GeoAI’17: GeoAI’17:1st ACM SIGSPATIAL Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery, November 7–10, 2017, Los Angeles Area, CA, USA

Finding equivalent keys in OpenStreetMap: semantic similarity computation based on extensional definitions

Ivan Majic
The University of Melbourne
Parkville, VIC 3010, Australia
imajic@student.unimelb.edu.au

Stephan Winter
The University of Melbourne
Parkville, VIC 3010, Australia
winter@unimelb.edu.au

Martin Tomko
The University of Melbourne
Parkville, VIC 3010, Australia
tomkom@unimelb.edu.au

ABSTRACT

Volunteered Geographic Information (VGI) projects, such as OpenStreetMap (OSM) enable the public to contribute to the collection of spatial data. In OSM, users may deviate from spatial feature annotation guidelines and create new tags (i.e. *key=value* pairs), even if recommended tags exist. This is problematic, as undocumented tags have no set meaning, and they potentially contribute to the dataset heterogeneity and thus reduce usability. This paper proposes an unsupervised approach to identify equivalent documented attribute keys to the used undocumented keys. Based on their extensional definitions through their values, co-occurring keys and geometries of the features they annotate, the semantic similarity of OSM keys is evaluated. The approach has been tested on the OSM dataset for the state of Victoria, Australia. Results have been evaluated against a set of manually detected equivalent keys and show that the method is plausible, but may fail if some assumptions about tag use are not enforced, e.g., semantically unique tags.

CCS CONCEPTS

• **Information systems → Geographic information systems**; *Data cleaning*; *Similarity measures*;

KEYWORDS

Volunteered Geographic Information, VGI, OpenStreetMap, OSM, semantic similarity, data quality, data cleaning, spatial database

ACM Reference format:
Ivan Majic, Stephan Winter, and Martin Tomko. 2017. Finding equivalent keys in OpenStreetMap: semantic similarity computation based on extensional definitions. In *Proceedings of GeoAI'17:1st ACM SIGSPATIAL Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery, Los Angeles Area, CA, USA, November 7–10, 2017 (GeoAI'17),* 10 pages.
https://doi.org/10.1145/3149808.3149813

1 INTRODUCTION

Collaborative mapping, known as Volunteered Geographic Information (VGI) [8], led to a significant change in the availability of spatial data – and led to a substantial increase in annotation of space. The most prominent VGI project is currently OpenStreetMap[1] (OSM). In OSM, contributors use attributes expressed as tags (i.e. *key=value* pairs) to annotate geographical features. The OSM model is non-restrictive, as contributors can create and assign tags as they see fit. As an attempt to standardize feature descriptions, OSM provides contributors with *tagging guidelines*[2] through its OSM Wiki website. However, contributors do not always comply to these guidelines and sometimes introduce attributes not covered by the guidelines to annotate mapped features.

While using undocumented attributes may lead to a richer annotation of space, it can also be problematic due to the lack of shared definitions of the attribute‘s meaning and usage. Undocumented tags alone (e.g. OSM tag *period=1970*) may not be sufficient for other users to understand the intended meaning of the attribute. Some of the undocumented attributes may even have an equivalent key documented in the OSM Wiki, which should have been used instead. By creating new attributes instead of using the existing ones, semantic heterogeneity of the data is increasing which reduces the dataset's usability [15] and may cause users to completely ignore or discard potentially useful information, or stop them from using the dataset at all [12].

The aim of this paper is to analyse undocumented keys in OSM (e.g. *period*) and, if they exist, detect equivalent documented keys (e.g. *historic:period*) that should have been used instead. In this paper, two OSM keys are considered equivalent if they are used to attribute the same information to the feature, if they are interchangeable without the loss of information and if their joint use would result in redundant information. Semantic similarity is utilized in order to find and rank potential equivalent keys. The hypothesis of this research is that attribute keys that represent same or equivalent concepts are used in a similar way by the contributors. The lack of documentation for undocumented keys makes existing methods for computation of semantic similarity in OSM [2, 3] unusable in this case. Instead of using the structure of the OSM Wiki website [2] or their lexical definitions [3], the approach presented in this paper assesses the actual keys' usage to evaluate their similarity. The presented approach evaluates usage statistics of key values, of other co-occurring keys, and of the types of geometries capturing the annotated features to identify equivalent keys. This approach can be considered intrinsic because it has no need for either external data or explicit key definitions to compute key similarity.

The main contribution of this paper is the intrinsic approach to computing the semantic similarity between attributes. By computing the similarity from the data itself, the proposed approach offers an alternative to existing methods for cases where they cannot

GeoAI'17, November 7–10, 2017, Los Angeles Area, CA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5498-1/17/11…$15.00
https://doi.org/10.1145/3149808.3149813

[1] https://www.openstreetmap.org/

[2] http://wiki.openstreetmap.org/wiki/Category:Tagging_guidelines

be applied. As such, this paper contributes to the field of spatial data cleaning by enabling to improve the quality and semantic homogeneity of loosely-controlled VGI datasets.

The remainder of this paper is organised as follows. Section 2 contains an overview of related work. The proposed approach and underlying methodology are presented in Section 3. The experimental setup of a use-case applied to OSM data of Victoria, Australia, is explained in Section 4. The results are presented in Section 5, while Section 6 contains their evaluation and discussion. Finally, conclusions and directions for future work are stated in Section 7.

2 RELATED WORK

As VGI datasets were originally produced by untrained individuals, the need to evaluate and monitor their quality has been recognised early on. First VGI quality assessment methods were based on comparing VGI to authoritative datasets [7, 10, 18]. These assessments were mostly focused on completeness and aspects of positional accuracy of VGI data, in particular OSM. With rapid growth, in certain areas OSM soon had richer data than many authoritative datasets. Because of this, intrinsic quality assessment methods with no need for external data are recently being explored [1, 4, 9, 11, 12]. Although the semantics have not received as much attention as other aspects of VGI, there have been several studies on this topic in the recent years.

Ballatore et al. [2] have developed the OSM Semantic Network[3] which captures semantic relationships between geographic concepts from their descriptions in the OSM Wiki website. The OSM Wiki contains definitions and usage guidelines for OSM tags. The authors conceptualize the OSM Wiki website as a directed graph where vertices are the web pages and edges are the hyperlinks between them. This directed graph is extracted from the OSM Wiki website by OSM Wiki Crawler, an open source tool developed specifically for this purpose. They then use a generic co-citation algorithm *P-Rank* [17] to calculate semantic similarity between OSM tags. This algorithm is based on assumptions that two entities are similar if they are referenced by similar entities, and that two entities are similar if they reference similar entities [17]. Finally, they have evaluated the cognitive plausibility (i.e. the ability to replicate human judgement) of their method by comparing the results against an evaluation dataset which contains human-generated similarity judgements for geographic concepts. Their results show that co-location similarity measures can be used with semantic network to calculate semantic similarity between OSM tags. However, the authors have considered only tag-to-tag similarity and have left other cases for future work. Also, because this method uses OSM Wiki pages for similarity calculation, it cannot be applied to keys and tags that are not documented in the OSM Wiki.

In their other research [3], Ballatore et al. have used volunteered lexical definitions and paraphrase-detection techniques to compute semantic similarity of geographic terms. This work is based on the intuition that similar terms tend to be defined with similar terms. They have obtained lexical definitions of OSM terms by extracting them from the OSM Semantic Network. After filtering and weighting the definitions' corpus, they have constructed semantic vectors from the definitions' terms. To overcome the issue that occurs when there is little or no overlap between the vectors, they have utilized a term-to-term similarity function based on the WordNet lexical database. By doing so, they have captured continuous semantic distance between terms in semantic vectors. Similarity between the lexical definitions of geographic terms was finally calculated as a vector-to-vector similarity between their corresponding semantic vectors. As in previous study, they have evaluated their method against human-generated similarity judgements and showed that their method can achieve high cognitive plausibility. Their approach still requires lexical definitions of geographic terms and is therefore not applicable for the identification of undocumented attribute keys that lack explicit definitions.

Vandecasteele and Devillers [15, 16] have taken a more practical approach to improving semantic quality of OSM data. They have developed a plugin for the Java OpenStreetMap Editor (JOSM) called *OSMantic*[4]. The plugin aims to improve the semantic homogeneity of the OSM data by suggesting tags to the contributor during the editing process and by displaying a warning when both too high or too low semantic similarity between tags has been detected. Semantic similarity is computed by utilizing the OSM Semantic Network and a P-Rank co-citation algorithm described in [2]. Inspired by Tobler's first law of geography [14], the authors assume a higher semantic homogeneity for objects that are closer in space. Thus, the plugin is only analysing the semantics of objects within the neighbourhood of the one that is being edited or created. The warnings that are being displayed are meant to reduce semantic heterogeneity during the editing process. They notify the user when semantic similarity between existing and new tags on the OSM object is lower than the arbitrary threshold. When there is a high semantic similarity between a tag on the object being edited and tags on objects in its neighbourhood, a warning is displayed to notify the user that a different tag is being used for that attribute in that area. Although this plugin is based on the method from [2] and same limitations apply, the authors propose the similarity between sets of tags that describe a geographic object as a future improvement of semantic similarity results.

3 METHODOLOGY

This paper proposes an intrinsic approach for computing the semantic similarity of OSM keys that has no need for external data or explicit definitions of keys, inspired by methods for assessing similarity of documents in information retrieval.

In logic, two different types of definitions of concept are recognised: intensional and extensional. In [5], an intensional concept definition is defined as one that *"provides the meaning of an expression by specifying necessary and sufficient conditions for correct application of the expression"*. This is equivalent to formal specifications of meanings of terms in dictionary definitions, or in our case, feature and attribute definitions in the OSM Wiki guidelines. In contrast to [3], where intensional lexical definitions were used, this paper explores extensional definitions of OSM keys for the identification of equivalent keys. An extensional definition *"provides a list of those instances in which the expression being defined is applicable"*. In our case, this can be the set of values that an attribute takes in the dataset.

[3] http://wiki.openstreetmap.org/wiki/OSM_Semantic_Network

[4] http://wiki.openstreetmap.org/wiki/JOSM/Plugins/OSMantic

OSM represents physical features of the real world with three basic data structures (nodes, ways and relations), described by attributes in form of tags[5]. Nodes are defining features with point geometries, ways are defining linear features and boundaries, while relations capture how complex arrangements of features relate together. In this paper, attribute data of node and way features are analysed, while relation features are left for future work because of their complexity and relative underrepresentation in the dataset [12]. Features in OSM can have an unlimited number of tags, as well as no tags at all. The only restriction is that all attribute keys used on the same feature have to be unique (they form a set). Yet, as keys are only *syntactically* distinct, this limitation is weak and does not assure that no two tags with the same meaning are used within the set describing a single feature. *Common attributes*[6] (*id*, *user*, *uid*, *timestamp*, *visible*, *version*, *changeset*) are not created by contributors, but automatically generated for every OSM feature, and are thus excluded from the analysis in this paper.

3.1 Extensional definition of keys based on attribute statistics

The first step in the proposed approach is to analyse and describe attribute keys. Following extensional logic, we consider all features annotated with the attribute key k. The key k is then defined by all of its instances k_i. If attribute key k has n instances, an extensional definition of k will be a set $K = \{k_1, k_2, \ldots, k_n\}$. Thus, the similarity between two keys is in fact the similarity of their instances. The entire set of a key's instances in an area of interest is analysed, with the aim of describing how the key is used.

Three properties of keys have been identified as relevant for this task. First property of every attribute key are the values it can have. The second property being considered are other keys that are being used with it to describe the features. In the remainder of the paper, this property is called co-occurring keys or co-keys. The third and final property are geometries of features the key is being used with. In that sense, node, way and closed way geometries are distinguished in this paper, and this property is referred to as geometry type.

The values of key k will be a set $V_k = \{v_1, v_2, \ldots, v_n\}$ where v_i is the value of instance k_i. Since every instance k_i can have multiple co-occurring keys, the co-keys of k will be a set CK of sets of co-keys describing each instance of k, such that $CK_k = \{C_1, C_2, \ldots, C_n\}$, where $C_i = \{c_1, c_2, \ldots, c_m\}$ is a set of co-occurring keys for instance k_i. In other words, a key k has n instances and each instance has m co-keys, where m can differ per instance of k.

Geometry types of k can be written as $G_k = \{g_1, g_2, \ldots, g_n\}$ where g_i is the geometry type of instance k_i. In aspect of a key's usage, co-occurring keys and geometry type properties describe the context in which a key is being used (e.g. key *period* is most often used with key *building* on closed way geometries). Result of this statistical analysis is a set of three histograms for each key that has been analysed. The values histogram (Figure 1) shows all the values that occur with a given key and their corresponding counts. Same analogy goes for the co-keys histogram (Figure 2), while histogram describing the geometry types which key is being used with can only have three values - nodes, ways and closed ways (e.g. all 673 instances of key *period* occur on features with closed way geometries).

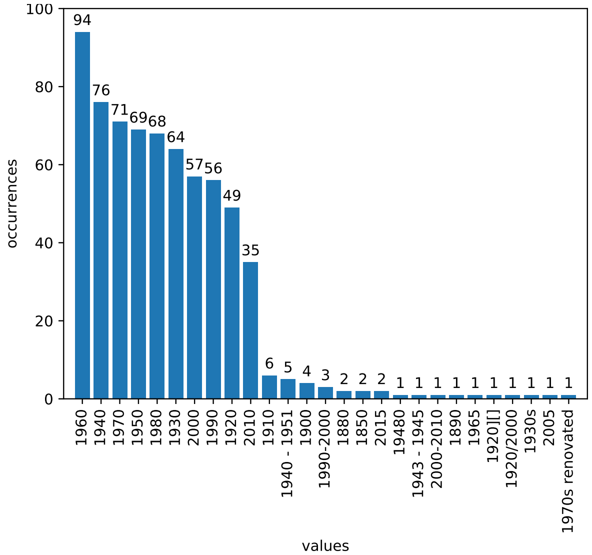

Figure 1: Values histogram for key *period* in the state of Victoria, Australia.

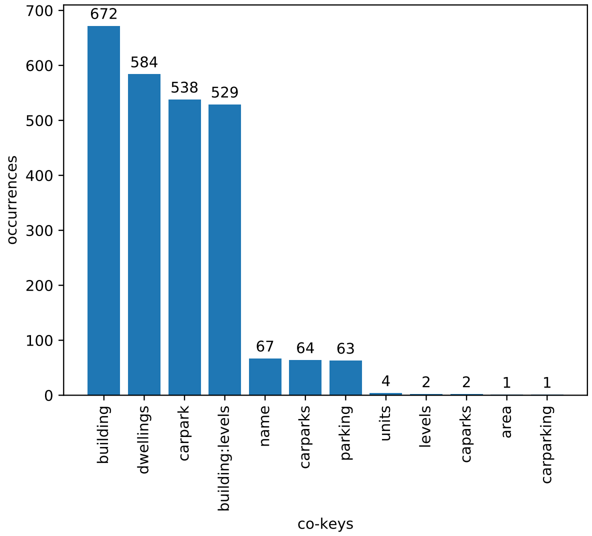

Figure 2: Co-keys histogram for key *period* in the state of Victoria, Australia.

3.2 Computing the semantic similarity

The proposed approach is inspired by methods for assessing the similarity of documents from information retrieval [6, 13]. In information retrieval, it is custom to represent a set of documents as a set of vectors in vector space. To create such vectors, term frequency that shows how frequently does each unique term (i.e. word in a textual document) occur in a document can be used. The usual

[5] http://wiki.openstreetmap.org/wiki/Elements
[6] http://wiki.openstreetmap.org/wiki/Elements

way of obtaining term frequencies of documents is to first convert each document to a bag of terms representation. Then, the counts (i.e., distribution) of each unique key is captured and is equivalent to the term frequency vectors in information retrieval. As each document has a corresponding term frequency vector, each axis in vector space corresponds to one term. The similarity between two documents is then a similarity between their term frequency vectors in a vector space. Standard measure of similarity between term frequency vectors is cosine similarity. This approach excludes the effect that document lengths may have on the result, and essentially compares relative distributions of terms in documents. If two documents are denoted as d_1 and d_2, and their vector representations are denoted as $\vec{V}_{d_1}$ and $\vec{V}_{d_2}$, their cosine similarity can be computed as

$$sim(d_1, d_2) = \frac{\vec{V}_{d_1} \cdot \vec{V}_{d_2}}{||\vec{V}_{d_1}|| \cdot ||\vec{V}_{d_2}||}$$

where the numerator is a dot product of document vectors $\vec{V}_{d_1}$ and $\vec{V}_{d_2}$, and the denominator is a dot product of their Euclidean lengths $||\vec{V}_{d_1}||$ and $||\vec{V}_{d_2}||$ [13]. In the same notation, cosine distance between documents d_1 and d_2 can be expressed as

$$cos\_dist(d_1, d_2) = 1 - sim(d_1, d_2)$$

where cosine distance is $cos\_dist(d_1, d_2) \in [0, 1]$. Cosine distance between two documents will be 0 if relative distributions of their terms are identical, while cosine distance of 1 indicates maximum dissimilarity of documents.

Thus, the values, co-keys and geometry types of OSM keys are described with three respective term frequency vectors. If two keys are denoted as A and B, their values are represented with vectors $\vec{A}_{values}$ and $\vec{B}_{values}$, their co-keys are represented with vectors $\vec{A}_{co\text{-}keys}$ and $\vec{B}_{co\text{-}keys}$, and their geometry types are represented with vectors $\vec{A}_{geometry}$ and $\vec{B}_{geometry}$. Each OSM key is thus described with values, co-keys and geometry types equivalent to a document. To compute the similarity of the keys, similarities of their values, co-keys and geometry type vectors are computed individually:

$$values\_distance(A, B) = 1 - sim(\vec{A}_{values}, \vec{B}_{values}),$$

$$co\text{-}keys\_distance(A, B) = 1 - sim(\vec{A}_{co\text{-}keys}, \vec{B}_{co\text{-}keys}),$$

$$geometry\_types\_distance(A, B) = 1 - sim(\vec{A}_{geometry}, \vec{B}_{geometry}).$$

Since the goal of this method is to find potential equivalent keys for a specific undocumented OSM key, a new three-dimensional vector space with the undocumented key in its origin can be defined (Figure 3). Each axis of this vector space represents one of the dimensions that are used to compute the similarity between keys - values, co-keys and geometry types. If the undocumented key is denoted as U, every documented key D will be represented as a three-dimensional point (X_D, Y_D, Z_D) where

$$X_D = values\_distance(U, D),$$

$$Y_D = geometry\_types\_distance(U, D),$$

$$Z_D = co\text{-}keys\_distance(U, D)).$$

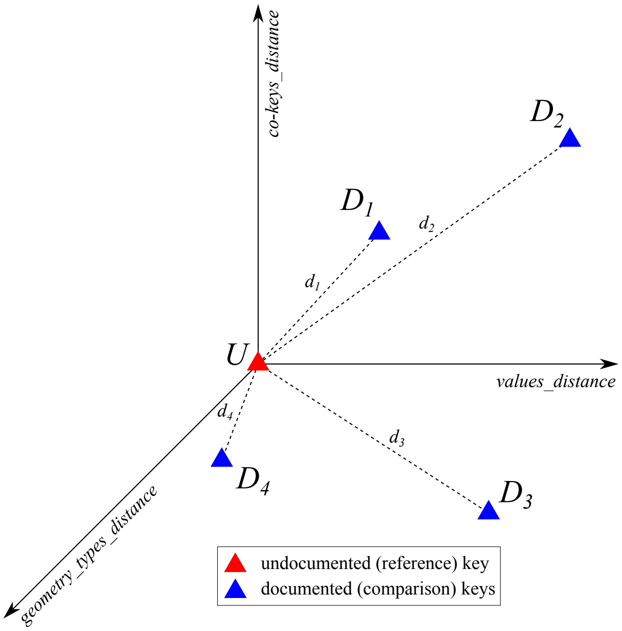

Figure 3: Three-dimensional vector space for similarity computation.

Finally, the overall similarity between undocumented OSM key $U(0, 0, 0)$ and documented key D is equal to their distance in this three-dimensional space

$$similarity(U, D) = \sqrt{X_D^2 + Y_D^2 + Z_D^2}.$$

4 EXPERIMENTAL SETUP

In order to test the hypothesis, semantic similarity between undocumented and documented OSM keys has been computed, using only their extensional definitions. The setup consists of a *PostgreSQL*[7] database enabled with *PostGIS*[8] and *hstore*[9] extensions. Raw OSM data were imported using *osm2pgsql*[10]. The workflow was implemented in the *Python* programming language.

4.1 Datasets

The region on which this experiment was carried out is the state of Victoria, Australia. The OSM dataset for this region was acquired using the OSM *Overpass API*[11] on 18 July 2017. Excluding relation features, this dataset contains 8,323,773 OSM features described with 2,777,902 tags (Table 1). The OSM Wiki documentation pages for tags and keys are available through the *taginfo*[12] system. This system provides detailed information about OSM tags, projects and the Wiki. *Taginfo* data were downloaded on the same date as OSM dataset, and used as an OSM Wiki reference.

[7] https://www.postgresql.org/
[8] http://postgis.net/
[9] https://www.postgresql.org/docs/9.6/static/hstore.html
[10] http://wiki.openstreetmap.org/wiki/Osm2pgsql
[11] http://wiki.openstreetmap.org/wiki/Overpass_API
[12] https://taginfo.openstreetmap.org/

Table 1: Statistics of the OSM dataset for the state of Victoria, Australia.

Nodes		7,510,661
Ways		813,112
Tags		2,777,902
Keys	*documented*	2,766,057
	undocumented	11,845
Unique keys	*documented*	715
	undocumented	887

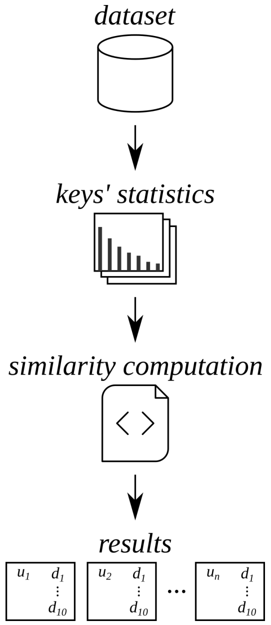

Figure 4: Steps in the workflow of the experiment.

4.2 Experimental workflow

Workflow for finding potential equivalent documented keys for each undocumented key consists of three steps (Figure 4). Starting point is the OSM dataset for selected region (here, stored in a relational database with support for key-value stores). All the keys in the database are analysed and their usage statistics extracted. Each key is thus described through the set of its *values*, *co-keys* and the associated *geometry type* vectors. These vectors are inputs for the similarity computation. Here, every undocumented key is compared with every documented key in the dataset. If there are n undocumented keys $u_{1...n}$, and m documented keys $d_{1...m}$ in the dataset, the resulting dissimilarity matrix has the dimensions $n \times m$, with each entry capturing the dissimilarity (distance) between the keys compared. Finally, for every undocumented key, a list of documented keys ranked by their similarity to the undocumented key is presented.

5 RESULTS

In this section, the experiment results for ten most frequent undocumented keys in the selected region are presented (Table 2). For each of these keys, the ten most similar documented keys (i.e. smallest distance in the comparison space) have been selected as their potential equivalents.

Table 2: Ten most frequent undocumented keys in the OSM dataset for the state of Victoria.

Key	Occurrences	Key	Occurrences
dwellings	756	ref:unimelb	316
period	673	ref:ptv	288
carpark	635	source:sport	275
full_name	333	name:source	272
lamp:type	329	source:details	222

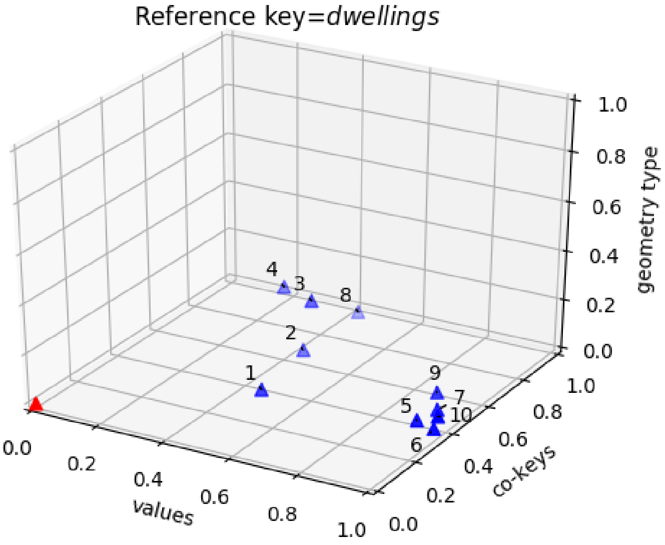

Figure 5: Potential equivalent documented keys for key *dwellings*, shown in its comparison space.

Table 3: Potential equivalent documented keys for key *dwellings*, ranked by similarity.

		Distances			
Rank	Key	Values	Co-keys	Geometry	Overall
1	building:levels	0.466	0.372	0.000	0.596
2	addr:flats	0.429	0.656	0.000	0.784
3	rooms	0.431	0.696	0.186	0.840
4	platforms	0.298	0.779	0.162	0.850
5	height	0.915	0.401	0.000	0.999
6	roof:levels	0.977	0.377	0.000	1.047
7	building:flats	0.928	0.488	0.000	1.048
8	capacity:disabled	0.429	0.958	0.002	1.050
9	building:min_level	0.875	0.586	0.000	1.053
10	source:building	1.000	0.353	0.072	1.063

Figure 5 shows the undocumented attribute key *dwellings* (i.e. red triangle) in its comparison space, together with its ten most similar documented keys (i.e. blue triangles). The distances between key *dwelling* and these documented keys are shown in Table 3. The theme of the detected keys and their co-keys distances indicate that key *dwellings* is most likely being used to describe buildings.

Table 4: Potential equivalent documented keys for key *period*, ranked by similarity.

				Distances	
Rank	Key	Values	Co-keys	Geometry	Overall
1	historic:period	0.275	0.277	0.000	0.390
2	source:building	1.000	0.379	0.072	1.072
3	roof:shape	1.000	0.418	0.000	1.084
4	roof:levels	1.000	0.425	0.000	1.086
5	height	1.000	0.427	0.000	1.087
6	building:type	1.000	0.427	0.000	1.087
7	building:roof	1.000	0.427	0.000	1.087
8	building:levels	1.000	0.429	0.000	1.088
9	roof:colour	1.000	0.448	0.000	1.096
10	building:material	1.000	0.455	0.000	1.099

Table 4 shows potential equivalent keys for key *period*. There is a clear distinction between the highest ranked key *historic:period* and all other detected keys. With smallest overall distance, *historic:period* is the most similar key to *period* in every singular dimension as well. While their values distances show there is no similarity in the values, the co-keys distances of other keys indicate that *period* is being used to describe buildings.

Table 5: Potential equivalent documented keys for key *carpark*, ranked by similarity.

				Distances	
Rank	Key	Values	Co-keys	Geometry	Overall
1	building:parts	0.044	0.517	0.000	0.519
2	building:part	0.050	0.642	0.002	0.644
3	headframe	0.038	0.719	0.106	0.728
4	bar	0.044	0.732	0.000	0.733
5	motorcycle_friendly	0.044	0.745	0.000	0.746
6	real_fire	0.044	0.765	0.000	0.766
7	ruins	0.048	0.737	0.243	0.777
8	disused:shop	0.396	0.685	0.143	0.803
9	generator:output:hot_water	0.044	0.804	0.000	0.805
10	toilets:handwashing	0.044	0.818	0.000	0.820

Results for key *carpark* are shown in Table 5. All the detected documented keys, except key *disused:shop*, have very similar values to key *carpark*. In terms of overall distance, results are grouped in a small range. In contrast to results for *dwellings* and *period*, documented keys that were detected here are thematically diverse.

Table 6: Potential equivalent documented keys for key *full_name*, ranked by similarity.

				Distances	
Rank	Key	Values	Co-keys	Geometry	Overall
1	brand	0.997	0.123	0.059	1.006
2	gambling	1.000	0.169	0.053	1.016
3	wikipedia	1.000	0.205	0.013	1.021
4	fuel:discount	1.000	0.249	0.053	1.032
5	name:fr	1.000	0.219	0.131	1.032
6	amenity	1.000	0.246	0.074	1.033
7	website	1.000	0.269	0.042	1.036
8	school	1.000	0.203	0.215	1.043
9	shop	1.000	0.238	0.195	1.046
10	wikidata	1.000	0.307	0.047	1.047

Results for keys *full_name* (Table 6), *lamp:type* (Table 7), *ref:unimelb* (Table 8), *ref:ptv* (Table 9) and *source:details* (Table 12) all have very similar characteristics. With exception of the key *brand* that was detected as a potential equivalent for key *full_name*, all the detected documented keys in these results have a maximum distance of 1 in the values dimension. This indicates that these 5 undocumented keys have unique values, which do not occur with other documented keys in the dataset. Furthermore, in terms of overall distance, all these keys have tightly grouped results with no documented keys standing out. Distances in the geometry types dimension tend to be very small. Because of this, overall ranking of detected documented keys mostly depends on distances in the co-keys dimension.

Table 7: Potential equivalent documented keys for key *lamp:type*, ranked by similarity.

				Distances	
Rank	Key	Values	Co-keys	Geometry	Overall
1	reg_ref	1.000	0.293	0.000	1.042
2	traffic_signals	1.000	0.294	0.000	1.042
3	crossing	1.000	0.296	0.001	1.043
4	direction	1.000	0.345	0.030	1.058
5	monitoring:water_level	1.000	0.374	0.000	1.068
6	tower:type	1.000	0.377	0.003	1.069
7	lamp_type	1.000	0.412	0.000	1.081
8	traffic_calming	1.000	0.419	0.000	1.084
9	exit_to	1.000	0.439	0.000	1.092
10	traffic_signals:direction	1.000	0.448	0.000	1.096

Table 10 shows the results for key *source:sport*. In terms of overall similarity, there is no documented key that stands out. With exception of *source:geometry*, distances in the geometry types dimension tend to be very small. For the values and co-keys dimension, results that have small distance in one dimension have a large distance in the other. By values, closest to key *source:sport* are keys

Table 8: Potential equivalent documented keys for key *ref:unimelb*, ranked by similarity.

		Distances			
Rank	Key	Values	Co-keys	Geometry	Overall
1	faculty	1.000	0.134	0.000	1.009
2	architect	1.000	0.147	0.000	1.011
3	club	1.000	0.233	0.011	1.027
4	name_1	1.000	0.248	0.080	1.033
5	department	1.000	0.292	0.000	1.042
6	owner	1.000	0.292	0.016	1.042
7	plant:source	1.000	0.313	0.000	1.048
8	role	1.000	0.320	0.000	1.050
9	date	1.000	0.321	0.017	1.050
10	platforms	1.000	0.293	0.161	1.054

Table 9: Potential equivalent documented keys for key *ref:ptv*, ranked by similarity.

		Distances			
Rank	Key	Values	Co-keys	Geometry	Overall
1	not_served_by	1.000	0.073	0.000	1.003
2	bus	1.000	0.082	0.004	1.003
3	route_ref	1.000	0.089	0.013	1.004
4	bay	1.000	0.112	0.000	1.006
5	tram	1.000	0.116	0.006	1.007
6	ref_name	1.000	0.123	0.000	1.008
7	shelter	1.000	0.136	0.030	1.010
8	public_transport	1.000	0.146	0.004	1.011
9	bench	1.000	0.171	0.067	1.017
10	operator	1.000	0.195	0.037	1.020

Table 10: Potential equivalent documented keys for key *source:sport*, ranked by similarity.

		Distances			
Rank	Key	Values	Co-keys	Geometry	Overall
1	size	1.000	0.032	0.000	1.001
2	source:position	0.161	1.000	0.000	1.013
3	fenced	1.000	0.270	0.000	1.036
4	lines	1.000	0.300	0.000	1.044
5	sport	1.000	0.304	0.001	1.045
6	garden:type	1.000	0.340	0.000	1.056
7	source:geometry	0.161	0.933	0.477	1.060
8	leisure	1.000	0.464	0.006	1.103
9	golf	1.000	0.616	0.001	1.174
10	golf:course	1.000	0.700	0.000	1.221

source:position and *source:geometry*, which is interesting because they only differ in the suffix namespace used with key *source*.

Table 11: Potential equivalent documented keys for key *name:source*, ranked by similarity.

		Distances			
Rank	Key	Values	Co-keys	Geometry	Overall
1	source:name	0.048	0.012	0.000	0.049
2	source:ref	0.544	0.269	0.004	0.607
3	name:etymology:wikidata	1.000	0.027	0.000	1.000
4	maxspeed:source	0.993	0.121	0.000	1.000
5	construction	1.000	0.031	0.008	1.001
6	proposed	1.000	0.041	0.001	1.001
7	surface	1.000	0.054	0.000	1.001
8	history	1.000	0.066	0.002	1.002
9	abutters	1.000	0.083	0.000	1.003
10	maxspeed	1.000	0.095	0.000	1.004

Table 11 shows the results for key *name:source*. The distances of the highest ranked key, *source:name*, indicate that it is almost identical to *name:source* in all three dimensions. This high similarity is present in the keys' names as well, although they were not used in the similarity computation.

Table 12: Potential equivalent documented keys for key *source:details*, ranked by similarity.

		Distances			
Rank	Key	Values	Co-keys	Geometry	Overall
1	addr:suburb	1.000	0.004	0.000	1.000
2	addr:place	1.000	0.076	0.000	1.003
3	addr:country	1.000	0.110	0.002	1.006
4	addr:state	1.000	0.110	0.001	1.006
5	addr:city	1.000	0.115	0.010	1.007
6	addr:postcode	1.000	0.119	0.011	1.007
7	addr:housenumber	1.000	0.123	0.009	1.008
8	addr:street	1.000	0.124	0.010	1.008
9	entrance	1.000	0.251	0.001	1.031
10	rental	1.000	0.260	0.089	1.037

6 EVALUATION AND DISCUSSION

In order to evaluate the results, equivalent documented keys for the analysed undocumented keys have been manually detected. These keys have then been used as a reference against which the results were compared. Table 13 shows the undocumented keys analysed in this paper and their equivalent documented keys. The rankings show how the proposed approach has ranked these equivalent keys in the results of their corresponding undocumented keys. It can also be seen that for five undocumented keys from the results section, equivalent documented keys do not exist. As such, these keys have not been evaluated.

The most frequent undocumented OSM key in the state of Victoria is key *dwellings*. In almost all of its instances, this key has a low numerical value and is used alongside key *building*. As such, the equivalent documented key that has been detected for key *dwellings*

Table 13: Manually detected equivalent keys for the undocumented keys presented in results, and their rankings.

Key	Equivalent key	Rank
dwellings	building:flats	7
period	historic:period	1
carpark	—	—
full_name	name	457
lamp:type	highway	579
ref:unimelb	—	—
ref:ptv	—	—
source:sport	—	—
name:source	source:name	1
source:details	—	—

is *building:flats*[13], which was ranked seventh in the results. Since this is the only equivalent documented key for *dwellings*, the results achieve a recall of 1 at seven returned documented keys. At the same number of returned potential equivalent keys, the precision of the results is $\frac{1}{7}$. As it can be seen in Table 3, the distance between these two keys in the values dimension is large. The most probable reason behind this is that key *building:flats* has only one instance in the experiment region. Although the value of this instance is 7, and the same value occurs with key *dwellings* 19 times, other values that occur with key *dwellings* make cosine distance between the keys' values vectors large.

The two undocumented keys for which the equivalents were successfully detected are *period* and *name:source*. Their respective equivalent keys *historic:period*[14] and *source:name*[15] are both highest ranked among the results (Tables 4 and 11). Consequently, both of these keys' results yield precision and recall of 1 at one returned documented key. The separation of these equivalent keys from the rest of the results also contributes to the trustworthiness of the results.

The equivalent documented keys for the remaining two undocumented keys *full_name* and *lamp:type* are *name*[16] and *highway*[17]. These equivalent keys were ranked 457th and 579th in the results, respectively. Thus, at ten returned documented keys, the proposed approach was not able to detect these equivalent keys. OSM Wiki documentation for key *name* states that all names should be contributed in their full form, under the key *name*. In almost all of its instances in the experiment region, key *full_name* has been used alongside key *name*, on the same features. Because of this, the two had a large distance in the co-keys dimension due to the fact that where both of them occur together, they instantly lose one common co-key (i.e. itself). Similarly, the approach was not able to detect key *highway* as a documented equivalent for key *lamp:type*. Although it would at first seem that the equivalent of key *lamp:type* is key *lamp_type*, the data insight proves otherwise. All the instances of key lamp:type have a value *street lamp*, which should be used in combination with key *highway*. Also, all instances of key *lamp:type* (i.e. tag *lamp:type=street lamp*) in the experiment region co-occur with tag *highway=street lamp*. As a result, there is a large distance in the co-keys dimension between these keys.

7 CONCLUSIONS AND FUTURE WORK

The aim of this paper was to detect equivalent documented attribute keys for the undocumented keys in OSM. Thus, the hypothesis was that attribute keys that represent same or equivalent concepts are going to be used in a similar way. The experiment that was carried out on the OSM dataset for the state of Victoria has shown uneven results. For three out of five evaluated undocumented keys, their equivalents were successfully detected and were highly ranked based on their semantic similarity to the undocumented keys. For the remaining two undocumented keys, the proposed approach was not able to detect their equivalent documented keys, although they exist.

The first issue that can be noticed in the proposed approach is that information about documented keys is obtained from the same regional dataset as information about the undocumented keys. The problem arises when a certain documented key that may be the equivalent to some undocumented key does not occur in that region. Results for key *dwellings*, whose equivalent key *building:flats* was ranked 7th, are affected by this problem. Although key *building:flats* has one instance in the experiment region, the difference in these keys' frequencies has had negative effect on their similarity in the values dimension. The other related problem is that documented keys are defined based on the way contributors use them in the selected region, rather than their definitions in the OSM Wiki documentation. The example of this problem is key *historic:period* that has been manually selected as an equivalent to key *period*. In OSM Wiki, *historic:period* is defined as a more detailed classification of key *historic:civilization*. As such, these keys should be used together to describe which civilization is feature related to (e.g. *historic:civilization=prehistoric; historic:period=stone-age*). In the experiment region, this key is used with numerical values to denote the year in which a feature was constructed, and is mostly used with buildings. There is a key *start_date*[18] that is intended for that purpose, but is not used often in the experiment region. A possible solution for these problems would be to obtain the information about documented keys globally, or to use their definitions from the OSM Wiki instead.

The second issue that is visible in the results is that the proposed approach is not able to detect equivalent keys if they are being used together on same features. Examples of this issue are keys *full_name* and *lamp:type* and their equivalent documented keys *name* and *highway*. With regards to attribute data, these cases can be seen as production of redundant or duplicate information which is certainly not desired. In the case of key *full_name*, it is always used in combination with key *name* to expand the name of the feature (e.g. *name=7 Eleven; full_name=7 Eleven Karingal*). OSM guidelines for names[19] state that all names should be contributed in the full form, since there are methods to shorten or abbreviate them, but it is impossible to produce full names if data is not available. Similar problem occurs with key *lamp:type*, which is always used

[13] http://wiki.openstreetmap.org/wiki/Key:building:flats
[14] http://wiki.openstreetmap.org/wiki/Key:historic:period
[15] http://wiki.openstreetmap.org/wiki/Key:source:name
[16] http://wiki.openstreetmap.org/wiki/Key:name
[17] http://wiki.openstreetmap.org/wiki/Key:highway
[18] http://wiki.openstreetmap.org/wiki/Key:start_date
[19] http://wiki.openstreetmap.org/wiki/Names

together with key *highway*, but contributes identical information (e.g. *lamp:type=street lamp; highway=street lamp*). It is hard to propose a solution for this problem, other than expect the contributors to be careful of potential redundancy in attribute data.

Regarding the computation of attribute keys' similarity, there are several ways in which the approach proposed in this paper can be improved. When analysing the values each key can have, there is no preprocessing that would categorise values based on their common properties. Consequently, cosine similarity of values vectors is only sensitive to the exact matches between individual values. Instead, it may be feasible to divide values into numerical and textual, or maybe even detect if numerical values are elements of a certain range (e.g. values of key *dwellings* would be defined as numbers in range 1–48). Furthermore, the geometry types that were distinguished in this paper are nodes, ways and closed ways. As such, this categorisation does not distinguish a building from a state boundary within closed ways. This problem can also be noticed in the results, where almost all distances in the geometry types dimension are very low and similar to each other. Possible improvement may be a more detailed categorisation that takes the lengths of the ways and the areas encapsulated by the closed ways into account. Regarding feature types, support for OSM relations needs to be added in order for this approach to be complete. Relations in OSM are annotated with tags, but so are their members - nodes, ways and other relations. With such structure, relations can be very complex and their preprocessing will be very important. Also, it would be interesting to assign different weights to each of the three comparison dimensions - values, co-keys and geometry types, and see how it would influence the results. Weighting could also be applied in the formulation of term frequency vectors for individual dimensions. Finally, the experiment workflow presented in this paper could also be applied as an iterative process. Upon detecting the equivalent documented key for one undocumented key, all the instances of the undocumented key in the dataset can be rectified. This would change the dataset and the results for the following undocumented keys that are to be analysed. In addition, the detected pairs of undocumented and their equivalent documented keys can be stored for later use. Such knowledge can be applied in the editing process to prevent undocumented keys from being used unnecessarily, and recommend an equivalent documented key.

ACKNOWLEDGMENTS

Support by the Australian Research Council (DP170100153) is acknowledged.

REFERENCES

[1] Ahmed Loai Ali, Falko Schmid, Rami Al-Salman, and Tomi Kauppinen. 2014. Ambiguity and Plausibility: Managing Classification Quality in Volunteered Geographic Information. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*. ACM, New York, NY, USA, 143–152. https://doi.org/10.1145/2666310.2666392

[2] Andrea Ballatore, Michela Bertolotto, and David C Wilson. 2013. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems* 37, 1 (2013), 61–81. https://doi.org/10.1007/s10115-012-0571-0

[3] Andrea Ballatore, David C Wilson, and Michela Bertolotto. 2013. Computing the semantic similarity of geographic terms using volunteered lexical definitions. *International Journal of Geographical Information Science* 27, 10 (2013), 2099–2118. https://doi.org/10.1080/13658816.2013.790548

[4] Christopher Barron, Pascal Neis, and Alexander Zipf. 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS* 18, 6 (2014), 877–895. https://doi.org/10.1111/tgis.12073

[5] Roy T Cook. 2009. *A dictionary of philosophical logic. [electronic resource]*. Edinburgh : Edinburgh University Press, ©2009.

[6] AnHai Doan, Alon Halevy, and Zachary Ives. 2012. *Principles of Data Integration* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[7] Jean-François Girres and Guillaume Touya. 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14, 4 (8 2010), 435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

[8] Michael F Goodchild. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 4 (11 2007), 211–221. https://doi.org/10.1007/s10708-007-9111-y

[9] Simon Gröchenig, Richard Brunauer, and Karl Rehrl. 2014. Digging into the history of VGI data-sets: results from a worldwide study on OpenStreetMap mapping activity. *Journal of Location Based Services* 8, 3 (2014), 198–210.

[10] Mordechai Haklay. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37, 4 (2010), 682–703. https://doi.org/10.1068/b35097

[11] Musfira Jilani, Padraig Corcoran, and Michela Bertolotto. 2014. Automated Highway Tag Assessment of OpenStreetMap Road Networks. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*. ACM, New York, NY, USA, 449–452. https://doi.org/10.1145/2666310.2666476

[12] S. Maguire and M. Tomko. 2017. Ripe for the Picking? Dataset Maturity Assessment based on Temporal Dynamics of Feature Definitions. *International Journal of Geographical Information Science* 31, 7 (2017), 1334–1358. https://doi.org/10.1080/13658816.2017.1287370

[13] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 100–123. https://doi.org/10.1017/CBO9780511809071.007

[14] Waldo R Tobler. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46 (6 1970), 234–240. https://doi.org/10.2307/143141

[15] Arnaud Vandecasteele and Rodolphe Devillers. 2013. Improving Volunteered Geographic Data Quality Using Semantic Similarity Measurements. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-2/W1, May 2013 (2013), 143–148. https://doi.org/10.5194/isprsarchives-XL-2-W1-143-2013

[16] Arnaud Vandecasteele and Rodolphe Devillers. 2015. *Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap*. Springer International Publishing, Cham, 59–80. https://doi.org/10.1007/978-3-319-14280-7_4

[17] Peixiang Zhao, Jiawei Han, and Yizhou Sun. 2009. P-Rank: A Comprehensive Structural Similarity Measure over Information Networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. ACM, New York, NY, USA, 553–562. https://doi.org/10.1145/1645953.1646025

[18] Dennis Zielstra and Alexander Zipf. 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany, In 13th AGILE International Conference on Geographic Information Science. *13th AGILE International Conference on Geographic Information Science* 1, 1–15.