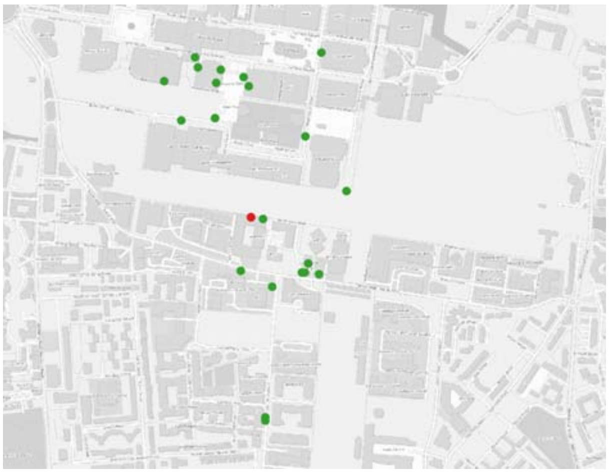

Figure 3 - Amenities within a radius *R* around a search feature (restaurant).

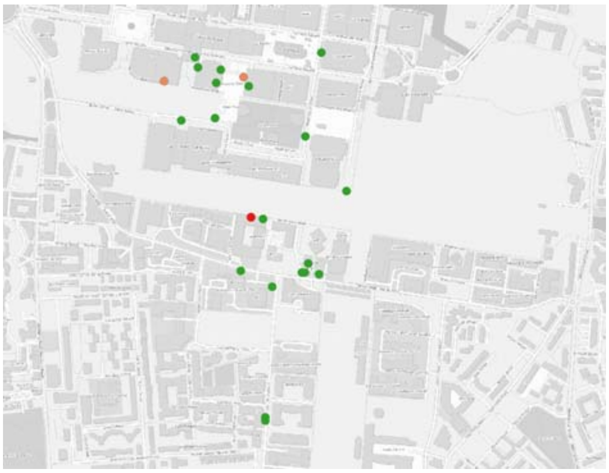

Figure 4 - A classic selection operation where features are selected based directly on a given classification attribute (amenity=restaurant in this case) or ranking attribute. Selected features shown in orange.

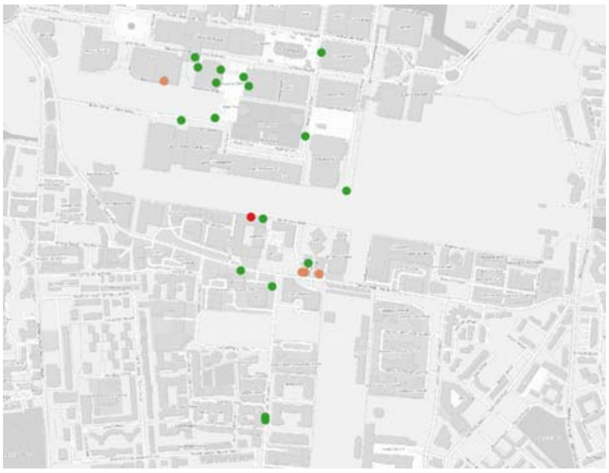

Figure 5 - Selection based on similarity to the search feature (here *α* is equal to 0.75, 0.81, 0.85, 0.41 for Jaccard, dice, cosine, KeyValue similarity, respectively). Selected features are shown in orange.

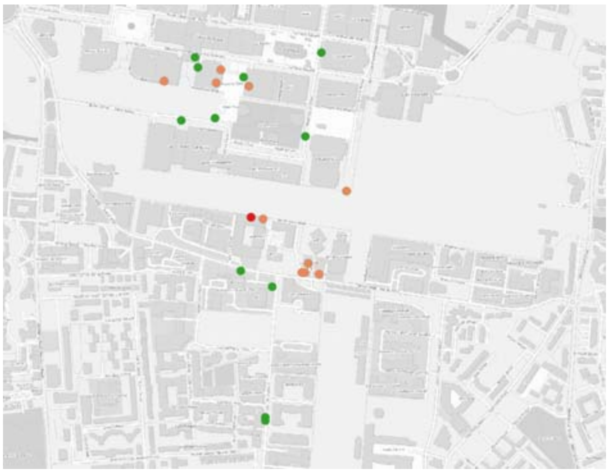

Figure 6 - Selection based on similarity to the search feature (here α is equal to 0.5, 0.667, 0.71, 0.33 for Jaccard, dice, cosine, KeyValue similarity, respectively). Selected features are shown in orange.

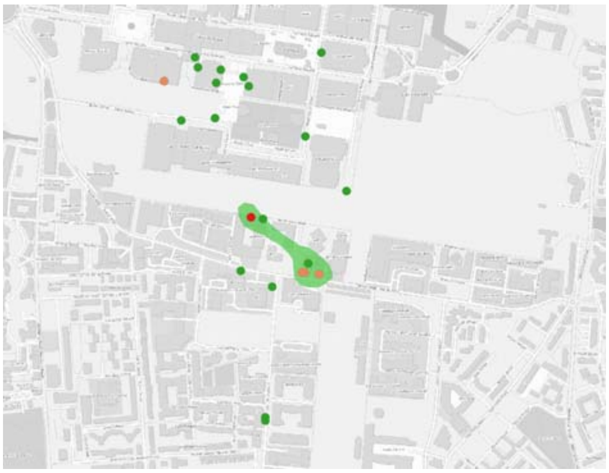

Figure 7 - Aggregation based on similarity to search feature (α equals 0.75, 0.81, 0.85, 0.41 for Jaccard, dice, cosine, KeyValue, respectively). A new feature is generated that inherits the common attributes but generally includes meanings like restaurant and cuisine.

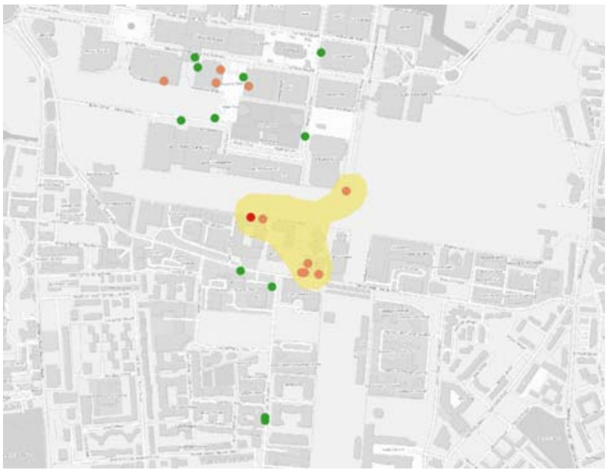

Figure 8 - Aggregation based on similarity to search feature (α equals 0.5, 0.667, 0.71, 0.33 for Jaccard, dice, cosine, KeyValue, respectively). A new feature is generated that inherits the common attributes but generally includes meanings like restaurant and cuisine.

Acknowledgements

This paper represents part of the PhD project of the first author. Funding by the Swiss National Science Foundation under project PlaceGen (200021_149823) is gratefully acknowledged.

References

[1] **Harrie, L., & Weibel, R. (2007).** Modelling the overall process of generalisation. Generalisation of geographic information: cartographic modelling and applications, 67-87.

[2] **Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006).** Emergent Semantics in BibSonomy. GI Jahrestagung (2), 94, 305-312.

[3] **International Cartographic Association (1973).** Multilingual Dictionary of Technical Terms in Cartography. F. Steiner.

[4] **Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. (2009).** Evaluating similarity measures for emergent semantics of social tagging. In Proceedings of the 18th Intl. Conference on World Wide Web (pp. 641-650). ACM.

[5] **McMaster, Robert Brainerd, and K. Stuart Shea. (1992).** ”Generalization in Digital Cartography.” Washington, DC: Assoc. of American Geographers.

[6] **Steiniger, S., and Weibel, R. (2007).** Relations among map objects in cartographic generalization. Cartography and Geographic Information Science, 34(3), 175-197.

[7] **Trant, J. (2009).** Studying social tagging and folksonomy: A review and framework. Journal of Digital Information, 10(1).

[8] **Vander Wal, T. (2005).** Folksonomy. Presented at Online Information, 2005. Accessed at http://www.vanderwal.net/essays/051130/folksonomy.pdf on 31 March 2016.