Expert Systems With Applications 243 (2024) 122902

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Integrating social media data: Venues, groups and activities

Thiago H. Silva [a,b,*], Mark S. Fox [a]

[a] University of Toronto, 55 St. George Street, Toronto, M5S 0C9, Canada
[b] Universidade Tecnológica Federal do Paraná, Av. 7 de setembro 3165, Curitiba, 80230-901, Brazil

ARTICLE INFO

Keywords:
Data integration
Social media
Urban data
Ontology

ABSTRACT

Social media has been fuelling necessary research in different areas, including the large-scale study of urban societies. Most research is done with a single source of information. Integrating data from multiple sources provides several benefits; for instance, we can have more information about the venues or groups in the city. However, the integration of different sources of social media is a complex task. A critical task in the interoperability between different social media platforms is to provide an integration link. We focus on location-based social network platforms and present solutions to integration based on physical venues, groups of users interacting with them, and activities performed in those venues. Besides, we also propose an ontology (Social Media Integration Ontology — SMIO) that provides a target data model into which data from multiple sources can be mapped with more precise, shared semantics. Our proposed approaches and ontology can help to enhance the variety of data that describes a venue or group and foster research into urban societies.

1. Introduction

Over the last decade, social media has become a major data source, fuelling research and insights in sociology, urban geography, economics, etc. As we sift through the data, three types of information are routinely captured by many social media platforms: (i) Information about an activity (event) that is taking place; (ii) Information about people and groups of people that participate in the activity; and (iii) Information about the venue or place where the activity takes place.

Often, social media-based research uses data from a single platform (Ferreira, Silva, & Loureiro, 2020; Mueller, Silva, Almeida, & Loureiro, 2017; Santos, Silva, Loureiro, & Villas, 2020; Senefonte, Delgado, Lüders, & Silva, 2022; Silver & Silva, 2023). Nevertheless, integrating data from multiple platforms affords several benefits in different use cases, namely: (i) complementary data — certain information can be available in just one system, for instance, email or gender information, by performing an integration, we can have a richer picture of an entity; (ii) additive data - some entities might not be available in a certain system, thus, by integrating two systems, we might have a better comprehension of the entities in the city; (iii) confirmatory data — we can confirm certain entity information, for example, a telephone number, or the type of venue, where it appears in more than one source.

Integrating different sources of social media is not a straightforward task, not just due to the accessibility of the data, but even if data from multiple platforms are available, it poses several challenges: determining whether two or more sources refer to the same individual or group, location or venue and/or activity. A critical task in the interoperability between different social media platforms is to provide an integration/matching link (Ansell & Dalla Valle, 2023; Sun, Hu, Song, & Zhu, 2021; Sun, Zhu, & Song, 2019).

This study focuses on social media platforms that provide information from the physical world, such as location-based social networks (LBSNs), where users can share and interact with entities representing physical locations, such as coffee shops, restaurants, or universities. Under this perspective, we present different solutions to integration based on forms, i.e., physical venues, groups of users interacting with them, and activities performed in those physical spaces. In addition, we also propose an ontology that provides a target data model into which data from multiple sources can be mapped with more precise, shared semantics.

The contributions of this study can be summarized as:

- Venue integration solution — This first contribution is from the perspective of venues. We show that state-of-the-art approaches do a good job in most cases, except when we try to match venues from the same (business) chain geographically close to each other, such as in core downtown areas of major cities. Our solution builds on existing solutions by including a popularity feature, which can be extracted from the number of events performed by users, such as reviews and tips.

[*] Correspondence to: Av. 7 de setembro 3165, Curitiba, 80230-901, Brazil.
E-mail addresses: th.silva@utoronto.ca (T.H. Silva), msf@eil.utoronto.ca (M.S. Fox).

https://doi.org/10.1016/j.eswa.2023.122902
Received 15 March 2023; Received in revised form 6 December 2023; Accepted 7 December 2023
Available online 12 December 2023

- Group integration solution — There are available in the literature excellent efforts to match individuals in different systems; however, we focus on matching based on groups, where no previous effort was identified.
- Activity integration solution — Activities performed at a venue are implicitly implied by the categories of places. Multiple sources do not necessarily share the same categories of venues. Our approach explores semantically meaningful sentence embeddings associated with definitions of the terms composing the categories.
- Data integration solution — We propose an Ontology to support the integration of data from different location-based social media, considering the physical form itself, groups that interact with them, and the types of activities performed in those places.

The remainder of the study is organized as follows. Section 2 presents the motivation for this work. Sections 4.1, 4.2, and 4.3 discuss the integration by physical venues, groups, and activities, respectively, and our contributions to each of those parts, regarding the matching between different social media and the associated ontology for each entity. Section 6 presents the Social Media Integration Ontology. Section 7 presents the discussions and conclusions.

2. Motivation

Our focus on integrating data from multiple social media is driven by two needs: (1) to enhance the variety of data that describes a venue or group of users, and (2) to support research into the evolution of urban environments (Fox, Silver and Adler, 2022; Fox, Silver, Silva and Zhang, 2022; Silver, Adler and Fox, 2022; Silver, Fox and Adler, 2022).

2.1. Data enhancement

A concern in data science is the completeness and validity of data. Incomplete data provide an incomplete picture of the phenomenon being studied. Invalid data leads to invalid inferences. The aggregation of data from multiple social media sources offers several benefits:

1. **Complementary:** Some attributes, for the same entity, in one dataset are not found in another, such as check-ins in Foursquare not being available in Yelp. This complementation enriches the combination.
2. **Additive:** Data in each dataset is incomplete. For example, people or venues that appear in one source and not another. By combining data from two or more datasets, the data becomes more complete. It is assumed "same" attributes between the datasets.
3. **Confirmatory:** Data from one dataset can be used to confirm or refute what we know about a group, venue or activity found in other sources. For example, category labels of venues.

Nevertheless, integrating data from multiple sources poses many challenges, including:

- Correspondence: For each entity in one social medium, which entity does it correspond to a second social medium? Without the consistent use of unique identifiers across media, finding the corresponding entity can, in some cases, be a challenge. For example, the correspondence of venues across social media.
- Interpretation: The challenge of interpretation occurs when entity types and their attributes are not shared across social media. Entity types and attributes with the same name may be interpreted differently. For example, the categories with which venues are classified differ between Yelp and Foursquare. Determining what they are supposed to "mean" and how users interpret them presents a major challenge.

The development of algorithms for determining correspondence and ontologies for explicit representation of interpretations is key.

2.2. Urban evolution

A second motivation for integrating data from multiple social media is to provide data not normally available from other sources for the analysis of urban evolution. A model of urban evolution under development by the Urban Genome Project at the University of Toronto[1] "proposes the concept of the Formeme as the basic unit of urban evolution. A Formeme is a specific encoding of urban space as a combination of physical features and the groups and activities towards which they are oriented." (Fox, Silver and Adler, 2022; Fox, Silver, Silva et al., 2022; Silver, Adler et al., 2022; Silver, Fox et al., 2022). A Formeme is a way of physically organizing space for some sets of activities and groups.

A Formeme is composed of three components:

- P: the set of all possible types of physical forms in the domain
- A: the set of all possible types of activities (uses) in the domain.
- G: the set of all possible types of groups (users) in the domain.

Forms enable activities performed by groups. But the relationship is not uni-directional. Groups enact their own interpretation of forms to carry out activities for which the forms may not have been designed.

Formeme captures the relationship among elements of P, A and G. Formemes encode the information in a space, enabling their replication elsewhere, their maintenance into the future, or their recoding into new configurations. A Formeme f is defined to be a triple composed of P, A and G: $f = \langle f[p], f[a], f[g] \rangle$, where $f[p] \subseteq P \wedge f[g] \subseteq G \wedge f[a] \subseteq A$.

This is precisely the type of information captured by social media such as Yelp and Foursquare; each event is a combination of Form (i.e., venue), Activity (i.e., what takes place at the venue), and Group (i.e., the person participating in the event).

In the following sections, we present how we can integrate different sources by venues, groups, or activities. In the end, we introduce the main components of the SMIO ontology.

3. Related work

This section groups related studies regarding the different contributions of this work.

3.1. Studies on integrating venues

There are several approaches in the literature with good performance regarding the venue matching problem — for instance, Deng, Luo, Liu, and Wang (2019), Kim, Vasardani, and Winter (2017), Li, Liu, Dai, and Liu (2020), McKenzie, Janowicz, and Adams (2013), Piech, Smywinski-Pohl, Marcjan, and Siwik (2020), and Sun et al. (2019). Many of them share core characteristics, such as the type of features considered in the matching procedure and the distance metrics applied to them.

A simple way to tackle the integration of venue problem is to explore the values of attributes common to both datasets using a specific similarity measure. For instance, if two datasets both contain a name attribute for their venue, the Levenshtein distance could be used to match them. Similarly, other attributes, such as locations and categories, will also be compared using appropriate measures. However, such simple approaches will tend to work for a few cases.

Noticing this problem, McKenzie et al. (2013) proposed a weighted combination of shared attributes, showing considerable improvement compared to single attribute approaches. Along the same line, Deng et al. (2019) propose a multiattribute model built using the improved combination rule of the Dempster–Shafer evidence theory. More recently, to tackle the shortcomings of previous studies, Li et al. (2020)

[1] https://academic.daniels.utoronto.ca/urbangenome.

proposed a matching method integrating multiple determination constraints, which explores spatial topology, venue name role labelling, and bottom-up class constraints. Piech et al. (2020) evaluated six different classifiers for venue matching, performing experiments and follow-up comparisons to identify the most effective matching classifier. Their results indicate that the best venue matching classifier combines random forest algorithms that mix different similarity metrics for different venue attributes. Piech's approach has very good performance; however, there is a particular scenario that it could be improved, as we present in Section 4.1.

3.2. Studies on integrating activities

A group of works for the category similarity problem explores structure-based metrics (Deng et al., 2019; Zheng, Fen, Xie, Peng, & Fu, 2010; Zhu et al., 2017). For example, Deng et al. (2019) calculate the category similarity in two steps. First, they use a manual step to match all first-level category labels in two different categories hierarchy — for example, Catering Service (in System-1) and Food (in System-2). Next, the authors trace the category tree for the second and third-level categories and assign its parent category. They explore this trace to compute similarities/distances of categories in two systems. While this method could be interesting to match categories of a few venues, such as determining if two venues are identical, it is not useful when making a wide matching as proposed in this paper. Our approach does not demand manual matching and works with category hierarchies of diverse sizes, which could have any number of levels. The present study greatly builds upon our previous work (Silva & Fox, 2021) in several directions. We here shape our proposition to capture better the real world, where users can share and interact with entities that represent physical locations; thus, we position our solution to integration based on physical forms of the real world. We also present different solutions to integration based on groups of users interacting with forms and activities performed in those physical spaces. This considerably increases the integration power and enables more sophisticated information extraction. Besides, here we also propose a new ontology to support the data integration from different sources more precisely under this framework.

Another group of works on the problem of category similarity considers content-based metrics, which rely on semantic information of categories of entities (Ballatore, Bertolotto, & Wilson, 2015; Čerba & Jedlička, 2016; Chen, Song, & Yang, 2018). Chen et al. (2018) proposed an approach that blends a generic lexical database with a professionally supervised vocabulary to compute the relatedness of any two terms in the thesaurus. For example, "stream" and "river" are semantically similar, while "boat" and "river" are dissimilar but semantically related, so relatedness refers to this latter case. While this method points to a direction that potentially helps the category similarity problem, it meets some practical challenges — one of the most critical is the demand for controlled vocabulary for diverse contexts of interest, which are typically costly to obtain, thus, hampering generalization.

3.3. Studies on integrating individuals and groups

Mapping the same users/groups in different social media is still a challenging open issue in the literature (Silva et al., 2019). While some studies, such as Hristova, Musolesi, and Mascolo (2014), Hristova, Panzarasa, and Mascolo (2015), and Silva et al. (2014), show the potential of the integration by users or proposals regarding how it could be modelled, no solution for the matching problem is provided. However, there are some promising directions in the literature.

Zafarani, Tang, and Liu (2015) propose two approaches for identifying users across social media. The first one explores local link information, considering users across social media that share most mutual friends across systems as matches of the same individual. Note that this approach only considers users in the mapping that are one hop away; thus, the second approach mitigates this problem by considering multiple hops away. This method creates a k-dimensional vector for each user representing the number of users in the mapping that are 1 to k hops away from the user under study. The key intuition behind these approaches is the fact that users may join multiple social media, and when they do, it is more likely for them to become friends with users with a previous connection. This is an interesting method; however, in some media, especially on location-based ones, such as Yelp and Foursquare, the social network is less prevalent than in, for example, Facebook.

Another strategy is based on user behavioural patterns observed in different social media. The key idea is that unique behaviours due to, for example, personality or environment can provide redundant information across systems, which can be exploited to identify users across social media sites. In this direction, Zafarani and Liu (2013) use usernames to derive many features based on, for example, the way people write, that can be used by supervised learning to effectively connect users in different systems. Goga et al. (2013) also proposed a feature-based approach that combines several characteristics obtained from the user's shared content, e.g., timestamps, geolocation, and language, to perform user matching in different systems. In a different direction, Rodrigues, Boukerche, Silva, Loureiro, and Villas (2017) proposed an algorithm to identify probable matches between entities, e.g., users, from different systems, relying on spatiotemporal information from both systems for this task. Despite some limitations, the authors demonstrated that their technique could be useful for matching users on different systems based on mobility patterns.

There are also network-based approaches for feature representation in the user matching problem, obtaining satisfactory results (Zhang, Tang, Yang, Pei, & Yu, 2015; Zhou & Fan, 2019). For example, the study of Zhou and Fan (2019) jointly embeds both users and interactive behaviours of different social media into a low-dimensional representation space according to a set of known anchors, achieving promising results when matching users.

While there are several efforts to identify individual users across social media (Shu, Wang, Tang, Zafarani, & Liu, 2017), the matching of groups did not receive the same attention; to the best of our knowledge, no such study exists in the literature. For this reason, we focus our contribution on group matching in this study.

3.4. Studies on social media interoperability ontology

There are important ontologies for the problem of social media interoperability in the literature. For instance, Bojars and Breslin (2007) and Bojārs, Breslin, Finn, and Decker (2008) define SIOC as a solution to enable linking and reuse scenarios of data from social media sites. One key point here is the specification of the type of content, allowing the match between user posts and the content items created. Scerri, Cortis, Rivera, and Handschuh (2012) proposed an ontology that captures key facets of personal information shared, for instance, through online posts. Some of the most recent efforts and closer related to our proposal include the study of Sebei, Hadj Taieb, and Ben Aouicha (2020), where the authors propose the ontology SNOWL. This ontology primarily focuses on the users, for instance, profiles and actions, and targets user-generated content, such as check-ins and posts. Thus, SNOWL is intended to extend existing ontologies by instructing extra general concepts, relating mainly to the content itself, the interaction among users, and users and content.

4. Methods for integration

4.1. Integrating venues

Integrating venues is an important step that provides complementary, additive, and confirmatory benefits mentioned above, which, in turn, enables more sophisticated analyses. Integration by venue is responsible for merging two venues in different systems, for instance, Starbucks on Yelp and Foursquare. Venues are often the focus of social media postings. For instance, Yelp reviews and Foursquare check-ins are of venues. Integration by venue enables access to all information available in each system regarding this venue.

Our proposal builds on a recent effort presented by Piech et al. (2020). Similar to Piech et al. (2020), we also start by getting venue candidates to be matched. For a given venue to be matched from system X, we extract all venues within 300 m from it in system Y, forming the candidate set, in which distance metrics are calculated from the given point we are trying to match. In our case study, we consider the Levenshtein distance from the venue's names and the address name, the Euclidean distance from the venues' geolocations, and the Cosine distance from the categories set (Vijaymeena & Kavitha, 2016). The categories from the Yelp dataset were converted to Foursquare's taxonomy of categories using the proposed method in this study (SFox, presented in Section 4.2), where for each category in Yelp, we chose the first Foursquare category suggested by the method. In addition, we proposed incorporating the popularity of places to help improve (business) chain identification in dense urban areas; in our case, we consider the difference in popularity between Yelp reviews and Foursquare tips — see Section 5.1 for the rationale. Next, we supply these distances to a trained machine learning classifier using Random Forest (Tan, Steinbach, & Kumar, 2016) with 100 trees (the best result obtained by Piech et al.). After that, we identify venues classified as the most probable matches and sort them in descending order by the average value of the calculated similarity metrics. We take the first one from the sorted candidates as the matching venue, as in the study of Piech et al.

4.2. Integrating activities

Usually, activities are implied. In Foursquare, the explicit activity is the check-in, but the implied activity is what takes place at the venue, such as eating, social interaction, exercise, etc. These activities are often implied by the category of the venue; thus, categories are a proxy for activities in distinct systems.

The integration by activities expects connecting two venues by a certain category of place, e.g., a supermarket. Several urban studies use venue categories to comprehend city dynamics and urban social behaviour (Silva, de Melo, Almeida, Musolesi, & Loureiro, 2017; Tsutsumi, Fenerich, & Silva, 2019). For instance, Silva et al. (2017) present a new approach to identifying cultural boundaries between urban societies, considering users' food preferences measured by the category of venues visited in Foursquare. Tsutsumi et al. (2019) show a model that captures significant virtual relationships among businesses that are generated by users in the virtual world, which, by exploring the categories of venues, enables the identification of venues that represents non-obvious relations that might deserve particular attention of business owners, for instance, for new partnerships.

Commonly, each platform maintains its own categorization of venues; for example, this is the case for Yelp and Foursquare. Given these specificities, given a category of venues in a certain system, e.g., Yelp, we want to find the corresponding match in another system, e.g., Foursquare. As the number of categories is high on social media sites like those exemplified, manual identification is not practical (Silva & Fox, 2021).

We propose a solution called SFox that matches categories in two different systems based on words definitions and Sentence-BERT (sBERT), which is a modification of the BERT network using siamese and triplet networks that can derive semantically meaningful sentence embeddings, i.e., semantically similar sentences close in vector space (Reimers & Gurevych, 2019).

More specifically, for each category name c we get its definition def from WordNet (Miller, 1995); we use the Python library PyDictionary for that. Next, we compute sentence embeddings for c+def, using sBert based on two pre-trained sentence-transformer models (stsb-roberta-large and stsb-roberta-base) provided by the creators of sBert (Reimers & Gurevych, 2019). This step provides two sentence embeddings for all categories of both systems; one is provided by exploring stsb-roberta-large and the other by stsb-roberta-base. In possession of sentence embeddings, for each category in System-1, we find the ten most similar embeddings in System-2 – five provided by stsb-roberta-large and five by stsb-roberta-base. After that, we sort the ten candidates by the cosine similarity value.

Evaluated Approaches: We compared our proposal with three different ones. The problem tackled by all is the same: for a given category on System-1, find the ten most similar categories on System-2 in descending order. The first approach, namely Levenshtein, considers the Levenshtein similarity, a string metric for calculating the similarity between two sequences (Levenshtein et al., 1966), to compute this measure for a given category c in System-1 to all categories in System-2 to choose the 10 most similar ones to c.

The second approach, called Levenshtein+Structure, first conducts the Levenshtein approach. Having the ten most similar categories to c given by the prior step, it calculates the category similarity of c to this set using a structure-based approach, as Deng et al. (2019) did, and adds this similarity value to the Levenshtein similarity. Following the algorithm for the structure-based approach, we must establish a connection between first-level categories on the two systems studied, Foursquare and Yelp, in this study. For example, the 'Nightlife' category on Yelp was manually matched to 'Nightlife Spot' on Foursquare. Categories for other levels in the hierarchy do not need to be manually matched. The structure-based similarity category Sstruct is calculated according to: $S_{struct} = e^{-D/2\alpha}$, where $D = p_1 + p_2$, with $p_1 + p_2$ representing the distance from their shared root parent node to the node representing a certain category for System-1 and System-2, respectively, α is the maximum distance, which could be derived either from System-1 or System-2, in this study $\alpha = 4$. If there is no shared root node, the similarity is 0.

The third approach computes word embeddings for a particular category using the sBert, computing next the cosine similarity using the embeddings. For a given category in System-1, it finds the ten most similar ones in System-2 (five using the pre-trained model stsb-roberta-base and five using stsb-roberta-large).

4.3. Integrating individuals and groups

Assuming that we do not have unique identifiers for individuals and groups, the focus on integrating individuals and groups is based on similarity metrics. There are two parts to the problem:

1. Identifying the same individual across two or more social media, and
2. Identifying similar groupings of individuals across social media.

Our approach to group matching has four steps:

1. User representation. We represent users by categories of business by describing each user by the names of categories of places they performed an action, e.g., a review on Yelp or a check-in on Foursquare. Thus, users can be seen as "documents", and the category names, which can be repeated, are words in these documents. This is an example: $d_1 = \{$Coffee Shop, University, Coffee Shop, Bookstore$\}$, where $d_1 \in D$ represents a document describing the categories of places visited by $user_1$.
2. Profiles extraction and representation. We apply standard text pre-processing steps on all documents D representing users, and use these cleaned documents to identify latent topics using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA allows for several topics per document, and each of them can be interpreted as user profiles; thus, users can be associated with

different profiles, as in real life. LDA demands the number of topics to be found as a parameter. One criterion commonly used to find a good number of topics is a coherence measure, such as the UMass (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) one used in our experiments in this study. After identifying the topics, we map each document (which represents users) on the space defined by those topics: $u = \{x_1, x_2, \dots, x_n\}$, where $u \in U$ is a feature vector representing a particular document (user), and xi represents the probability that this user is associated with a certain topic (profile) among n topics considered.

3. Create groups by users' profiles. We find groups of users $G = \{g_1, \dots, g_k\}$ in the space represented by the feature vector U. For this task, we use the k-means algorithm (Tan et al., 2016), where k is identified according to the silhouette heuristic (Tan et al., 2016).
4. Representation of geographical areas by a distribution of groups' visits. Each geographical area a_i of a certain city, such as a neighbourhood or a census tract, is represented by the number of unique visits of each group (f_{gi}): $a_i = \{f_{g1}, \dots, f_{gk}\}$.

Note that a central element in this approach is the categories alignment between different systems. Suppose two systems share different categories' taxonomy, as is the case of Foursquare and Yelp. In that case, we have two options: (i) integrate by venues and then merge the categories in different systems, or (ii) integrate by categories. We exemplify in the next section the latter option.

5. Experimental results

5.1. Results on integrating venues

To exemplify our approach, we got the 1000 most popular venues in terms of check-ins in our Foursquare dataset and manually found the matches in our Yelp dataset to build the validation set — details about our datasets explored in this study are in Appendix A. Since our Foursquare dataset is from 2014, we considered data from Yelp regarding the same year. We identified 560 matches, representing our final test set. In our analysis, we only consider features present in all datasets. The accuracy of this approach was 0.95, which is in the range expected according to the baseline study.

However, when evaluating the results, we identified some mismatches regarding business chains in dense areas. This was the case, for example, for some of the Starbucks venues. It is common for metropoles like Toronto to have several places of the same chain nearby each other. For instance, Fig. 1 shows the locations of two Starbucks in a particular area in Toronto. As we can see, they are located very close to each other, a situation that happens considerably in this city.

Piech et al. (2020) approach can achieve good results regarding the venue matching problem, especially considering the venue's name, categories, address, and geolocation, as demonstrated above. Note that for chains, names and categories are typically expected to be the same; this also tends to be the case for websites, telephones, and other attributes. Thus, in this case, spatial features are important points of differentiation. First, note that the pin in the figure does not correspond exactly with the address; the owners opted to use the address of the mall's main entrance where it is located. Potentially, other businesses can use the same geolocation in this case. Approaches that consider string proximity based on the address description may also suffer from differences in different systems. This is the address for venue A (see Fig. 1) in Foursquare: 10 Dundas St E (at Yonge St.) Toronto. And in Yelp: 10 Dundas St E Toronto. This implies that similarities between those attributes will not be perfect. Thus, uncertainty will be introduced in the matching decision in the exemplified scenario; Venue B would be a candidate for matching with venue A because it is in the range of 300 m, and they could be wrongly matched in different systems.

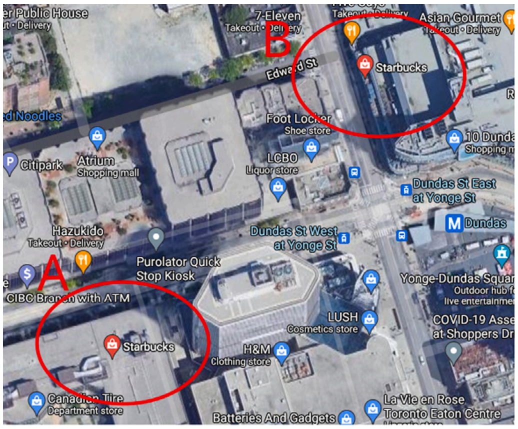


Fig. 1. Geographical locations of two Starbucks in a particular area in downtown Toronto [Image from Google Street View: http://maps.google.com/].

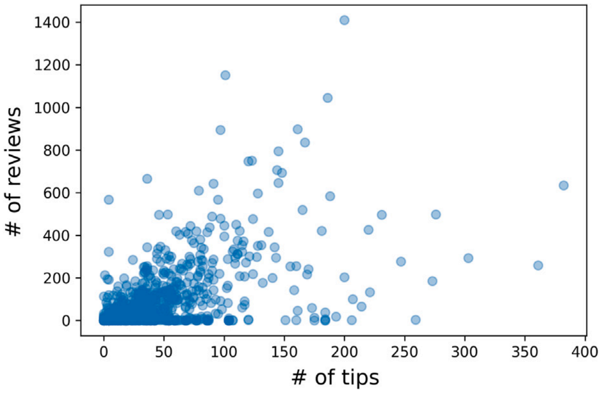


Fig. 2. Scatter plot considering the number of tips and reviews for the same venue. For this data, we have a statistically significant Spearman correlation of 0.79.

For this reason, we proposed incorporating an extra feature regarding places' popularity to help improve this chain limitation of previous proposals. In our case study, we found that Foursquare tips and Yelp reviews tend to be highly correlated. Considering all venues in our evaluation dataset this is the correlation we have a Pearson correlation = 0.64 (p-value = 1.387e–62) and a Spearman correlation = 0.79 (p-value = 2.465e–117). Fig. 2 shows a scatter plot considering the number of tips and reviews for the same venue, where we can see this clear trend quantified by the correlation coefficients.

In our Starbucks scenario exemplified above, the number of Foursquare tips and Yelp reviews for venue A is, respectively, 28 and 7, whereas, for venue B, we observe 15 tips and 3 reviews. Let us assume we want to find a match for the Foursquare venue A in our example to its corresponding one in Yelp – i.e., venue A at Yelp. When calculating the distance between those two popularity indices (tips and reviews) from Foursquare to Yelp, we have (without normalizations, to favour understanding): correct match = 28 − 7 = 21 and wrong match = 28 − 3 = 25. Thus, we have a smaller distance for the correct match. Note that introducing this popularity feature suggests that the problem can be minimized in this case.

In fact, this is true in our case study; when considering the same matching approach described above with the inclusion of popularity (with 0–1 normalization for each system), the accuracy improves to around 98% (versus 95%). Thus, the popularity feature not considered by previous efforts is recommended to be considered when available to help minimize this type of problem highlighted.

Table 1
Results of the evaluation of all regarded approaches for category matching in three different experiments.

	Levenshtein	Levenshtein + Structure	sBert	SFox
Experiment 1: Match with the first candidate	0.71	0.72	0.71	**0.82**
Experiment 2: Match with any of the top 3	0.79	0.77	0.83	**0.88**
Experiment 3: Match with any of the top 10	0.82	0.79	0.88	**0.89**

5.2. Results on integrating activities

For each approach, our evaluation assesses three distinct experiments. Experiment 1 considers the match with the best candidate suggested by the approach under examination, i.e., with the highest similarity score. This experiment refers to the case of an automatic method, where the matching follows what is suggested. The following two experiments consider the case where there is human intervention in the matching process. In experiment 2, the approach chooses the three best candidate categories so the user can pick the most suitable one. Experiment 3 is analogous to 2 but presents ten options to users. Experiment 3 is more costly because it supplies more cases to be examined; nevertheless, it could be worth the price depending on the case.

To assess the approaches defined in Section 4.2, we designed a test set having 300 random Foursquare categories. We manually tried to match these categories on the Yelp category system. Categories with no match found were disregarded, 25 categories in total. Thus, 275 words were evaluated — this is our ground truth.

Table 1 presents the results of the evaluation of all regarded approaches for category matching, presented in Section 4.2, in the three different experiments mentioned above. According to Table 1, SFox (our proposal) is better in all evaluations. Notably, it is considerably superior in the match with the first candidate, the most challenging one, because it assumes automatic evaluation. This means SFox is a promising approach to performing an automatic evaluation without human intervention in the assignment process.

The structure-based approach alone does not produce satisfactory results in the investigated problem, below 0.1 in all three experiments (omitted in the analysis). This is because the range of possibilities to match is extensive, and several options have the same similarity but are unrelated. The problem is partly related to the fact that, in some cases, more than one first-level category in Yelp had to be matched with one in Foursquare. For instance, 'Shop & Service' in Foursquare were matched with 'Local Services' and 'Shopping' on Yelp because they refer to similar places. Thus, instead of regarding the results of this approach by itself, we applied it after selecting a set of candidates. This is a common practice; for instance, Deng et al. (2019) also used a similar strategy. This strategy helps the Levenshtein approach in experiment 1, being better than sBert; however, it does not help in experiments 2 and 3. With more candidates (as in experiments 2 and 3), the chances of adding noise due to the problems of structure-based similarity increase leading to errors, especially in a gray area of the decision space.

5.3. Results on integrating individuals and groups

We integrated categories from Foursquare and Yelp for Toronto using our datasets for these cities, described above. After this integration, all categories in Foursquare were translated into a corresponding one in Yelp; here, we followed the automatic version of the approach SFox described in Section 4.2. After that, we identified groups following the method described in Section 4.3.

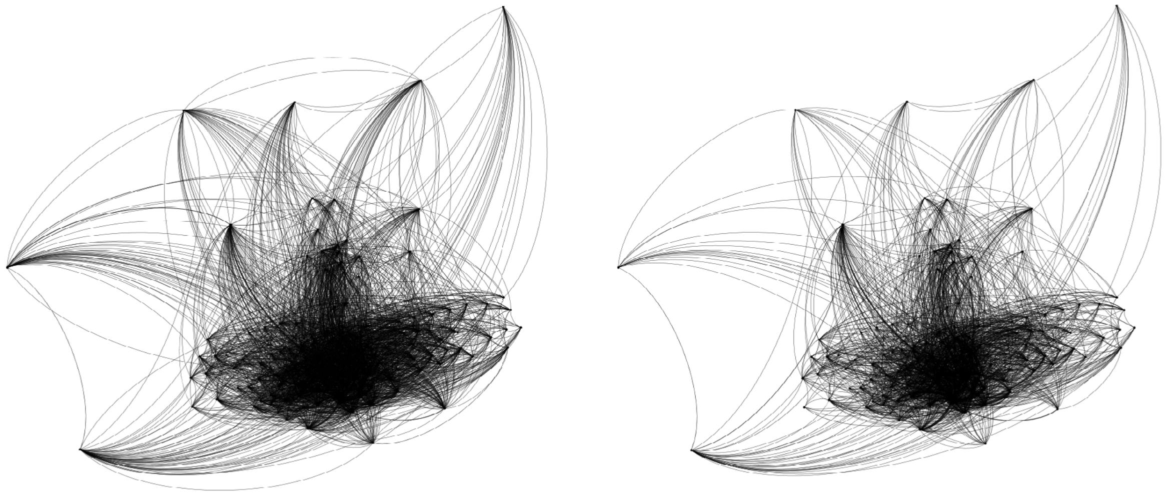
Once we have groups, we can perform the evaluation of interest. For example, one might be interested in confirming if the mobility pattern on census tracts is similar regarding the same group in different systems. Taking this as an example, we identified two groups using Foursquare and Yelp data. We focus on Group 1, which contains more users from both systems – 506 from Foursquare and 663 from Yelp. Next, we created two mobility networks, one for each system, representing the mobility of Group 1. In these networks, nodes are census tracts vi — those with at least 30 venues. Undirected edges connect all visited tracts by a particular user. Weights are the number of links between two different tracts considering all users in the analysis — in our example, those that made at least 10 Foursquare check-ins or Yelp reviews.

Table 2
Pearson correlation values between edges weights regarding the mobility of Group 1 for each tract on Foursquare and Yelp. For example, tract "0041.00" has a correlation of 0.63, meaning that the edge weights from this tract performed by Group 1 in Foursquare and Yelp mobility networks present a high positive relationship. Bold values have p-value smaller than 0.05.

0041.00 = 0.63	0098.00 = 0.14	**0122.00 = 0.36**	**0136.02 = 0.58**
0032.00 = 0.41	**0063.01 = 0.81**	0020.00 = 0.06	0072.01 = 0.09
0064.00 = 0.28	**0127.00 = 0.45**	0097.01 = 0.14	0096.00 = 0.13
0247.01 = 0.58	**0129.00 = 0.27**	**0069.00 = 0.37**	0046.00 = 0.02
0021.00 = 0.35	0126.00 = 0.15	**0030.00 = 0.18**	**0088.00 = 0.60**
0376.06 = 0.01	0081.00 = 0.08	0082.00 = 0.07	**0051.00 = 0.23**
0045.00 = 0.34	**0194.04 = 0.29**	0053.00 = 0.14	**0010.01 = 0.49**
0091.01 = 0.60	0100.00 = 0.08	**0023.00 = 0.51**	**0011.00 = 0.86**
0099.00 = 0.13	**0091.02 = 0.26**	0121.00 = 0.18	**0008.00 = 0.77**
0090.00 = 0.29	0095.00 = 0.08	**0040.00 = 0.70**	0087.00 = 0.15
0038.00 = 0.76	**0043.00 = 0.77**	**0105.00 = 0.34**	**0070.00 = 0.33**
0058.00 = 0.41	**0017.00 = 0.74**	**0014.00 = 0.87**	0068.00 = −0.07
0035.00 = 0.71	**0028.00 = 0.56**	0071.00 = 0.10	**0263.02 = 0.42**
0012.02 = 0.82	**0055.00 = 0.63**	**0037.00 = 0.76**	**0128.02 = 0.52**
0047.02 = 0.29	**0039.00 = 0.82**	**0136.01 = 0.35**	**0308.01 = 0.27**
0026.00 = 0.30	**0101.00 = 0.32**	**0052.00 = 0.34**	**0286.00 = 0.72**
0116.00 = 0.07	**0010.02 = 0.83**	**0018.00 = 0.36**	**0048.00 = 0.42**
0089.00 = 0.77	0103.00 = 0.13	**0034.02 = 0.83**	**0016.00 = 0.61**
0066.00 = 0.14	0141.02 = 0.03	**0036.00 = 0.85**	**0042.00 = 0.44**
0114.00 = 0.05	0080.01 = −0.06	0128.03 = −0.04	**0080.02 = 0.35**
0059.00 = 0.69	**0044.00 = 0.75**	**0050.02 = 0.27**	0074.00 = 0.13
0135.00 = 0.62	**0060.00 = 0.55**	0073.00 = 0.04	**0213.00 = 0.46**
0057.00 = 0.56	0142.00 = 0.05	**0029.00 = 0.59**	**0085.00 = 0.31**
0104.00 = 0.42	0050.01 = 0.00	0124.00 = 0.19	0079.00 = 0.12
0092.00 = 0.44	0047.01 = 0.13	**0001.00 = 0.56**	**0002.00 = 0.37**
0005.00 = −0.09	**0083.00 = 0.26**	0019.00 = 0.01	**0062.02 = 0.70**
0022.00 = 0.39	0134.00 = 0.10	**0137.00 = 0.28**	**0034.01 = 0.43**
0015.00 = 0.78	**0061.00 = 0.38**	**0094.00 = 0.23**	**0311.06 = 0.53**
0007.02 = 0.52	**0024.00 = 0.34**	**0054.00 = 0.41**	**0063.02 = 0.72**
0106.00 = 0.30	0084.00 = 0.15	0303.00 = −0.02	**0056.00 = 0.59**
0062.01 = 0.57	**0013.00 = 0.77**	0139.00 = 0.09	0110.00 = −0.04
0027.00 = 0.18	**0093.00 = 0.74**	0113.00 = 0.20	

These networks are shown in Fig. 3; nodes respect the central geographical coordinates of the tract they represent. Visually, they look quite similar. To quantify this similarity, for each node v_i, we extract $Foursquare_{vi} = \{Fw_{v1}, Fw_{v2}, \ldots, Fw_{vn}\}$ and $Yelp_{vi} = \{Yw_{v1}, Yw_{v2}, \ldots, Yw_{vn}\}$ representing edge weights from vi for all other remaining n nodes in the Foursquare (Fw) and Yelp (Yw) mobility network, respectively; self-loops are disregarded. In possession of $Foursquare_{vi}$ and $Yelp_{vi}$, we compute the Pearson correlation among those two vectors. Table 2 presents the correlation values for all tracts studied in Toronto. For example, tract "0041.00" has a correlation of 0.63, meaning that the edge weights from this tract performed by Group 1 in Foursquare and Yelp mobility networks present a high positive relationship.

Note that virtually all correlations are positive — those that are not are very close to 0. Note also that most of the correlations are high and significant with at least 95% of confidence (bold values). This result helps to validate that Group 1 represents similar users in terms of mobility in both systems. Thus, this illustrated integration enables understanding city dynamics in a way that would be hard (if possible) looking at just one system. For example, in the Yelp dataset, we do



(a) Foursquare

(b) Yelp

Fig. 3. Mobility network for Group 1 in different social media.

not have the gender information of users, but we do on Foursquare. In our running example, we know that in Group 1, approximately 41% are women, disregarding 20 that did not have gender information. The top 10 categories in the most important topic in Group 1 are “Restaurants”, “Food”, “Shopping”, “Nightlife”, “Bars”, “Coffee/Tea”, “Beauty Spas”, “Breakfast/Brunch”, “Canadian New”, and “Bakeries”, helping to understand what type of places attract more attention of this group.

6. Social Media Integration Ontology

We first define the Social Media Integration Ontology (SMIO) by what it is not. The primary focus of previous related efforts is to capture user interactions with a social media site, the actions they perform, such as posting, replying, etc., and the content to which their actions are applied. SMIO ontology aims to provide a semantically precise representation to map venues, individuals, groups, and activities. This way, we can have a rich representation of venues by integrating information from multiple social media sites. User interactions are less important and are represented sufficiently by SIOC (Bojārs et al., 2008) and SNOWL (Sebei et al., 2020).

SMIO also seeks to represent the Formemes embedded in social media information. It aims to make explicit the patterns of interactions among groups, activities, and forms. Therefore, SMIO must support the identification of Formemes that emerge from this data. For example, to support a tourism use case by identifying frequently occurring combinations of forms (venues), groups, and activities as potential places of interest.

In the remainder of this section, we present the ontology pattern for each of Form, Group, and Activity.

6.1. Form ontology

According to the identified use cases and competency questions derived from them, SMIO has a central goal of providing a rich representation of venues. Integrating information from different social media sites is an essential step for that. SMIO inspires on the concept of Formeme (Silver, Adler et al., 2022) to capture venue information. Formemes have forms as central elements because they represent physical features where groups of people perform activities.

Physical place/venue description is an important representation in the abstraction intended by SMIO. The first way SMIO does that is by representing raw data from social media sites, such as names, phone numbers, pictures, and categories of venues. It also goes beyond that, representing more sophisticated information derived from raw data to help comprehend people’s profiles/interests who visit specific venues/places. In this way, SMIO enables the answer to more complex questions. Form Ontology is described in Table 3.

6.2. Group ontology pattern

Groups can contain one or more persons and/or other groups, resulting in having groups of groups, where the subgroups are identifiable. Formally, a group $g \subseteq \mathcal{P}(G)$ where $\mathcal{P}(G)$ is the power set of G. A person can be a member of more than one group, just as a group can be a member of more than one (larger) group.

Groups perform activities, have interests, and have roles. Roles are defined by the social media the group participates in. They have members that can be other groups and persons. They can be members of one or more other groups.

We define a group as presented in Table 4. We define a person as in Table 5. A Role, Table 6, defines how a Person or Group interacts with a Social Medium.

6.3. Activity ontology

An Event in SMIO is equivalent to a Formeme in the sense that it reflects the occurrence of an activity, situated at a form, performed or attended by a group, over some time interval. In addition, it specifies the broader spatial location where the event occurs. The following describes the properties of an Event (see Table 7):

- **hasForm:** specifies the Form where the Event took place. More than one form can be specified.
- **hasGroup:** specifies the Group that participated in the Event. More than one group can be specified.
- **hasActivity:** specifies the Activity that was performed in the Event. More than one activity can be specified.
- **time:hasTime:** specifies a time point or interval at or during which the event occurred.

Table 3
Form class definition.

Class	Property	Value restriction
FoursquareFormCategory	formCatID	exactly 1 xsd:string
YelpFormCategory	formCatID	exactly 1 xsd:string
Form	formID	exactly 1 xsd:string
YelpForm	rdfs:subClassOf	Form
	hasAddress	only ic:Address
	ic:hasOperatingHours	Min 0 ic:HoursOfOperation
	ofSameCorporation	only YelpForm
	hasTip	Only YelpTip
	hasCheckins	only YelpCheckin
	hasReview	only YelpReview
	hasRating	Only YelpRating
	isOpen	Max 1 xsd:boolean
	hasYelpParking	Only YelpParking
	takeout	Max 1 xsd:boolean
	hasName	Only name
	overalFormRating	Max 1 xsd:float
	numberReviews	Max 1 xsd:integer
	hasYelpCategory	Only YelpFormCategory
YelpParking	rdfs:subClassOf	Parking
	street	Max 1 xsd:boolean
	garage	Max 1 xsd:boolean
	valet	Max 1 xsd:boolean
	lot	Max 1 xsd:boolean
	validated	Max 1 xsd:boolean
FoursquareForm	rdfs:subClassOf	Form
	hasAddress	only ic:Address
	ofSameCorporation	only FoursquareForm
	ic:hasOperatingHours	Min 0 ic:HoursOfOperation
	hasTip	only FoursquareTip
	hasCheckins	only FoursquareCheckin
	hasPhoto	only FoursquarePhoto
	hasName	Only Name
	hasEmail	min 0 xsd:string
	hasTelefone	min 0 ic:Phone Number
	description	only xsd:string
	formUrl	Max 1 xsd:string
	formVerified	Max 1 xsd:boolean
	menu	only xsd:string
	createdAt	Max 1 xsd:integer
	mayor	Max 1 Foursquarer
	formShortUrl	Max 1 xsd:string
	overalFormRating	Max 1 xsd:float
	numberTips	Max 1 xsd:integer
	numberPhotos	Max 1 xsd:integer
	numberLikes	Max 1 xsd:integer
	hasPrice	Max 1 xsd:integer
	currentMayor	Max 1 sxd:string
	formCanonicalUrl	Max 1 xsd:string
	formBestPhoto	Max 1 xsd:string
	formAtributes	only xsd:string
	hasFoursquareCategory	Only FoursquareFormCategory
TwitterForm	hasTweet	only TwitterTweet
	hasVideo	Only TwitterVideo
	ID	max 1 xsd:string
	url	Max 1 xsd:string
	placeType	Only xsd:string
	hasName	Only Name
	hasAddress	only ic:Address
	hasBoundingbox	Only 1 FormBounding
FormBounding	coordinates	Only Array of Array of Array of Float
	type	Only xsd:string
Name	nameAcronym	max 1 xsd:string
	nameLanguage	max 1 xsd:string
	namePhonetic	max 1 rdf:PlainLiteral
	nameType	max 1 xsd:string
	nameValue	max 1 xsd:string

Table 4
Group class definition.

Class	Property	Value restriction
Group	performsActivity hasInterest hasRole org:memberOf org:hasMember	only Activity only Interest only Role only Group only (Group or Person)

Table 5
Person class definition.

Class	Property	Value restriction
Person	foaf:firstName foaf:lastName performsActivity hasInterest hasRole org:memberOf ic:hasAddress ic:hasPhoneNumber	only xsd:string only xsd:string only Activity only Interest only Role only Group only ic:Address only ic:PhoneNumber

Table 6
Role class definition.

Class	Property	Value restriction
Role	sch:id (SM ID) performedBy	exactly 1 xsd:string Only (Group or Person)

Table 7
Event class definition.

Class	Property	Value restriction
Event	rdfs:subClassOf hasForm hasGroup hasActivity time:hasTime hasLocation	Formeme only Form only Group only Activity exactly 1 time:ProperTimeInterval only geo:Feature

- **hasLocation:** specifies the location of the event a Geo Sparql Feature. More than one location can be specified.

The activities we refer to are not the activities of check-in, posting, etc., but the activities that occur at a venue (in the physical world). The categories used by social media to classify venues can be used as a proxy for the activities that occur at a venue, as these categories attribute activity-related characteristics to the venue (Quercia, Aiello, & Schifanella, 2018).

For the basic representation of activity, we adopt the definition as found in the TOVE Activity ontology (Gruninger & Fox, 1994; Katsumi & Fox, 2017), which has been reproduced in the ISO/IEC 5087-1 "City Data Model: Part 1 Foundation Level Concepts" standard currently under development by the ISO/IEC JTC1 Working Group 11 on Smart Cities.

The core of the activity class is the activity cluster which consists of an activity connected to an enabling and caused state, each of which may be a state tree that defines complex states via decomposition into conjunctions and disjunctions of states — see Fig. 4.

Table 8 reproduces the definition of Activity and State for reference. Note that the time prefix denotes the OWL-Time ontology (http://www.w3.org/2006/time).

An Activity represents the most general description of an action that is captured by social media. Actions are divided into two subclasses (see Table 9):

- **MediaActivity**, which captures the actions of the social media member in the context of a social medium. This includes posting, commenting, etc., within a social medium.

Table 8
Activity and State definitions.

Object	Property	Value restriction
Activity	hasSubactivity enabledBy causes hasResource	only Activity only State only State only Resource
State	enables causedBy achievedAt	only Activity only Activity only time:TemporalEntity
TerminalState	subClassOf disjointWith hasResource	State NonTerminalState max 1 Resource
NonTerminalState	subClassOf disjointWtih hasSubstate	State TerminalState only State
ConjunctiveState	subClassOf disjointWith	NonTerminalState DisjunctiveState
DisjunctiveState	subClassOf disjointWith	NonTerminalState ConjunctiveState
ConsumeState	subClassOf hasResource	TerminalState exactly 1 Resource
ProduceState	subClassOf hasResource	NonTerminalState exactly 1 Resource
UseState	subClassOf hasResource	NonTerminalState exactly 1 Resource
ReleaseState	subClassOf hasResource	NonTerminalState exactly 1 Resource

Table 9
MediaActivity and Form Activity definitions.

Class	Property	Value restriction
Activity	rdfs:subClassOf	act:Activity
MediaActivity	rdfs:subClassOf disjointWith hasPerson	Activity FormAction exactly 1 Person
FormActivity	rdfs:subClassOf hasGroup	Activity only Group

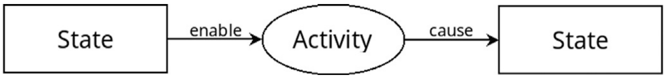


Fig. 4. Diagram expressing the relation between State and Activity.

- **FormActivity**, which captures the action performed by some individual or group, such as having dinner, riding the subway, etc., as captured by the social medium.

Media Activity - As described earlier, MediaActivity defines the taxonomy of social media-related actions that a member of a Social Media platform can perform. Where appropriate, we note which actions in SNOWL and SIOC they correspond to — see Table 10.

As can be seen above, a Post corresponds to a SIOC Post and SNOWL Content, but we provide a taxonomy of posts corresponding to the types found in various social media. The taxonomy is important as we are interested in the content of each — see Fig. 5.

Form Activity - FormActivity captures the activity performed by an individual or group, such as having dinner, riding the subway, etc., as captured by the social medium. A taxonomy of FormActivity is specified in Appendix C.

7. Final discussion and conclusion

As contributions of this study, we proposed several strategies for integrating social media data. In particular, we focus integration on

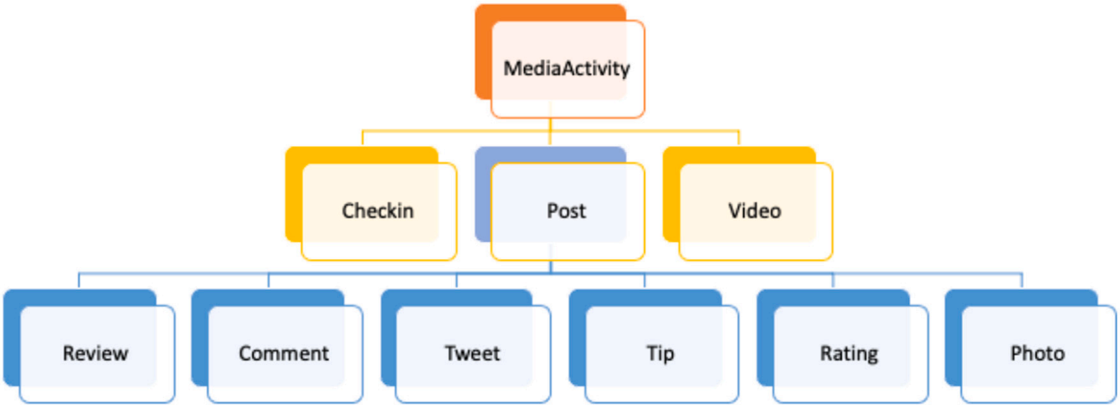


Fig. 5. MediaActivity taxonomy.

Table 10
MediaActivity taxonomy definitions.

Class	Property	Value restriction	SIOC	SNOWL
Post	rdfs:subClassOf	MediaActivity	Post	Content
Review	rdfs:subClassOf reviewText reviewID reviewLanguage	Post max 1 xsd:string max 1 xsd:string max 1 rdf:plainLiteral		
Tip	rdfs:subClassOf tipText tipID tipLanguage	Post max 1 xsd:string max 1 xsd:string max 1 rdf:plainLiteral		
Photo	rdfs:subClassOf photoText photoID photoURL	Post max 1 xsd:string max 1 xsd:string max 1 xsd:string		
Checkin	rdfs:subClassOf	MediaActivity		
Comment	rdfs:subClassOf commentText commentID commentLanguage	Post max 1 xsd:string max 1 xsd:string max 1 rdf:plainLiteral		
Video	rdfs:subClassOf videoID videoURL	MediaActivity max 1 xsd:string max 1 xsd:string		
Tweet	rdfs:subClassOf tweetText tweetID tweetLanguage	Post max 1 xsd:string max 1 xsd:string max 1 rdf:plainLiteral		
Rating	rdfs:subClassOf ratingValue ratingMaxLevel ratingMinLevel ratingLevel	Post exactly 1 xsd:double exactly 1 xsd:double exactly 1 xsd:double only xsd:double		

the perspective of (i) venues, improving state-of-the-art solutions, (ii) groups of users, providing the first matching based on groups, (iii) activity, providing an approach that explores semantically meaningful sentence embeddings associated with definitions of the terms composing the categories. As another contribution, we also present an ontology (SMIO) to support integrating data from different location-based social networks. SMIO is inspired by the recent concept of Formeme, which has form as a central element because it represents physical features where groups of people perform activities, thus valuable to capture venue information (Fox, Silver and Adler, 2022; Silver, Adler et al., 2022). This way, SMIO helps to achieve a semantically precise representation to map venues, individuals, groups, and activities. It supplies a rich representation of venues by integrating information from multiple social media sources.

The task of integrating data from multiple social media platforms is a challenge. The strategies presented here are valuable resources for integrating social media platforms that provide information from the physical world, such as LBSNs, where users can share and interact with entities representing physical locations. Our solutions could help foster more sophisticated solutions with the possibility of richer information due to data integration, for instance:

1. **Venue information enrichment.** If you only have Foursquare data, you do not have the information on venue ambiance or basic profile of users visiting the venue as we have on Yelp. If we only have Yelp data, we miss the level of check-in activity as we have in Foursquare.
2. **Improvement of venue labelling (semantic ambiguities reduction).** Venues can have different labels on different systems; an integration could help with labelling consolidation.
3. **More accurate ratings.** Integration helps to have a better idea of users' ratings and opinions. Different systems could represent particular types of users.

LBSNs offer solid data that can help improve understanding of different phenomena related to urban societies (Ferreira et al., 2020; Santala, Costa, Gomes, Gadda, & Silva, 2020; Silva et al., 2017; Skora, Senefonte, Delgado, Lüders, & Silva, 2022). Yet, it is essential to consider possible limitations in LBSN data. For instance, it may reflect the behaviour of a fraction of users, and data might be based on a limited sample of data (Silva et al., 2019). While our propositions do not solve those issues, they could be used in this direction in future work. For example, data quality, another possible issue in LBSN data, could be minimized with data integration in different sources, as exemplified above.

Our study opens up several other possible avenues for future work. For instance, other strategies regarding our approaches to integration could be proposed and evaluated. Take, for instance, our approach to integrating individuals and groups. We proposed a group matching strategy, but others could be contrasted against ours. The SMIO ontology could be expanded in several ways as well. For instance, FormActivity taxonomy is inspired by Quercia et al. (2018); nevertheless, this taxonomy could be extended according to specific application requirements.

CRediT authorship contribution statement

Thiago H. Silva: Conceptualization, Data curation, Methodology, Software, Analysis, Writing. **Mark S. Fox:** Conceptualization, Methodology, Ontology, Analysis, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was supported in part by a University of Toronto Connaught Global Challenge Award, and the School of Cities Urban Challenge Fund. This study is also partially supported by the project SocialNet (grant 2023/00148-0 from Sao Paulo Research Foundation - FAPESP, Brazil), and National Council for Scientific and Technological - CNPq, Brazil (grant 310998/2020-4), and CNPq GoodCity, Brazil (process 314603/2023-9).

Appendix A. Datasets explored in this study

We consider two datasets in this study – one from Yelp and the other from Foursquare – all for the city of Toronto, Canada, considering data from 2014 (the year of the intersection of both datasets).

Foursquare dataset: In the experiments, we explore the same Foursquare dataset of the studies (Senefonte et al., 2022; Silva et al., 2017). Foursquare is a location-based social network where users can use their smartphones to perform a check-in — the act of disclosing their current location to users in the system.

Each venue in Foursquare has a category with subcategories according to a hierarchical taxonomy provided by Foursquare. For instance, a given venue may have Food as a category with Pizza Place as its subcategory. A complete list of venue categories and subcategories in Foursquare is available on the Foursquare developers' website.[2]

Each check-in in the dataset comprises user ID, date and time of the check-in, location (latitude/longitude), venue ID, user gender, venue category, and venue subcategory. The dataset was collected in 2014. In Toronto, we have 16,295 check-ins from 2820 unique users in 4843 unique venues.

Yelp dataset: We use an official dataset provided by Yelp [https://www.yelp.com/dataset], explored in Silver and Silva (2023). Yelp is a location-based social network that publishes crowd-sourced reviews about businesses (venues). This public dataset represents a subset of Yelp venues, reviews, and user data from 2008 to 2018. We have information about each venue's category (e.g. "coffee shop") and all venue reviews. Yelp categories are also organized in a hierarchy, available on the Yelp developers' website.[3]

Each review is also indexed to the reviewer who wrote it. Finally, we have a venue ID, a user ID, and a location (latitude/longitude) associated with each review. For Toronto in 2014, we have 57,830 reviews, 19,179 unique users and 9.502 unique venues.

Appendix B. Other specifications

Table B.11 shows the prefixes used in the specifications.

Table B.11
Prefixes used in the study.

Prefix	URL
geo	http://www.opengis.net/ont/geosparql#
ic	http://ontology.eil.utoronto.ca/icontact#
time	http://www.w3.org/2006/time#

Table B.12
FormActivity taxonomy definitions.

Class	Property	Value restriction
Transport	rdfs:subClassOf	FormActivity
Driving	rdfs:subClassOf	Transport
	vehicleID	max 1 xsd:string
Commuting	rdfs:subClassOf	Transport
	commutingID	max 1 xsd:string
	commutingType	max 1 xsd:string
Eating	rdfs:subClassOf	FormActivity
Dining	rdfs:subClassOf	Eating
Cooking	rdfs:subClassOf	Eating
Shopping	rdfs:subClassOf	FormActivity
Market	rdfs:subClassOf	Shopping
Trading	rdfs:subClassOf	Shopping
Protest	rdfs:subClassOf	FormActivity
Riot	rdfs:subClassOf	Protest
Occupy	rdfs:subClassOf	Protest
Spiritual	rdfs:subClassOf	FormActivity
Funeral	rdfs:subClassOf	Spiritual
Praying	rdfs:subClassOf	Spiritual
WorkStudy	rdfs:subClassOf	FormActivity
Study	rdfs:subClassOf	WorkStudy
Teaching	rdfs:subClassOf	Study
Work	rdfs:subClassOf	WorkStudy
Office	rdfs:subClassOf	Work
Sports	rdfs:subClassOf	FormActivity
Teams	rdfs:subClassOf	Sports
Football	rdfs:subClassOf	Teams
Individual	rdfs:subClassOf	Sports
Gymnastics	rdfs:subClassOf	Individual
Running	rdfs:subClassOf	Sports
Jogging	rdfs:subClassOf	Running
Outdoors	rdfs:subClassOf	Sports
Climbing	rdfs:subClassOf	Outdoors
Shows	rdfs:subClassOf	FormActivity
Music	rdfs:subClassOf	Shows
Singer	rdfs:subClassOf	Music
Dance	rdfs:subClassOf	Shows
Ballroom	rdfs:subClassOf	Dance
Exhibitions	rdfs:subClassOf	Shows
Art museum	rdfs:subClassOf	Exhibitions
Performance	rdfs:subClassOf	Shows
Theater	rdfs:subClassOf	Performance
Costumes	rdfs:subClassOf	Shows
Cosplay	rdfs:subClassOf	Costumes
Public Speech	rdfs:subClassOf	Shows
Speaker	rdfs:subClassOf	Public Speech
Self	rdfs:subClassOf	FormActivity
Walking	rdfs:subClassOf	Self
Sightseeing	rdfs:subClassOf	Waking
Exploring	rdfs:subClassOf	Walking
HobbiesHomecare	rdfs:subClassOf	Self
Knitting	rdfs:subClassOf	HobbiesHomecare
Bathing	rdfs:subClassOf	HobbiesHomecare
Meet	rdfs:subClassOf	FormActivity
Partying	rdfs:subClassOf	Meet
Clubbing	rdfs:subClassOf	Partying
Housewarm	rdfs:subClassOf	Partying
Meeting	rdfs:subClassOf	Meet
Networking	rdfs:subClassOf	Meeting
Meetup	rdfs:subClassOf	Meeting
Sex	rdfs:subClassOf	FormActivity
Self-pleasure	rdfs:subClassOf	Sex
Love-making	rdfs:subClassOf	Sex

[2] https://developer.foursquare.com/docs/resources/categories.

[3] https://docs.developer.yelp.com/docs/resources-categories.

Appendix C. FormActivity taxonomy

Table B.12 presents how the FormActivity taxonomy is integrated with the SMIO ontology. This taxonomy is inspired by the Urban Activity Taxonomy proposed by Quercia et al. (2018), which is interesting because it was obtained through social media data — please refer to this publication for a visualization of the taxonomy. However, this taxonomy could be extended according to particular application needs.

References

Ansell, L., & Dalla Valle, L. (2023). A new data integration framework for Covid-19 social media information. *Scientific Reports*, *13*(1), 6170.

Ballatore, A., Bertolotto, M., & Wilson, D. C. (2015). A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS International Journal of Geo-Information*, *4*(2), 471–492.

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

Bojars, U., & Breslin, J. G. (2007). *SIOC core ontology specification. WC member submission*. World Wide Web Consortium (W3C).

Bojārs, U., Breslin, J. G., Finn, A., & Decker, S. (2008). Using the Semantic Web for linking and reusing data across Web 2.0 communities. *Journal of Web Semantics*, *6*(1), 21–28.

Čerba, O., & Jedlička, K. (2016). Linked Forests: Semantic similarity of geographical concepts "forest". *Open Geosciences*, *8*(1), 556–566.

Chen, Z., Song, J., & Yang, Y. (2018). An approach to measuring semantic relatedness of geographic terminologies using a thesaurus and lexical database sources. *ISPRS International Journal of Geo-Information*, *7*(3), 98.

Deng, Y., Luo, A., Liu, J., & Wang, Y. (2019). Point of interest matching between different geospatial datasets. *ISPRS International Journal of Geo-Information*, *8*(10), 435.

Ferreira, A. P., Silva, T. H., & Loureiro, A. A. (2020). Uncovering spatiotemporal and semantic aspects of tourists mobility using social sensing. *Computer Communications*, *160*, 240–252.

Fox, M. S., Silver, D., & Adler, P. (2022). Towards a model of urban evolution: Part II: Formal model. *Urban Science*, *6*(4), 88.

Fox, M. S., Silver, D., Silva, T., & Zhang, X. (2022). Towards a model of urban evolution part IV: Evolutionary (formetic) distance—An interpretation of yelp review data. *Urban Science*, *6*(4), 86.

Goga, O., Lei, H., Parthasarathi, S. H. K., Friedland, G., Sommer, R., & Teixeira, R. (2013). Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on world wide web* (pp. 447–458).

Gruninger, M., & Fox, M. S. (1994). *An activity ontology for enterprise modelling*. Department of Industrial Engineering, University of Toronto.

Hristova, D., Musolesi, M., & Mascolo, C. (2014). Keep your friends close and your facebook friends closer: A multiplex network approach to the analysis of offline and online social ties. In *Proceedings of the international AAAI conference on web and social media, Vol. 8* (pp. 206–215).

Hristova, D., Panzarasa, P., & Mascolo, C. (2015). Multilayer brokerage in geo-social networks. In *Proceedings of the international AAAI conference on web and social media, Vol. 9* (pp. 159–167).

Katsumi, M., & Fox, M. (2017). Defining activity specifications in OWL. In *WOP@ ISWC*.

Kim, J., Vasardani, M., & Winter, S. (2017). Similarity matching for integrating spatial information extracted from place descriptions. *International Journal of Geographical Information Science*, *31*(1), 56–80.

Levenshtein, V. I., et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady, Vol. 10* (pp. 707–710). Soviet Union.

Li, C., Liu, L., Dai, Z., & Liu, X. (2020). Different sourcing point of interest matching method considering multiple constraints. *ISPRS International Journal of Geo-Information*, *9*(4), 214.

McKenzie, G., Janowicz, K., & Adams, B. (2013). Weighted multi-attribute matching of user-generated points of interest. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 440–443).

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).

Mueller, W., Silva, T. H., Almeida, J. M., & Loureiro, A. A. (2017). Gender matters! analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science*, *6*, 1–21.

Piech, M., Smywinski-Pohl, A., Marcjan, R., & Siwik, L. (2020). Towards automatic points of interest matching. *ISPRS International Journal of Geo-Information*, *9*(5), 291.

Quercia, D., Aiello, L. M., & Schifanella, R. (2018). Diversity of indoor activities and economic development of neighborhoods. *PLoS One*, *13*(6), Article e0198441.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Rodrigues, D. O., Boukerche, A., Silva, T. H., Loureiro, A. A., & Villas, L. A. (2017). SMAFramework: Urban data integration framework for mobility analysis in smart cities. In *Proceedings of the 20th ACM international conference on modelling, analysis and simulation of wireless and mobile systems* (pp. 227–236).

Santala, V., Costa, G., Gomes, L., Jr., Gadda, T., & Silva, T. H. (2020). On the potential of social media data in urban planning: Findings from the beer street in Curitiba, Brazil. *Planning Practice & Research*, 1–16.

Santos, F. A., Silva, T. H., Loureiro, A. A., & Villas, L. A. (2020). Automatic extraction of urban outdoor perception from geolocated free texts. *Social Network Analysis and Mining*, *10*, 1–23.

Scerri, S., Cortis, K., Rivera, I., & Handschuh, S. (2012). Knowledge discovery in distributed social web sharing activities. In *# MSM* (pp. 26–33).

Sebei, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2020). SNOWL model: social networks unification-based semantic data integration. *Knowledge and Information Systems*, *62*(11), 4297–4336.

Senefonte, H. C. M., Delgado, M. R., Lüders, R., & Silva, T. H. (2022). Predictour: Predicting mobility patterns of tourists based on social media user's profiles. *IEEE Access*, *10*, 9257–9270.

Shu, K., Wang, S., Tang, J., Zafarani, R., & Liu, H. (2017). User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter*, *18*(2), 5–17.

Silva, T. H., De Melo, P. O. V., Almeida, J. M., Viana, A. C., Salles, J., & Loureiro, A. A. (2014). Participatory sensor networks as sensing layers. In *2014 IEEE fourth international conference on big data and cloud computing* (pp. 386–393). IEEE.

Silva, T. H., & Fox, M. (2021). Towards interoperability of social media: Venue matching by categories. In *XIV seminar on ontology research in Brazil* (pp. 126–137).

Silva, T. H., de Melo, P. O. V., Almeida, J. M., Musolesi, M., & Loureiro, A. A. (2017). A large-scale study of cultural differences using urban data about eating and drinking preferences. *Information Systems*, *72*, 95–116.

Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., et al. (2019). Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys*, *52*(1), 1–39.

Silver, D., Adler, P., & Fox, M. S. (2022). Towards a model of urban evolution—Part I: Context. *Urban Science*, *6*(4), 87.

Silver, D., Fox, M. S., & Adler, P. (2022). Towards a model of urban evolution—Part III: Rules of evolution. *Urban Science*, *6*(4), 89.

Silver, D., & Silva, T. H. (2023). Complex causal structures of neighbourhood change: Evidence from a functionalist model and yelp data. *Cities*, *133*, Article 104130.

Skora, L. E., Senefonte, H. C., Delgado, M. R., Lüders, R., & Silva, T. H. (2022). Comparing global tourism flows measured by official census and social sensing. *Online Social Networks and Media*, *29*, Article 100204.

Sun, K., Hu, Y., Song, J., & Zhu, Y. (2021). Aligning geographic entities from historical maps for building knowledge graphs. *International Journal of Geographical Information Science*, *35*(10), 2078–2107.

Sun, K., Zhu, Y., & Song, J. (2019). Progress and challenges on entity alignment of geographic knowledge bases. *ISPRS International Journal of Geo-Information*, *8*(2), 77.

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Tsutsumi, D. P., Fenerich, A. T., & Silva, T. H. (2019). Towards business partnership recommendation using user opinion on Facebook. *Journal of Internet Services and Applications*, *10*, 1–23.

Vijaymeena, M., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, *3*, 19–28.

Zafarani, R., & Liu, H. (2013). Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 41–49).

Zafarani, R., Tang, L., & Liu, H. (2015). User identification across social media. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *10*(2), 1–30.

Zhang, Y., Tang, J., Yang, Z., Pei, J., & Yu, P. S. (2015). Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1485–1494).

Zheng, Y., Fen, X., Xie, X., Peng, S., & Fu, J. (2010). Detecting nearly duplicated records in location datasets. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 137–143).

Zhou, J., & Fan, J. (2019). Translink: User identity linkage across heterogeneous social networks via translating embeddings. In *IEEE INFOCOM 2019-IEEE conference on computer communications* (pp. 2116–2124). IEEE.

Zhu, Y., Zhu, A.-X., Song, J., Yang, J., Feng, M., Sun, K., et al. (2017). Multidimensional and quantitative interlinking approach for linked geospatial data. *International Journal of Digital Earth*, *10*(9), 923–943.