KnowSite: Leveraging Urban Knowledge Graph for Site Selection

Yu Liu
liuyu2419@126.com
BNRist, Department of Electronic Engineering, Tsinghua University
Beijing, China

Jingtao Ding*
dingjt15@tsinghua.org.cn
BNRist, Department of Electronic Engineering, Tsinghua University
Beijing, China

Yong Li
liyong07@tsinghua.edu.cn
BNRist, Department of Electronic Engineering, Tsinghua University
Beijing, China

ABSTRACT

Site selection determines optimal locations for new stores, which is of crucial importance for business success and urban development. Especially, the wide application of artificial intelligence with multi-source urban data makes intelligent site selection promising. Nevertheless, existing data-driven approaches heavily rely on feature engineering, which cannot take the complex relationships as well as the diverse influences of various semantics among data into consideration. Further, most approaches fail to reveal underlying factors for site decisions. To get rid of the dilemma, in this work, leveraging the knowledge graph (KG) technique, we propose a knowledge-driven model for site selection, short for KnowSite. Specifically, by empowering rich semantics in KG, we firstly construct an urban KG (UrbanKG) for site selection knowledge discovery with cities' key elements and complex relationships captured. Based on UrbanKG, we apply pre-training for semantic representations, and then design a generalized encoder-decoder structure for site decisions. KnowSite designs a graph neural network based encoder to adaptively model diverse influences, and further builds a relation path based decoder revealing the reasons behind site decisions. Extensive experiments on two datasets demonstrate that KnowSite outperforms representative baselines by more than 9% on precision. Moreover, KnowSite provides intuitive and convincing explanations for site decisions and sheds light on the site selection understanding.

KEYWORDS

Site selection, knowledge graph, relation path, graph neural network.

ACM Reference Format:
Yu Liu, Jingtao Ding, and Yong Li. 2023. KnowSite: Leveraging Urban Knowledge Graph for Site Selection. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3589132.3625640

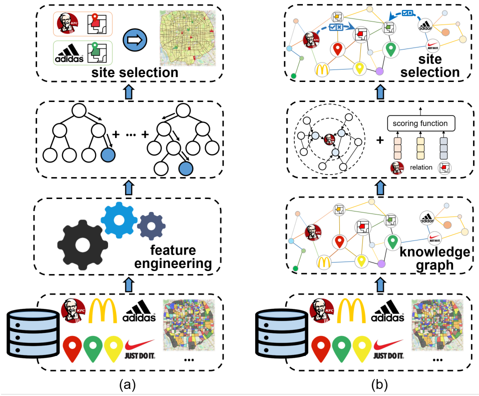


Figure 1: Illustration of (a) data-driven paradigm and our proposed (b) knowledge-driven paradigm for site selection.

1 INTRODUCTION

The task of site selection, which selects optimal locations for opening new stores, is of crucial importance for business success. A good choice of location always brings substantial profits while an inappropriate one could lead to store closure, such as opening a Starbucks store in a business area versus a residential one. Generally, site selection for a specific brand requires a comprehensive consideration of both its own characteristics and potential urban regions, e.g., the brand's category and the region's human flow and function. Traditional solution for most corporations is to employ expert consultants and conduct manual surveys [4, 17, 27, 32], which are expensive, labor-intensive, and time-consuming.

Owing to the rapid development of location-based services [15] and wide availability of multi-source urban data [55], recent studies introduce the data-driven paradigm for site selection [16, 18, 24, 42, 43]. As shown in Figure 1(a), these data-driven approaches typically extract various features from the multi-source urban data, which are then fed into a machine learning model like XGBoost [6] to calculate the score for site decision. However, the manually defined feature involves just one or two aspects (store density, human flow, etc.), failing to exploit complex relationships as well as diverse influences among the multi-source urban data. Moreover, such approaches merely provide an importance score for each feature without underlying factors, which is insufficient to persuade corporations [32, 48].

In comparison to the site selection still in data-driven paradigm, several other areas of artificial intelligence have further introduced knowledge-driven paradigm for superior performance, such as question answering [13], natural language understanding [53] and recommender systems [10]. The core of such knowledge-driven

[*] Corresponding Author.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0168-9/23/11…$15.00
https://doi.org/10.1145/3589132.3625640

paradigm is knowledge graph (KG) [11]. With domain entities as nodes and semantic relations as edges, KG could integrate multi-source data into a graph structure, and then powerful knowledge representation learning (KRL) methods are developed to avoid complex feature engineering [14]. Hence, knowledge-driven paradigm stands out as a promising solution for site selection, but it is still under-explored due to following three critical challenges:

- **The difficulty of knowledge discovery from multi-source urban data.** The target knowledge for site selection lies in complex relationships among multi-source urban data, e.g., attribute, affiliation, spatiality, mobility and etc., which increases the difficulty to discover the structured knowledge from such data.
- **The complexity of knowledge refinement for diverse influences.** The influences of various knowledge are diverse for site selection, e.g., for KFC opening stores, the site decision of McDonald's has more reference value than store density indicator at regions. Thus, refining task-specific knowledge is non-trivial considering the rich while diverse urban contexts.
- **The necessity of knowledge explainability to site decision understanding.** Although feature importance is provided in data-driven paradigm [16, 42], the reasons for site decisions remain unknown, e.g., finding new sites that have significant flow transition with existing sites. Thus a challenge is how to clearly explain the underlying factors behind corresponding site decisions for convincing and practical applications.

To overcome the above challenges as well as explore the potential capability of KG, in this paper, we propose a generalized knowledge-driven paradigm for site selection. As shown in Figure 1(b), we first construct the KG from multi-source urban data (referred to as UrbanKG), based on which a generalized encoder-decoder structure is proposed for site selection. Specifically, knowledge discovery is achieved in UrbanKG, i.e., the key elements of the city such as regions, point of interests (POIs), corporation brands, etc. are identified as entities, while their complex relationships on attribute, affiliation, spatiality, mobility, etc. are modeled as relations. To obtain semantic representations for entity and relation initialization, we adopt pre-training techniques on UrbanKG. Furthermore, we design a graph neural network (GNN) based encoder on UrbanKG, such that knowledge refinement for diverse influences is adaptively modeled via multi-relational message passing. As for the decoder part, we design a relation path based scoring function for knowledge explainability, which measures the plausibility of site decisions between corporation brands and regions with the underlying factors revealed. The scoring function firstly introduces multiple multi-hop relation paths based on different site selection criteria, then generates relation path representations via semantic composition of relations, and finally obtains corresponding scores using the attention mechanism. The overall model is termed as KnowSite for Knowledge-driven Site selection. Our key contributions are summarized as follows:

- We are the first to propose the knowledge-driven paradigm for site selection, and propose a model KnowSite generalized for various types of businesses. Especially, KnowSite leverages urban knowledge via KG, and builds an encoder-decoder structure to explore the knowledge for effective and explainable site selection.
- We conduct a systematic study of knowledge discovery from multi-source urban data via KG construction, which identifies key elements and complex relationships in the city as entities and relations, respectively.
- Under the proposed encoder-decoder structure, we design a multi-relational message passing mechanism with GNN based encoder for knowledge refinement, and develop multi-hop relation path based decoder, which achieves knowledge explainability with the reasons behind site decisions.
- We conduct extensive experiments on two real-world datasets and the proposed KnowSite outperforms state-of-the-art data-driven approaches by more than 18% on precision, which demonstrates the effectiveness of knowledge-driven paradigm. Further visualization results and dynamic case study shed light on understanding critical mechanism behind different brands' site decisions as well as demonstrate high practicality in terms of efficiency and scalability.

The rest of this paper is organized as follows. Section 2 introduces the research problem, while Section 3 presents the details of our proposed knowledge-driven framework. The empirical results are discussed in Section 4. We review the related works in Section 5, followed by a conclusion in Section 6.

2 PROBLEM STATEMENT

Typically, the multi-source urban data for site selection can be categorized into three aspects [9, 16, 24].

Spatial Data. They include the road network data $\mathcal{D}_{\text{RN}}$ and business area (Ba) data $\mathcal{D}_{\text{Ba}}$. $\mathcal{D}_{\text{RN}}$ is a collection of road segments connecting each other and $\mathcal{D}_{\text{Ba}}$ collects core areas of business and commercial activities, e.g., *Sanlitun*[1] in Beijing, China.

Store Data. They include the POI data $\mathcal{D}_{\text{POI}}$, brand data $\mathcal{D}_{\text{Brand}}$ and site selection data $\mathcal{D}_{\text{Site}}$. $\mathcal{D}_{\text{POI}}$ and $\mathcal{D}_{\text{Brand}}$ are the collection of venues and corporation brands respectively in the city. $\mathcal{D}_{\text{Site}}$ is a collection of brand-region pairs for site selection.

User Behavior Data. They include trajectory data $\mathcal{D}_{\text{Traj}}$ with user trajectories, check-in data $\mathcal{D}_{\text{Check}}$ with users' self-reported check-in records and click data $\mathcal{D}_{\text{Click}}$ of aggregated clicking POIs records using map services.

Then we state the knowledge-driven site selection problem.

Problem 1. **Knowledge-driven Site Selection Problem.** *Given the multi-source urban data, the knowledge-driven site selection problem can be divided into two sub-problems of KG construction and site selection. The KG construction sub-problem requires to construct KG $\mathcal{G} = f(\mathcal{D}_{\text{RN}}, \mathcal{D}_{\text{Ba}}, \mathcal{D}_{\text{POI}}, \mathcal{D}_{\text{Brand}}, \mathcal{D}_{\text{Site}}, \mathcal{D}_{\text{Traj}}, \mathcal{D}_{\text{Check}}, \mathcal{D}_{\text{Click}})$ with construction method f. Then the site selection sub-problem is formulated as a link prediction problem on $\mathcal{G}$, predicting if there exists a site decision link between brand b and region a, i.e., $(b, ?, a)$.*

3 METHODOLOGY

3.1 Framework Overview

To overcome the challenges of applying knowledge-driven paradigm for site selection, we present the framework of our proposed method in Figure 2, including UrbanKG construction and

[1] https://en.wikipedia.org/wiki/Sanlitun

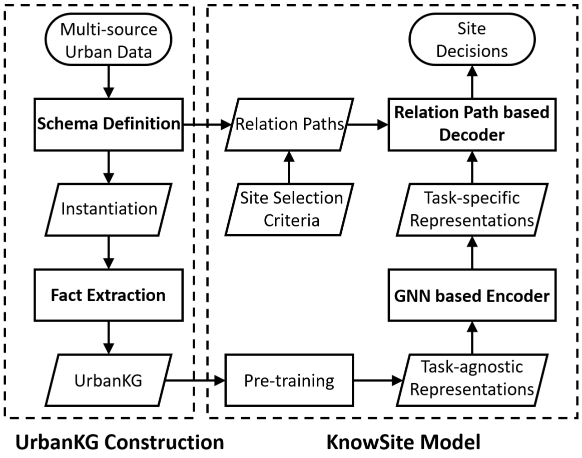
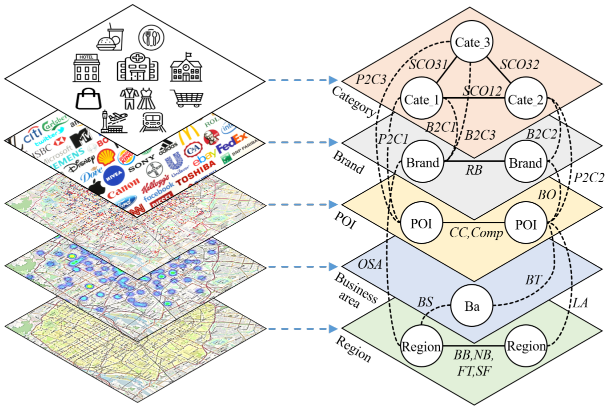


Figure 2: The framework of our proposed knowledge-driven site selection methodology.



(a) Multi-source urban data (b) Schema

Figure 3: The schema of urban knowledge graph. In (b), the dash lines represent inter-ontology relations and the solid lines for intra-ontology ones.

the KnowSite model for KG construction and site selection subproblems, respectively. Specifically, to discover knowledge from multi-source urban data, we firstly construct UrbanKG for structured urban knowledge, which is comprised of two major components: schema definition and fact extraction. As for the KnowSite model, we exploit the pre-training on UrbanKG for task-agnostic but knowledgeable representations. To further refine knowledge for diverse influences, we propose a GNN based encoder with task-specific representations learned. Finally, to make knowledge explainable and identify reasons behind site decisions, we design a relation path based decoder with effective performance achieved.

3.2 UrbanKG Construction

To discover knowledge from multi-source urban data, we construct UrbanKG for structured urban knowledge. Formally, a KG is defined as a graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$, where $\mathcal{E}$ is the node set of entities and $\mathcal{R}$ is the edge set of relations, while $\mathcal{F}$ corresponds to the fact set $\{(s, r, o) \mid s, o \in \mathcal{E}, r \in \mathcal{R}\}$ [14, 40]. The triplet (s, r, o) denotes the directional edge from node s to node o via the edge of relation r.

3.2.1 Schema Definition. At first, by investigating the multi-source urban data as well as expert knowledge from urban computing and planning [23, 26, 49, 55], we build the schema of UrbanKG, as shown in Figure 3. It defines the high-level structure for UrbanKG with ontologies and relations [11], where the ontologies determine the types of entities, including key elements in cities, i.e., *Region*, *Ba*, *POI*, *Brand* and *Category*, mainly identified from $\mathcal{D}_{\text{RN}}$, $\mathcal{D}_{\text{Ba}}$, $\mathcal{D}_{\text{POI}}$ and $\mathcal{D}_{\text{Brand}}$. Since the category is an important property of POIs and brands, we further divide the category into coarse-level, mid-level, and fine-grained categories, referred to as Cate_1/2/3.

Moreover, we identify the underlying relations to capture the complex relationships among city elements, as presented in Table 1. For intra-ontology relations, we describe them layer by layer, from bottom to up in Figure 3(b). At the first layer of Region, *BorderBy* and *NearBy* define the spatial relationships of two regions, while *SimilarFunction* link regions with similar POI distributios. By analyzing $\mathcal{D}_{\text{Traj}}$, we devise *FlowTransition* to link regions with significant crowd flow transitions. At POI layer, based on $\mathcal{D}_{\text{Check}}$, *CoCheckin* reveals the geographical influence among POIs with check-in concurrence [5] and *Competitive* models the competitive relationship among POIs [19]. At Brand layer, *RelatedBrand* describes relatedness of brands. At Category layer, *SubCateOf_ij* defines the taxonomy among three-level categories. As for inter-ontology relations, *BaServe*, *BelongTo* and *LocateAt* define the spatial relationships between different ontologies, especially *BaServe* describes regions are in service range of business area. Moreover, *POIToCate_i* and *BrandToCate_i* represent the attribute relationships, while *BrandOf* describes the affiliation relationship between POI and brand. *OpenStoreAt* represents site selection records in $\mathcal{D}_{\text{Site}}$. Besides, for asymmetric relations $\{r \in \mathcal{R} \mid (s, r, o) \nLeftrightarrow (o, r, s), \forall (s, r, o) \in \mathcal{F}\}$, we introduce a new inverse relation r' into UrbanKG schema.

Table 1: The details of defined relations in UrbanKG.

Relation	Abbrev.	Subject & Object Ontologies	Symmetry	Data Source
BorderBy	*BB*	(Region, Region)	✓	$\mathcal{D}_{\text{RN}}$
NearBy	*NB*	(Region, Region)	✓	$\mathcal{D}_{\text{RN}}$
FlowTransition	*FT*	(Region, Region)	✗	$\mathcal{D}_{\text{Traj}}$
SimilarFunction	*SF*	(Region, Region)	✓	$\mathcal{D}_{\text{POI}}$, $\mathcal{D}_{\text{RN}}$
CoCheckin	*CC*	(POI, POI)	✓	$\mathcal{D}_{\text{Check}}$, $\mathcal{D}_{\text{POI}}$
Competitive	*Comp*	(POI, POI)	✓	$\mathcal{D}_{\text{Brand}}$, $\mathcal{D}_{\text{POI}}$
RelatedBrand	*RB*	(Brand, Brand)	✓	$\mathcal{D}_{\text{Brand}}$
SubCateOf_ij	*SCOij*	(Cate_*i*, Cate_*j*)	✗	$\mathcal{D}_{\text{POI}}$
BaServe	*BS*	(Ba, Region)	✗	$\mathcal{D}_{\text{Ba}}$, $\mathcal{D}_{\text{RN}}$
BelongTo	*BT*	(POI, Ba)	✗	$\mathcal{D}_{\text{Ba}}$, $\mathcal{D}_{\text{POI}}$
LocateAt	*LA*	(POI, Region)	✗	$\mathcal{D}_{\text{POI}}$, $\mathcal{D}_{\text{RN}}$
POIToCate_i	*P2Ci*	(POI, Cate_*i*)	✗	$\mathcal{D}_{\text{POI}}$
BrandToCate_i	*B2C_i*	(Brand, Cate_*i*)	✗	$\mathcal{D}_{\text{Brand}}$, $\mathcal{D}_{\text{POI}}$
BrandOf	*BO*	(Brand, POI)	✗	$\mathcal{D}_{\text{Brand}}$, $\mathcal{D}_{\text{POI}}$
OpenStoreAt	*OSA*	(Brand, Region)	✗	$\mathcal{D}_{\text{Site}}$

3.2.2 Fact Extraction. Based on the defined schema above, we instantiate facts from the data, i.e., mapping ontologies to specific entities and linking entities via semantic relations. First, we introduce the mapping step. For mapping *Region* ontology, we partition the city into disjointed regions according to the main road network with $\mathcal{D}_{\text{RN}}$. Compared with grid partition of equal size [43], our partition is much closer to people's movement and urban functional units. For *Ba* and *POI* ontologies, we obtain the entities from $\mathcal{D}_{\text{Ba}}$ and $\mathcal{D}_{\text{POI}}$, respectively. For *Brand* ontology, we adopt a text segmentation tool[2] and name matching to obtain entities. For *Category* ontology, the three-level categories are divided by domain experts, e.g., Cate_3 entity *Beijing Cuisine* belongs to Cate_2 entity *Chinese Food* and Cate_1 entity *Food*. Then, in the second step, the entities are further linked via relations defined in Table 1 with corresponding data sources. Here we highlight the link details for brand-related relations. For *RelatedBrand*, the facts are obtained from a public KG zhishi.me with the "relatedPage" relation. For *BrandOf*, POI entities and their corresponding brand entities are linked together, based on which the *BrandToCate_i* facts are obtained by brands' connected POIs. Other relational links follow the definitions above, and are obtained by data mapping, aggregation and calculation methods. In this way, the constructed UrbanKG successfully presents the structured knowledge among multi-source urban data.

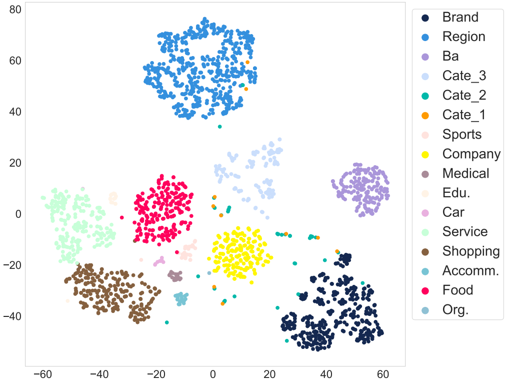
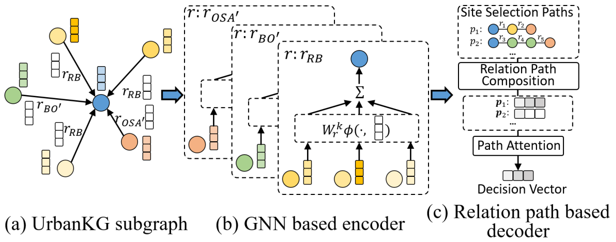


Figure 4: t-SNE of pre-trained entity embeddings of beijing's UrbanKG (better viewed in color). Ba denotes business area.

Using the data introduced in Section 4.1.1 later, we construct UrbanKGs for two of the largest cities, which contains over 20k/40k entities and 300k/500k triplet facts in Beijing/Shanghai. Note that the original UrbanKGs are significantly large, and we only report the statistics of subgraphs utilized in this work. It is worth mentioning that we utilize pre-training for knowledgeable representations of entities and relations. Specifically, we leverage the KRL model, TuckER [2] for pre-training, which measures the plausibility of triplets in UrbanKG with embeddings learned. Note that the pre-training process is task-agnostic and captures the global semantic information. To validate the representation capability of UrbanKG, we visualize the pre-trained entity embeddings using t-SNE. Especially, we randomly sample 1000 POI entities and all of other entities for visualization, showing in Figure 4. It can be observed that, entities of the same ontology are clustered in space. Moreover, POIs of different categories are also separated in visualization. Such results indicate the effectiveness of our constructed UrbanKG with the underlying semantics captured.



(a) UrbanKG subgraph (b) GNN based encoder (c) Relation path based decoder

Figure 5: The illustration of KnowSite model with a subgraph of UrbanKG.

3.3 The KnowSite Model

3.3.1 GNN based Encoder. To fully explore the potential of UrbanKG and model diverse influences of various knowledge, we design a GNN based encoder for knowledge refinement. For a node/entity v in KG $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$, d denotes the embedding dimension, $\boldsymbol{h}_v^k \in \mathbb{R}^d$ denotes its representation after k layers GNN, while $\mathcal{N}_v^r$ denotes its neighbors under relation $r \in \mathcal{R}$. The relation r's representation at layer k is denoted as $\boldsymbol{h}_r^k \in \mathbb{R}^d$. The number of GNN layers is denoted as K. Especially, the representation of node v at layer $k+1$, $\boldsymbol{h}_v^{k+1}$ is obtained via three steps [28, 35, 36, 50].

- **Message calculation**, which defines the function *MSG* to calculate the message for triplet (u, r, v): $m_{urv}^{k+1} = MSG(\boldsymbol{h}_u^k, \boldsymbol{h}_r^k, \boldsymbol{h}_v^k)$.
- **Message aggregation**, which defines the function *AGG* to aggregate messages from node v's neighbors: $M_v^{k+1} = AGG(m_{urv}^{k+1} | r \in \mathcal{R}, u \in \mathcal{N}_v^r)$.
- **Representation update**, which defines the function *UPD* to update v's representation from the aggregated messages M_v^{k+1} and v's previous layer representation $\boldsymbol{h}_v^k$: $\boldsymbol{h}_v^{k+1} = UPD(\boldsymbol{h}_v^k, M_v^{k+1})$.

In terms of message calculation, for a node v with the triplet (u, r, v), our proposed GNN based encoder adopts the composition of neighbor node and linked relation [28, 35]:

$$MSG\left(\boldsymbol{h}_u^k, \boldsymbol{h}_r^k, \boldsymbol{h}_v^k\right) = \boldsymbol{W}_r^k \phi\left(\boldsymbol{h}_u^k, \boldsymbol{h}_r^k\right), \quad (1)$$

where $\boldsymbol{W}_r^k$ is the relation-specific projection matrix, while $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the entity-relation composition operation, e.g., element-wise subtraction and element-wise multiplication.

Moreover, the message aggregation and the representation update are defined as relation-specific mean pooling and nonlinear transformation, respectively. Thus, the representation of node v at layer $k+1$ can be expressed as follows,

$$\boldsymbol{h}_v^{k+1} = f\left(\sum_{r \in \mathcal{R}} \frac{1}{|\mathcal{N}_v^r|} \sum_{u \in \mathcal{N}_v^r} \boldsymbol{W}_r^k \phi\left(\boldsymbol{h}_u^k, \boldsymbol{h}_r^k\right)\right), \quad (2)$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ denotes the nonlinear activation function. Such relation-specific message passing is illustrated from Figure 5(a) to (b). Besides, in each layer the relation representation is obtained

[2] https://github.com/fxsjy/jieba

via linear projection,

$$\boldsymbol{h}_r^{k+1} = \boldsymbol{W}_{\text{rel}}^{k+1}\boldsymbol{h}_r^k, \tag{3}$$

where $\boldsymbol{W}_{\text{rel}}^{k+1}$ denotes the relational projection matrix at layer $k+1$. The pre-trained embeddings are initialized for $\boldsymbol{h}_r^0, \boldsymbol{h}_u^0, \boldsymbol{h}_v^0$.

Compared with task-agnostic pre-training, the GNN based encoding is task-specific, where the learnable projection matrices determine the influences of various messages and refine useful knowledge for site selection, supervised by the task loss introduced in the following.

3.3.2 Relation Path based Decoder. With knowledgeable representations obtained in GNN based encoder, to explore the explainability of knowledge, we further propose a relation path based decoder for both effective and explainable site decisions. Here we first introduce the relation path in KG [20, 57].

Definition 3.1. **Relation Path.** A relation path in KG $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$ is defined as $p = (r_1, \cdots, r_{|p|})$, where $|p|$ denotes the number of hops and $r_1, \cdots, r_{|p|} \in \mathcal{R}$.

Obviously, the relation path provides rich semantic contexts and can be used to explain the underlying factors of site decisions with UrbanKG. For example, the relation path Brand $\xrightarrow{r_{OSA}}$ Region $\xrightarrow{r_{SF}}$ Region focuses on the criteria of region function, i.e., opening the new store at the region with similar functions, while Brand $\xrightarrow{r_{RB}}$ Brand $\xrightarrow{r_{OSA}}$ Region indicates the logic that the brand learns from its related brand and opens the new store at the same region. Thus, based on existing studies in decision making for business site selection, especially analytic hierarchy process [48, 49], which identifies 5 criteria to guide site selection according to the brand, geographical distance and region function, etc. Also, we further follow existing feature engineering studies [16, 24, 42] to select other 3 criteria related to competitiveness, human flow and shop category. Therefore, guided by these criteria, we use 8 relation paths on UrbanKG to model site selection criteria. We summarize the relation paths for site selection in Table 2.

Table 2: Relation paths for site selection in UrbanKG. Relations of $r'_{\text{BS\_1}}$, $r'_{\text{B2C\_1}}$ $r'_{\text{P2C\_1}}$ represent inverse relations.

Criteria	Relation Paths with Ontologies
Region Distance	Brand $\xrightarrow{r_{OSA}}$ Region $\xrightarrow{r_{NB}}$ Region
Region Function	Brand $\xrightarrow{r_{OSA}}$ Region $\xrightarrow{r_{SF}}$ Region
Region Flow	Brand $\xrightarrow{r_{OSA}}$ Region $\xrightarrow{r_{FT}}$ Region
Business Area	Brand $\xrightarrow{r_{OSA}}$ Region $\xrightarrow{r'_{BS}}$ Ba $\xrightarrow{r_{BS}}$ Region
Related Brand	Brand $\xrightarrow{r_{RB}}$ Brand $\xrightarrow{r_{OSA}}$ Region
Brand Category	Brand $\xrightarrow{r_{B2C\_1}}$ Cate_1 $\xrightarrow{r'_{B2C\_1}}$ Brand $\xrightarrow{r_{OSA}}$ Region
Competitiveness	Brand $\xrightarrow{r_{BO}}$ POI $\xrightarrow{r_{Comp}}$ POI $\xrightarrow{r_{LA}}$ Region
Store Category	Brand $\xrightarrow{r_{B2C}}$ Cate_1 $\xrightarrow{r'_{P2C\_1}}$ POI $\xrightarrow{r_{LA}}$ Region

Based on the relation paths, we introduce the design of relation path based decoder, as shown in Figure 5(c). First, we obtain the representation of each relation path by semantic composition [20]. Specifically, given a relation path $p = (r_1, \cdots, r_{|p|})$ and the brand b, the brand-specific path representation can be calculated via following three ways,

$$\text{Add}: \boldsymbol{p} = \boldsymbol{h}_b^K + \boldsymbol{h}_{r_1}^K + \cdots + \boldsymbol{h}_{r_{|p|}}^K, \tag{4}$$

$$\text{Mult}: \boldsymbol{p} = \boldsymbol{h}_b^K \odot \boldsymbol{h}_{r_1}^K \odot \cdots \odot \boldsymbol{h}_{r_{|p|}}^K, \tag{5}$$

$$\text{GRU}: \boldsymbol{p} = \text{GRU}([\boldsymbol{h}_{r_1}^K, \cdots, \boldsymbol{h}_{r_{|p|}}^K], \boldsymbol{h}_b^K), \tag{6}$$

where $\odot$ is the element-wise product, and $\boldsymbol{h}_b^K$ in (6) is the initial hidden state for gated recurrent unit (GRU) input.

Since multiple factors/criteria are comprehensively considered in site selection [34], we further applies the attention mechanism [35] on relation paths for brand-specific site decision vector,

$$z_b = \text{Attention}(\boldsymbol{W}^{\text{Query}}\boldsymbol{h}_b^K, \boldsymbol{W}^{\text{Key}}\boldsymbol{P}, \boldsymbol{W}^{\text{Value}}\boldsymbol{P}), \tag{7}$$

where $\boldsymbol{P} = [\boldsymbol{p}_1; \cdots; \boldsymbol{p}_{n_p}]$ is the concatenated relation path representation matrix and n_p is the number of relation paths for site selection ($n_p = 8$ in our case). $\boldsymbol{W}^{\text{Query}}, \boldsymbol{W}^{\text{Key}}$ and $\boldsymbol{W}^{\text{Value}}$ are learnable parameters in the attention mechanism. The attention weights provide explainable results behind site decisions, especially the relationship between brands and various criteria.

For pairwise data $(b, a) \in \mathcal{D}_{\text{site}}$ (b is the brand and a is the region), the decision vector is multiplied with region embedding vector for the path based score. Additionally, for relatedness maximization, we utilize the bilinear product[3] to obtain the link based score via direct relation *OpenStoreAt*. The two parts are fused by a hyperparameter α for final link prediction score on site selection.

$$y_{ba} = (1-\alpha) \cdot \boldsymbol{z}_b^\top \boldsymbol{h}_a^K + \alpha \cdot \langle \boldsymbol{h}_b^K, \boldsymbol{h}_{r_{OSA}}^K, \boldsymbol{h}_a^K \rangle. \tag{8}$$

Accordingly, we adopt the cross-entropy loss for model parameter learning, and formulate the objective function as follow,

$$\min_{\Theta} \sum_{(b_i, a_j) \in \mathcal{D}_{\text{Site}}} -\log \frac{e^{y_{b_i a_j}}}{\sum_{a_k \in \mathcal{A}} e^{y_{b_i a_k}}} + \lambda \cdot \|\Theta\|, \tag{9}$$

where Θ includes the learnable parameters in GNN based encoder and relation path based decoder. $\mathcal{A}$ represents the set of candidate regions. λ is used to regularize the model parameters. The proposed KnowSite model is trained in a mini-batch way to minimize the objective formulation above.

Overall, with task-specific loss and end-to-end training, the proposed KnowSite model designs the multi-relational GNN based encoder for site selection related message passing, and further learns the relation path based decoder to explicitly model the logic of site decisions, achieving both effective and explainable performance.

4 EVALUATION

4.1 Experimental Setup

4.1.1 Datasets. Several sources of urban data are collected and crawled from map service, life service platform, social media as well as Internet service provider. Besides, the user data has been anonymized for privacy protection.

Built upon these multi-source urban data, we collect two datasets for evaluation as shown in Table 3.

[3] $\langle \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \rangle = \sum_i a_i \cdot b_i \cdot c_i$

Table 3: Dataset statistics. #triplet denotes the number of triplets in the corresponding UrbanKGs.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#triplet	Brand	Region	Train	Valid	Test
Beijing	23,754	35	330,652	398	528	15,022	5,007	5,008
Shanghai	41,338	36	589,852	441	2,042	29,006	9,669	9,669

- **Beijing**: This dataset focuses on the area within the Fifth Ring Road, Beijing, China.
- **Shanghai**: This dataset focuses on the whole area of Shanghai, China.

The brands with over 20 stores are selected for dataset construction, and the site selection data are randomly split into train/valid/test sets by a proportion of 6:2:2. The details of site selection data and UrbanKG can be found in Appendix B.

4.1.2 Baselines. We compare our proposed KnowSite model with two types of models. First, following the feature engineering and framework in [16, 18, 24, 42, 43], we choose five traditional data-driven models, Lasso [31], XGBoost [6], Geo-Spotting [16], D2S3 [42] and the state-of-the-art neural network-based model, NeuMF-RS [18]. All data sources have been utilized for feature extraction. Due to the model generalization issue to various brands, we train and test the first four models brand by brand, and report the average performance. Second, we further compare with four typical KRL models on UrbanKG, TransE [3], DistMult [46], ComplEx [33], TuckER [2] and CompGCN [35]. All the baselines are tuned with their reported settings (in site selection works, if applicable), and the weights of *OpenStoreAt* links in KRL models are increased to 10 for the site selection task.

4.1.3 Evaluation Metrics. We evaluate the site selection performance with five standard metrics of NDCG@k, Hit@k, Precision@k, Recall@k and MAP@k [16, 18, 42, 43], defined in Appendix C.1. We evaluate the performance with $k = 5, 10, 20$. Due to the space limitation, some results with $k = 5, 20$ are omitted, which are in accord with other metrics.

4.1.4 Implementation. For the proposed KnowSite model learning, the batch size is set to 128 and the embedding dimension d is set to 64. Besides, batch normalization and dropout are used for regularization. We use the rotate composition operator [8] in GNN based encoder, and the number of GNN layers ranges from 1 to 3. We tune other hyper-parameters with early stopping mechanism on validation NDCG@10. The learning rate and the dropout are searched from $\{0.0005, 0.001, 0.003, 0.005\}$ and $\{0.1, 0.3, 0.5\}$, respectively. The fusion parameter α ranges from 0.0 to 1.0. As for the pre-training step, we train the TuckER model with early stopping mechanism on training loss. All models are run 10 times and the average results are reported to prevent extreme cases. Besides, the stores (POIs) as well as *OpenStoreAt* links in valid & test sets are removed from UrbanKG to avoid test leakage. The implementation code and dataset are available at the given link[4]

Next, we present the performance comparison on two datasets, and then analyze the effectiveness of each module in KnowSite with ablation study. Several explainable results are further investigated for the logic of site selection. Finally, we validate the robustness of the proposed model in dynamic environment.

4.2 Performance Comparison

Table 4 presents the site selection performance comparison on both datasets. For KnowSite, all three composition operations of addition (Add), multiplication (Mult) and GRU are considered for relation path representation. In general, our proposed KnowSite outperforms all baselines across five evaluation metrics. Specifically, the improvement in Beijing dataset ranges from 5.9%~16.5%, while the improvement in Shanghai dataset is from 6.5%~10.7%. For example, for Precision@10 in Bejing dataset, the improvement is 12.0%. The considerable improvements demonstrate the effectiveness of our proposed knowledge-driven paradigm as well as systematic encoder-decoder framework. Besides, KnowSite models with three composition operations achieve comparable performance, and we select the GRU operation for detailed studies later.

Moreover, we have following three observations. First, knowledge-driven models of DistMult, ComplEx, TuckER and KnowSite perform more competitively than left data-driven ones, which owes to the knowledge discovery on UrbanKG. For example, the best data-driven baseline NeuMF-RS formulates the problem as matrix completion, which is easily affected by limited brand-region samples and cannot exploit rich semantics in multi-source urban data as UrbanKG does. Second, knowledge-driven models show strong robustness to various cities with knowledge refinement. For the two datasets, Shanghai dataset contains much more brands and candidate regions, and thus is more challenging. Due to the incompleteness of feature engineering and diverse influences, the performance gap of data-driven models between the two datasets are significant, e.g., a gap of over 0.150 on Hit@10 for SVR/XGBoost. In comparison, the gap for knowledge-driven models is less than 0.080 with site selection knowledge learned. Third, the performance gap between KRL models and KnowSite implies that extending KRL models to site selection is nontrivial and needs further customized designs, e.g., multi-relational message passing for knowledge refinement, and site selection related relation paths as well as brand-specific attention mechanism for knowledge explainability. Besides, on the larger dataset Shanghai, a training epoch costs 36 seconds and total training is in 50 epochs, while the inference costs 24 seconds, which is acceptable in practice.

4.3 Ablation Study

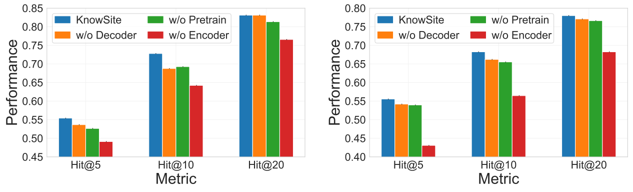
To evaluate the effectiveness of each module in KnowSite, Figure 6 shows the hit ratio performance of different model variants on both datasets. Specifically, we evaluate the KnowSite model without pre-training, GNN based encoder and relation path based decoder, respectively. Note that the variant without decoder (w/o Decoder) is equivalent to the KnowSite model with $\alpha = 1$ in (8).

According to the results, without the GNN based encoder, the model performance is reduced by 12% and 17% on Hit@10 for Beijing and Shanghai datasets, respectively. Thus, the GNN based encoder plays a quite important role in performance guarantee, which confirms the importance of knowledge refinement and the gain of task-specific message passing mechanism. Compared with other KRL models, the GNN based encoder successfully models

[4] https://github.com/tsinghua-fib-lab/UrbanKG-KnowSite

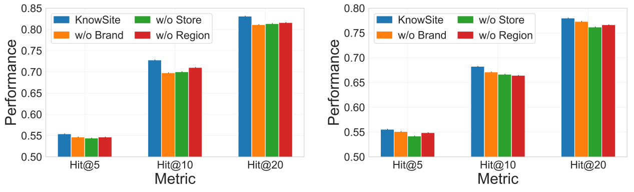
Table 4: Performance comparison w.r.t. test NDCG@k, Hit@k, Precision@k, Recall@k and MAP@k on two datasets. Best results are in bold and the best results (in baselines) are underlined. The last two rows show relative improvement in percentage and *p*-value compared with the best baseline with 10 runs of experiments.

	Beijing							Shanghai						
Model	N@5	N@10	H@5	H@10	P@10	R@10	M@10	N@5	N@10	H@5	H@10	P@10	R@10	M@10
Lasso	0.057	0.061	0.189	0.305	0.061	0.068	0.031	0.039	0.037	0.118	0.176	0.037	0.038	0.020
XGBoost	0.100	0.100	0.320	0.454	0.089	0.103	0.050	0.075	0.062	0.205	0.297	0.058	0.062	0.030
D2S3	0.094	0.093	0.301	0.435	0.082	0.096	0.046	0.064	0.059	0.211	0.299	0.054	0.058	0.028
Geo-Spotting	0.122	0.121	0.369	0.501	0.104	0.122	0.064	0.085	0.081	0.274	0.383	0.074	0.078	0.038
NeuMF-RS	0.180	0.178	0.501	0.653	0.155	0.182	0.097	0.178	0.168	0.478	0.615	0.148	0.163	0.090
TransE	0.080	0.084	0.297	0.460	0.075	0.089	0.036	0.064	0.063	0.244	0.372	0.058	0.064	0.026
DistMult	0.161	0.161	0.475	0.634	0.137	0.164	0.083	0.150	0.142	0.448	0.591	0.124	0.138	0.071
ComplEx	0.170	0.169	0.502	0.657	0.143	0.171	0.088	0.147	0.142	0.442	0.583	0.126	0.140	0.070
TuckER	0.183	0.183	0.518	0.673	0.156	0.187	0.098	0.188	0.174	0.502	0.620	0.150	0.166	0.094
CompGCN	0.196	0.194	0.503	0.668	0.166	0.198	0.109	0.203	0.188	0.506	0.630	0.161	0.178	0.105
KnowSite (Add)	0.218	0.217	0.556	0.707	0.185	0.222	0.125	0.218	0.200	0.541	0.653	0.171	0.191	0.113
KnowSite (Mult)	**0.221**	**0.219**	**0.565**	0.709	**0.186**	**0.224**	**0.127**	0.219	0.202	**0.543**	0.664	0.173	0.193	0.115
KnowSite (GRU)	0.220	**0.219**	0.557	**0.713**	**0.186**	0.223	**0.127**	**0.220**	**0.205**	**0.543**	**0.671**	**0.177**	**0.197**	**0.116**
Improv.	**12.7%**	**12.8%**	**9.1%**	**5.9%**	**12.0%**	**13.1%**	**16.5%**	**8.4%**	**9.0%**	**7.3%**	**6.5%**	**9.9%**	**10.7%**	**10.5%**
p−value	2.0e-10	1.5e-11	1.8e-6	1.6e-5	7.1e-12	1.1e-11	1.5e-11	1.1e-9	4.2e-11	6.9e-8	3.0e-10	1.9e-11	6.6e-11	1.2e-9



(a) Beijing (b) Shanghai

Figure 6: Performance comparison of different model variants on datasets.



(a) Beijing (b) Shanghai

Figure 7: Performance comparison of KnowSite models without different types of relation paths on datasets.

diverse knowledge with site selection, making the KnowSite model expressive. Besides, the pre-training step provides a task-agnostic but semantic initialization, contributing a performance gain of 5% on Hit@10 for datasets. Moreover, relation path based decoder further achieves 4%-5% improvement on Hit@10 with brand-specific choice of site selection criteria. Therefore, all three modules of pre-training, GNN based encoder and relation path based decoder are quite essential for effective site decisions.

4.4 Explainability Study

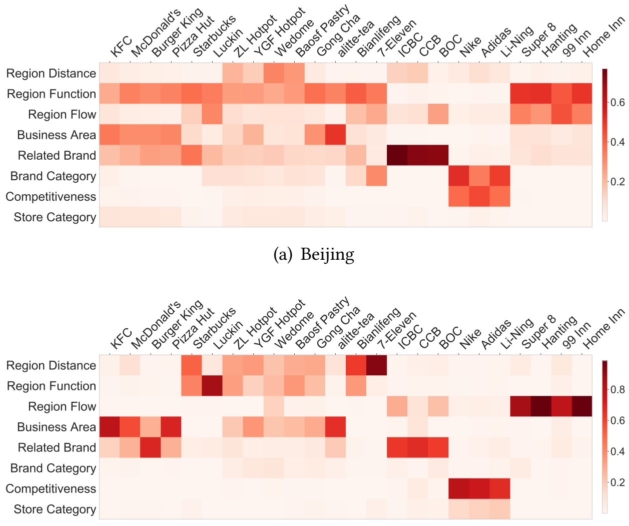
To further investigate the influence of relation paths in KnowSite as well as understand the reasons behind different brands' site decisions, we present several case studies in this part.

4.4.1 Influence of Relation Paths. The relation paths in Table 2 can be categorized into three types of region-based (the first four paths), brand-based (the 5th and 6th paths), and store-based (the last two paths) criteria, and we investigate their influence on model performance by removing any type of relation paths in decoder of KnowSite, as shown in Figure 7.

Overall, we observe the performance decrease in both datasets. For example, based on the evaluation metric of Hit@10, removing brand-based relation paths brings a drop of 4% for Beijing dataset, while removing region-based ones brings a drop of 3% for Shanghai dataset. More importantly, based on results in two datasets, we are able to identify different preferences to above relation paths for different cities, which may be caused by different city structures and other social factors. Specifically, the region-based relation paths are the most important type for Shanghai but the least important one for Beijing. This may partly owe to the different region structures. Due to numerous waterways in Shanghai, the regions are in irregular structure and thus own various functions and sizes, which further becomes a quite important factor for site selection. In contrast, regions in Beijing are arranged in grid structure with similar functions and sizes, which is less important than other factors like the characteristics of brands and stores. Hence, the influence of relation paths provides explainable site decisions in different cities.

4.4.2 Brands v.s. Site Selection Criteria. As described in Section 3.3.2, the attention weights in (7) show the relationship between brands and criteria. Thus, we present the attention weight visualization on two datasets in Figure 8. Several typical brands across food, leisure sports, accommodation and other categories are selected for visualization. A description of selected brands can be found in Appendix B.2 for better understanding.

By combining visualization results in different cities, i.e., Figure 8(a) and (b) together, we have similar findings regarding to brands' preference to selection criteria that are both insightful and



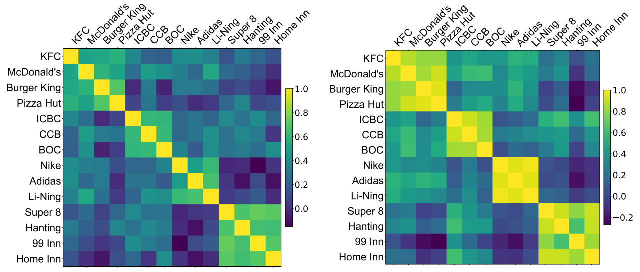
(a) Beijing

(b) Shanghai

Figure 8: Attention weight visualization of different brands to site selection criteria on datasets.

convincing. First, all fast-food chain brands like KFC, McDonald's, Burger King and Pizza Hut determine optimal locations with business area condition and related brand strategy considered, which is in accord with the location game between brands [30] as well as the commonsense that there always is one KFC store near one McDonald's store [12]. Second, similar attention on related brand strategy can also be observed among bank brands of ICBC, CCB and BOC, three large banks in China. These bank brands also focus on region flow for more customers. Moreover, the last four columns in figures represent the preference of four popular hotel chain brands to region flow, which determines the occupancy directly. Note that the slight difference between results in Figure 8(a) and (b) may be caused by different city conditions and noise in model learning. Overall, such results demonstrate the explainable capability of our proposed KnowSite model, which can provide a good reference for site selection understanding.

To further investigate the influence of site selection criteria on brand representations, Figure 9 visualizes the cosine distance between selected brands in Beijing, in which Figure 9(a) utilizes task-agnostic representations $\boldsymbol{h}_b^0$ of pre-training, while Figure 9(b) utilizes task-specific ones $\boldsymbol{h}_b^K$ of GNN based encoder output with end-to-end training. Since UrbanKG contains semantic information like *RelatedBrand* links, related brands' representations are closer compared with others, as shown in diagonal blocks of Figure 9(a). However, due to the task-agnostic learning' in pre-training step, such correlation is not that obvious. In comparison, a remarkable brand correlation is illustrated in Figure 9(b). Several highlight diagonal blocks indicate the closeness of brands in hidden space, such as the first block of four fast-food chain brands and the last block of four hotel chain brands. Besides, the brand correlations in off-diagonal parts are also enhanced in Figure 9(b), which also suggests the effectiveness of knowledge refinement with brand information encoding. Therefore, KnowSite successfully captures the semantic relatedness among brands and reveals the relationship between brands and various site selection criteria.

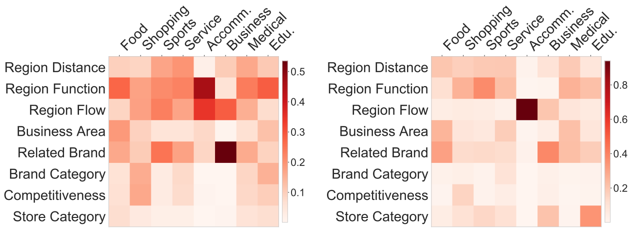


(a) Task-agnostic output

(b) Task-specific output

Figure 9: Cosine distance visualization of different brands' representations in Beijing.

4.4.3 Categories v.s. Site Selection Criteria. In Figure 10, we further reveal the relationship between categories and site selection criteria. For each dataset, eight typical categories are selected, and the attention weights of all brands under corresponding categories are averaged for visualization.



(a) Beijing

(b) Shanghai

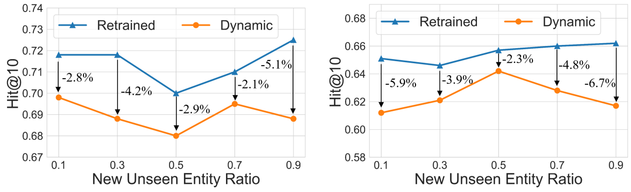
Figure 10: Attention weight visualization of different categories to site selection criteria on datasets. Sports, Service, Accomm., Edu. represent leisure sports, life service, accommodation and education, respectively.

Similar phenomenons in Figure 8 can be observed in Figure 10. The brands of food category focus on business area and related brand, while the brands of accommodation category pay more attention to region function and flow. Among the site selection criteria, the region factors of distance and function as well as related brand strategy are commonly considered across various categories. Such results again validate the importance of relation path based decoder in KnowSite, and explore its potential in site selection for both brand and category levels.

Throughout the experimental study, KnowSite achieves the state-of-the-art performance on site selection task, and the effectiveness of each designed module is validated. Moreover, with relation paths and attention mechanism utilized, KnowSite successfully reveals the influences of site selection criteria on various businesses.

4.5 Dynamic Case Study

In the framework design and aforementioned experiments, we mainly focus on the static case, where all the entities as well as their connections in UrbanKG are observed in training. However, in practical applications, the surrounding environment for site selection always dynamically changes, especially for the case of newly opened/deployed infrastructures (POIs), i.e., the UrbanKG can be dynamic with new POI entities added. In such case, the newly added POI entities are only available in testing step, which is also known as inductive setting [22, 38] or out-of-sample settting [1] in KG research. To cope with this, we integrate oDistMult [1] with the proposed KnowSite model, which obtains the embedding of new unseen entity by aggregating embeddings of all seen entities in its neighborhood. As a result, the modified KnowSite model can be applied in the dynamic case without retraining.



(a) Beijing

(b) Shanghai

Figure 11: Performance comparison of retrained and dynamic KnowSite models with different amounts of new unseen entities under dynamic setting.

Especially, we keep the half of POI entities in training UrbanKG, while add different ratios of POI entities into UrbanKG in testing. Figure 11 shows the model performance comparison, where "Retrained" denotes the KnowSite model retrained on the extended UrbanKG in testing, while "Dynamic" denotes the above modified KnowSite model. According to the results, the dynamic model achieves close performance with the retrained model in both datasets (no more than 7% performance drop) but avoids costly retraining in practice. Such results further validate the compatibility of KnowSite model with advanced methods in KG as well as its practicality to real-world site selection applications.

5 RELATED WORK

Closely related studies of our work include site selection methods, knowledge representation learning with KG, and KG applications in urban computing.

Site Selection. With multi-source urban data available, data-driven methods first extract features from data, and then learn regression/learning-to-rank models for the problem [16, 42]. Most of these methods follow the static prediction without considering the dynamic environment. Specifically, both Geo-Spotting [16] and DD3S [42] firstly investigate the predictive power of various features like density, competitiveness and area popularity, and then apply traditional SVR and RankNet to determine the optimal location. However, these methods learn individual models for each brand and cannot generalize to various businesses. Furthermore, several works also integrate deep network with feature engineering [18, 24, 43, 45]. For example, DeepStore [24] and AR^{2}Net [43] extracts features from commercial data, satellite images, etc., and further combine deep neural networks with attention mechanism for solution. UKG-NN [52] builds a relational graph with manually defined features, which are passed to the neural network for site decisions. NeuMF-RS [18] adds restaurants' and sites' attributes to neural collaborative filtering for site selection. O^2-SiteRec [45] develops a multi-graph attention network model for online-to-offline store site recommendation. However, deep models suffer from explainability issues with black-box neural networks. Since both traditional and deep models fail to extract vital knowledge like site selection criteria from data, their performance is easily affected by the quality of upstream feature engineering. In contrast, our work leverages knowledge-driven paradigm for both effective and explainable performance.

Knowledge Representation Learning. As for KRL to learn embeddings of entities and relations, though complete structures like GNN have been introduced [28, 35], tensor decomposition models still achieve the best performance [14], such as DistMult [46] and TuckER [2]. Here we argue that the proposed GNN encoder is more suitable for representing specific knowledge of site selection, as it can flexibly control the information sharing among diverse factors. Meanwhile, multi-hop relation paths have been introduced in KRL for more accurate representations [20, 57]. In our proposed KnowSite, we adopt relation path based decoder to model site selection criteria for brands. Thus, it not only boosts the performance, but also provides explainable site decisions based on the relation path logic. Note that our relation path based on KG is different from the meta-path counterpart in heterogeneous graphs [47], which only learns node embeddings but ignores edge representations [41]. Thus, it is not applicable to this work.

KG for Urban Computing. Recently, there are some attempts to apply KG for urban computing. For example, the construction of geographic KGs is investigated in [29, 44], where the spatial relationships between geographic components are extracted. Some works [7, 37, 39, 51, 54] introduce KG with two or three relations and ontologies for specific applications. Besides, KG is utilized to mine urban flow patterns in [21, 58], and socioeconomic prediction in [25, 56]. However, such developed KGs miss important knowledge for site selection like human flow, competitiveness, brand relatedness, etc. In comparison, our proposed UrbanKG contains rich site selection related knowledge with over 20k entities in city and over 300k facts between them, which is a promising backbone for various applications in urban computing.

6 CONCLUSION

In this work, we proposed KnowSite, a knowledge-driven model for site selection. By leveraging KG for urban knowledge representation, KnowSite develops a generalized encoder-decoder framework, where multi-relational message passing and criteria-based relation paths are adopted to understand different brands' site decisions. Extensive experiments demonstrate that KnowSite achieves superior performance with both effectiveness and explainability achieved.

For future works, we will combine KnowSite with the traditional data-driven paradigm, and utilize both KRL methods and feature engineering towards powerful site selection. Moreover, we plan to explore our proposed UrbanKG as well as the generalized encoder-decoder framework for other urban computing tasks such as flow prediction, socioeconomic indicator prediction, etc.

ACKNOWLEDGMENTS

This research has been supported in part by the National Key Research and Development Program of China under Grant 2022ZD0116402; in part by the National Natural Science Foundation of China under Grant U22B2057, Grant U20B2060, and Grant 61971267.

REFERENCES

[1] Marjan Albooyeh, Rishab Goel, and Seyed Mehran Kazemi. 2020. Out-of-sample Representation Learning for Knowledge Graphs. In *EMNLP*. 2657–2666.

[2] Ivana Balažević, Carl Allen, and Timothy M Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *EMNLP*. 5185–5194.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NeurIPS*. 2787–2795.

[4] Michael J Breheny. 1988. Practical Methods of Retail Location Analysis: A Review. In *Store Choice, Store Location and Market Analysis*. Routledge, London, 39–86.

[5] Buru Chang, Gwanghoon Jang, Seoyoon Kim, and Jaewoo Kang. 2020. Learning Graph-based Geographical Latent Representation for Point-of-Interest Recommendation. In *CIKM*. 135–144.

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *KDD*. 785–794.

[7] Amine Dadoun, Raphaël Troncy, Olivier Ratier, and Riccardo Petitti. 2019. Location Embeddings for Next Trip Recommendation. In *WWW*. 896–903.

[8] Mikhail Galkin, Priyansh Trivedi, Gaurav Maheshwari, Ricardo Usbeck, and Jens Lehmann. 2020. Message Passing for Hyper-Relational Knowledge Graphs. In *EMNLP*. 7346–7359.

[9] Bin Guo, Jing Li, Vincent W Zheng, Zhu Wang, and Zhiwen Yu. 2018. Citytransfer: Transferring Inter-and Intra-city Knowledge for Chain Store Site Recommendation Based on Multi-source Urban Data. *UbiComp* 1, 4 (2018), 1–23.

[10] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A Survey on Knowledge Graph-based Recommender Systems. *IEEE TKDE* (2020).

[11] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge Graphs. *ACM Comput. Surv.* 54, 4 (2021), 1–37.

[12] Wei Hu and Yuanyuan Xie. 2013. Comparative Study of McDonald's and Kentucky Fried Chicken (KFC) Development in China. *Sanovia: Internal Business Administration* (2013).

[13] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge Graph Embedding Based Question Answering. In *WSDM*. 105–113.

[14] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2021. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE TNNLS* (2021).

[15] Iris A Junglas and Richard T Watson. 2008. Location-based Services. *Commun. ACM* 51, 3 (2008), 65–69.

[16] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013. Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In *KDD*. 793–801.

[17] Vipin Kumar and Kiran Karande. 2000. The Effect of Retail Store Environment on Retailer Performance. *Journal of Business Research* 49, 2 (2000), 167–181.

[18] Nuo Li, Bin Guo, Yan Liu, Yao Jing, Yi Ouyang, and Zhiwen Yu. 2018. Commercial Site Recommendation Based on Neural Collaborative Filtering. In *UbiComp Adjunct*. 138–141.

[19] Shuangli Li, Jingbo Zhou, Tong Xu, Hao Liu, Xinjiang Lu, and Hui Xiong. 2020. Competitive Analysis for Points of Interest. In *KDD*. 1265–1274.

[20] Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015. Modeling Relation Paths for Representation Learning of Knowledge Bases. In *EMNLP*. 705–714.

[21] Jia Liu, Tianrui Li, Shenggong Ji, Peng Xie, Shengdong Du, Fei Teng, and Junbo Zhang. 2021. Urban Flow Pattern Mining based on Multi-source Heterogeneous Data Fusion and Knowledge Graph Embedding. *IEEE TKDE* (2021).

[22] Shuwen Liu, Bernardo Grau, Ian Horrocks, and Egor Kostylev. 2021. Indigo: Gnn-based Inductive Knowledge Graph Completion Using Pair-wise Encoding. *NeurIPS* 34 (2021), 2034–2045.

[23] Yu Liu, Jingtao Ding, Yanjie Fu, and Yong Li. 2023. UrbanKG: An Urban Knowledge Graph System. *ACM TIST* 14, 4 (2023), 1–25.

[24] Yan Liu, Bin Guo, Nuo Li, Jing Zhang, Jingmin Chen, Daqing Zhang, Yinxiao Liu, Zhiwen Yu, Sizhe Zhang, and Lina Yao. 2019. DeepStore: An Interaction-aware Wide&Deep Model for Store Site Recommendation with Attentional Spatial Embeddings. *IEEE Internet Things J.* 6, 4 (2019), 7319–7333.

[25] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused Contrastive Learning for Urban Imagery-based Socioeconomic Prediction. In *WWW*. 4150–4160.

[26] Yu Liu, Zhilun Zhou, Yong Li, and Depeng Jin. 2023. Urban Knowledge Graph Aided Mobile User Profiling. *ACM TKDD* 1, 1 (2023).

[27] Nicholas A Phelps and Andrew M Wood. 2018. The Business of Location: Site Selection Consultants and the Mobilisation of Knowledge in the Location Decision. *Journal of Economic Geography* 18, 5 (2018), 1023–1044.

[28] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *ESWC*. 593–607.

[29] Yuhan Sun, Jia Yu, and Mohamed Sarwat. 2019. Demonstrating Spindra: A Geographic Knowledge Graph Management System. In *ICDE*. 2044–2047.

[30] Presh Talwalkar. 2012. Why are McDonald's and Burger King Usually Located Near Each Other? Fast Food Location Game Theory. https://mindyourdecisions.com/blog/2012/10/23/why-are-mcdonalds-and-burger-king-usually-located-near-each-other-fast-food-location-game-theory/

[31] Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B* 58, 1 (1996), 267–288.

[32] Harry Timmermans. 1986. Locational Choice Behaviour of Entrepreneurs: An Experimental Analysis. *Urban Studies* 23, 3 (1986), 231–240.

[33] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *ICML*.

[34] Gülden Turhan, Mehmet Akalın, and Cemal Zehir. 2013. Literature Review on Selection Criteria of Store Location Based on Performance Measures. *Procedia-Social and Behavioral Sciences* 99 (2013), 391–402.

[35] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2019. Composition-based Multi-relational Graph Convolutional Networks. In *ICLR*.

[36] Hongwei Wang, Hongyu Ren, and Jure Leskovec. 2021. Relational Message Passing for Knowledge Graph Completion. In *KDD*. 1697–1707.

[37] Huandong Wang, Qiaohong Yu, Yu Liu, Depeng Jin, and Yong Li. 2021. Spatio-temporal Urban Knowledge Graph Enabled Mobility Prediction. *UbiComp* 5, 4 (2021), 1–24.

[38] Peifeng Wang, Jialong Han, Chenliang Li, and Rong Pan. 2019. Logic Attention based Neighborhood Aggregation for Inductive Knowledge Graph Embedding. In *AAAI*, Vol. 33. 7152–7159.

[39] Pengyang Wang, Kunpeng Liu, Lu Jiang, Xiaolin Li, and Yanjie Fu. 2020. Incremental Mobile User Profiling: Reinforcement Learning with Spatial Knowledge Graph for Modeling Event Streams. In *KDD*. 853–861.

[40] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE TKDE* 29, 12 (2017).

[41] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous Graph Attention Network. In *WWW*. 2022–2032.

[42] Mengwen Xu, Tianyi Wang, Zhengwei Wu, Jingbo Zhou, Jian Li, and Haishan Wu. 2016. Demand Driven Store Site Selection via Multiple Spatial-temporal Data. In *SIGSPATIAL*. 1–10.

[43] Yanan Xu, Yanyan Shen, Yanmin Zhu, and Jiadi Yu. 2020. AR2Net: An Attentive Neural Approach for Business Location Selection with Satellite Data and Urban Data. *ACM TKDD* 14, 2 (2020), 1–28.

[44] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Rui Zhu. 2019. A Spatially Explicit Reinforcement Learning Model for Geographic Knowledge Graph Summarization. *Transactions in GIS* 23, 3 (2019), 620–640.

[45] Hua Yan, Shuai Wang, Yu Yang, Baoshen Guo, Tian He, and Desheng Zhang. 2022. O^2-SiteRec: Store Site Recommendation under the O2O Model via Multi-graph Attention Networks. In *ICDE*. 525–538.

[46] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR*.

[47] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. 2020. Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark. *IEEE TKDE* (2020).

[48] Jeremy YL Yap, Chiung Ching Ho, and Choo-Yee Ting. 2018. Analytic Hierarchy Process (AHP) for Business Site Selection. In *AIP Conference Proceedings*. 020151.

[49] Nurdan Yıldız and Fatih Tüysüz. 2019. A Hybrid Multi-criteria Decision Making Approach for Strategic Retail Location Investment: Application to Turkish food retailing. *Socio-Economic Planning Sciences* 68 (2019), 100619.

[50] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks.. In *NeurIPS*. 9240–9251.

[51] Ningyu Zhang, Huajun Chen, Xi Chen, and Jiaoyan Chen. 2016. Semantic Framework of Internet of Things for Smart Cities: Case Studies. *Sensors* (2016).

[52] Ningyu Zhang, Shumin Deng, Huajun Chen, Xi Chen, Jiaoyan Chen, Xiaoqian Li, and Yiyi Zhang. 2018. Structured Knowledge Base as Prior Knowledge to Improve Urban Data Analysis. *ISPRS International Journal of Geo-Information* 7, 7 (2018), 264.

[53] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL*. 1441–1451.

[54] Ling Zhao, Hanhan Deng, Linyao Qiu, Sumin Li, Zhixiang Hou, Hai Sun, and Yun Chen. 2020. Urban Multi-Source Spatio-Temporal Data Analysis Aware Knowledge Graph Embedding. *Symmetry* 12, 2 (2020), 199.

[55] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM TIST* 5, 3 (2014), 1–55.

[56] Zhilun Zhou, Yu Liu, Jingtao Ding, Depeng Jin, and Yong Li. 2023. Hierarchical Knowledge Graph Learning Enabled Socioeconomic Indicator Prediction in Location-based Social Network. In *WWW*. 122–132.

[57] Yao Zhu, Hongzhi Liu, Zhonghai Wu, Yang Song, and Tao Zhang. 2019. Representation Learning with Ordered Relation Paths for Knowledge Graph Completion. In *EMNLP*. 2662–2671.

[58] Chenyi Zhuang, Nicholas Jing Yuan, Ruihua Song, Xing Xie, and Qiang Ma. 2017. Understanding People Lifestyles: Construction of Urban Movement Knowledge Graph from GPS Trajectory.. In *IJCAI*. 3616–3623.

A URBANKG CONSTRUCTION DETAILS

We present the urbanKG construction details here. Table 5 presents the source of multi-source data used in this work, which also correspond to entities and relations in UrbanKG construction. For region entities in UrbanKG for Beijing and Shanghai datasets, we partition the city into multiple regions by the road network, and each region entity is provided with a sequence of longitude-latitude pairs $L_a = \{(lng_a^1, lat_a^1), \cdots, (lng_a^k, lat_a^k)\}$ as region boundary. Brand entities are commonly brands opening stores. POI entities and business area entities are provided with location information of longitude-latitude pairs like $l_i = (lng_i, lat_i)$. The category entities are POI properties identified by experts, e.g., food, shopping, business, residence, education, etc.

Based on the entities above, the relational links defined in Table 1 can be extracted as follows.

- *BorderBy.* Given two regions a, b, they are connected by *BorderBy* if $|L_a \cap L_b| > 0$, i.e., sharing the same boundary points.
- *NearBy.* Given two regions a, b, they are connected by *NearBy* if $\|\bar{L}_a - \bar{L}_b\| \leq 1km$, where $\bar{L}_a, \bar{L}_b$ are center location of regions.
- *FlowTransition.* Given two regions a, b, they are connected by *FlowTransition* if the aggregated flow transition between two regions exceeds the threshold.
- *SimilarFunction.* Given two regions a, b and the category distribution vectors of POIs therein z_a, z_b, they are connected by *SimilarrFunction* if $cos(z_a, z_b) \geq 0.95$ with cosine similarity.
- *CoCheckin.* Given two POIs p_1, p_2, they are connected by *CoCheckin* if the number of records that consecutively visit p_1 and p_2 exceeds the threshold.
- *Competitive.* Given two POIs p_1, p_2, they are connected by *Competitive* if $\|l_{p_1} - l_{p_2}\| \leq 500m$ and they are in the same category.
- *RelatedBrand* Two brands are connected by *RelatedBrand* if they are connected by "relatedPage" relation in zhishi.me KG.
- *SubCateOf_ij.* The categories are connected by *SubcateOf_ij* according to taxonomy.
- *BaServe* Given a region a and a business area ba, they are connected by *BaServe* if $\|\bar{L}_a - l_{ba}\| \leq 3km$.
- *BelongTo.* Given a POI p and a business area ba, they are connected by *BelongTo* if $\|l_p - l_{ba}\| \leq 3km.$
- *LocateAt.* Given a POI p and a region a, they are connected by *LocateAt* if l_p is in the closure by region boundary L_a.
- *POIToCate_i.* A POI is connected to its associated category by *POIToCate_i.*
- *BrandToCate_i.* A Brand is connected to its associated category by *BrandToCate_i.*
- *BrandOf.* A POI is connected to its associated brand by *BrandOf.*
- *OpenStoreAt* Given a brand b and a region a, they are connected by *OpenStoreAt* if b opens the store at region a.

B DATASET DETAILS

B.1 Dataset Statistics

Table 6 introduces the ontology statistics of UrbanKG, i.e., the number of entities for corresponding ontology. As for POIs in the construction of UrbanKG, we only consider those belonging to selected brands in datasets.

Table 7 shows the relational fact statistics of UrbanKG in two cities for our work.

B.2 Details of Selected Brands for Visualization

Here we give a description of selected brands in experiments.

- **KFC**, **McDonald's**, **Burger King**, **Pizza Hut**. Fast-food chain brands around the world.
- **Starbucks**, **Luckin**. Coffeehouse chain brands. Luckin, founded in Beijing, manages more stores than Starbucks in China.
- **ZL(Zhangliang) Spicy Hotpot**, **YGF(Yang Guofu) Spicy Hotpot**. Two of the largest spicy hotpot (a.k.a. Mala Tang, Chinese snack) chain brands in China.
- **Wedome**, **Baosf Pastry**. Bakery chain brands in China, focus on cakes, bread, and bakery items.
- **Gong Cha**, **alittle-tea**. Tea chain brands, offering both original tea and milk tea.
- **Bianlifeng**, **7-Eleven**. Convenience store chain brands.
- **ICBC (Industrial and Commercial Bank of China)**, **CCB (China Construction Bank)**, **BOC (Bank of China)**. State-owned commercial bank companies in China, opening branch banks and ATMs throughout the country.
- **Nike**, **Adidas**, **Li-Ning**. Leisure sport chain brands.
- **Super 8 (Hotel)**, **Hanting (Hotel)**, **99 Inn**, **Home Inn**. Four of the largest hotel chain brands in China.

C EXPERIMENTAL DETAILS

C.1 Metrics

Given the region set $\mathcal{A}$, the brand set $\mathcal{B}$ and the i-th brand, we denote A^i and $\hat{A}^i$ as its true and model predicted region list based on popularity/predicted score, respectively. n_i denotes the number of regions in test set where the i-th brand opens the store. Then the metrics are calculated as follows,

- NDCG@k (Normalized Discounted Cumulative Gain), which measures the extent to which the top-k regions in A^i are highly ranked in $\hat{A}^i$.

$$\text{NDCG}@k = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\text{NDCG}_i@k, \quad \text{DCG}_i@k = \sum_{j=1}^{k}\frac{2^{rel(\hat{a}_j)} - 1}{\log_2(j+1)},$$

  where the relevance score $rel(\hat{a}_j)$ follows the definition in [16], i.e., $rel(\hat{a}_j) = \frac{|\mathcal{A}| - rank(\hat{a}_j) + 1}{|\mathcal{A}|}$ for ground truth and $rel(\hat{a}_j) = 0$ for invalid regions. NDCG$_i$@k is obtained by normalizing DCG$_i$@k via the ideal prediction IDCG$_i$@k.
- Hit@k, which describes the hit ratio of top-k regions in A^i.

$$\text{Hit}@k = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\mathbb{I}(|A^i_{1:k} \cap \hat{A}^i_{1:k}|),$$

  where $\mathbb{I}(\cdot)$ denotes the indicator function, i.e., $\mathbb{I}(x) = 1$ if $x > 0$, otherwise $\mathbb{I}(x) = 0$.
- Precision@k and Recall@k, which are defined as follows,

$$\text{Precision}@k = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{|A^i \cap \hat{A}^i_{1:k}|}{k}.$$

$$\text{Recall}@k = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{|A^i \cap \hat{A}^i_{1:k}|}{\min(n_i, k)}.$$

Table 5: The data source for UrbanKG construction.

Data	Description	Entities	Relations
$\mathcal{D}_{\text{RN}}$	The road network data crawled from Map Platform	Region	*BorderBy*, *NearBy*
$\mathcal{D}_{\text{Ba}}$	The business area data crawled from Life Service Platform	Ba	*BaServe*
$\mathcal{D}_{\text{POI}}$	The POI data crawled from Map Platform	POI, Category	*SimilarFunction*, *Competitive*, *SubCateOf_ij*, *BelongTo*, *LocateAt*, *POIToCate_i*
$\mathcal{D}_{\text{Brand}}$	The brand data crawled from BaiduBaike	Brand	*RelatedBrand*, *BrandToCate_i*, *BrandOf*
$\mathcal{D}_{\text{Site}}$	The site selection records derived from $\mathcal{D}_{\text{POI}}$, $\mathcal{D}_{\text{RN}}$ and $\mathcal{D}_{\text{Brand}}$	-	*OpenStoreAt*
$\mathcal{D}_{\text{Traj}}$	The mobility trajectories collected from Mobile Operator	-	*FlowTransition*
$\mathcal{D}_{\text{Check}}$	The check-in records collected from Social Platform	-	*CoCheckin*

Table 8: The hyper-parameters for KnowSite in experiments.

Hyper-parameter	Beijing	Shanghai
batch_size	128	128
embedding dimension d	64	64
learning_rate	0.001	0.003
dropout	0.3	0.1
GCN_layers	2	2
fusion parameter α	0.5	0.8

Table 6: The ontology statistics of UrbanKG for cities.

Dataset	#Brand	#Region	#Ba	#POI	#1-Cate	#2-Cate	#3-Cate
Beijing	398	528	168	22,468	10	39	143
Shanghai	441	2042	264	38,394	11	42	144

Table 7: The details of defined relations in UrbanKG.

Relation	Beijing	Shanghai
BorderBy	2,626	9,896
NearBy	7,232	29,942
FlowTransition	287	634
SimilarFunction	2,844	5,126
Competitive	1,968	2,576
RelatedBrand	296	352
SubCateOf_ij	325	330
BaServe	6,152	11,876
BelongTo	22,372	38,394
LocateAt	22,468	38,394
POIToCate_i	22,468*3	38,394*3
BrandToCate_i	398*3	441*3
BrandOf	22,468	38,394
OpenStoreAt	15,022	29,006

- MAP@k (Mean Average Precision), which measures the relative ranking quality of the top-k regions in $\hat{A}^i$.

$$\text{MAP@}k = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{1}{\min(n_i,k)}\cdot\sum_{j=1}^{k}\frac{|A^i\cap\hat{A}^i_{1:j}|}{j}\cdot rel(\hat{a}_j),$$

where $rel(\hat{a}_j)$ follows the same definition above.

C.2 Hyper-parameter Selection

Table 8 summarizes all the hyper-parameters for the KnowSite model on two datasets.