OpenSiteRec: An Open Dataset for Site Recommendation

Xinhang Li[*]
Tsinghua University
Beijing, China
xh-li20@mails.tsinghua.edu.cn

Xiangyu Zhao
City University of Hong Kong
Hong Kong
xianzhao@cityu.edu.hk

Yejing Wang
City University of Hong Kong
Hong Kong
yejing.wang@my.cityu.edu.hk

Yu Liu
Tsinghua University
Beijing, China
liuyu2419@126.com

Chong Chen
Huawei Cloud BU.
Beijing, China
chenchong55@huawei.com

Cheng Long[†]
Nanyang Technological University
Singapore
c.long@ntu.edu.sg

Yong Zhang[†]
Tsinghua University
Beijing, China
zhangyong05@tsinghua.edu.cn

Chunxiao Xing
Tsinghua University
Beijing, China
xingcx@tsinghua.edu.cn

ABSTRACT

As a representative information retrieval task, site recommendation, which aims at predicting the optimal sites for a brand or an institution to open new branches in an automatic data-driven way, is beneficial and crucial for brand development in modern business. However, there is no publicly available dataset so far and most existing approaches are limited to an extremely small scope of brands, which seriously hinders the research on site recommendation. Therefore, we collect, construct and release an open comprehensive dataset, namely OpenSiteRec, to facilitate and promote the research on site recommendation. Specifically, OpenSiteRec leverages a heterogeneous graph schema to represent various types of real-world entities and relations in four international metropolises. To evaluate the performance of the existing general methods on the site recommendation task, we conduct benchmarking experiments of several representative recommendation models on OpenSiteRec. Furthermore, we also highlight the potential application directions to demonstrate the wide applicability of OpenSiteRec. We believe that our OpenSiteRec dataset is significant and anticipated to encourage the development of advanced methods for site recommendation. OpenSiteRec is available online at https://OpenSiteRec.github.io/.

CCS CONCEPTS

• **Information systems → Data mining**; **Business intelligence**; **Learning to rank**; **Recommender systems**; **Location based services**.

KEYWORDS

site recommendation, dataset, heterogeneous graph, benchmark

ACM Reference Format:
Xinhang Li, Xiangyu Zhao, Yejing Wang, Yu Liu, Chong Chen, Cheng Long, Yong Zhang, and Chunxiao Xing. 2024. OpenSiteRec: An Open Dataset for Site Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657875

1 INTRODUCTION

In modern business, selecting an optimal site to open a new branch is definitely crucial for the development of a brand or an institution [18, 20, 75]. An appropriate site will bring substantial profits while an inappropriate site may lead to business failure [28, 57, 58]. Thus, properly determining the best choice from so many candidate sites is quite important yet complex, since it needs to take many factors into accounts [39, 40, 54], such as the brand types and the population surrounding the site. Typically, this task is mainly accomplished by the professional consulting or marketing departments of companies [50], which is usually labor-intensive and time-consuming. Meanwhile, human error and bias can also lead to suboptimal solutions. Therefore, it is difficult for such an artificial approach to fulfill the high demand of rapid development in modern business.

Thanks to the booming development in information retrieval, automatic data-driven approaches have been introduced to assist the decision-making and reduce the cost [30], i.e. site recommendation [7, 25, 35]. These approaches come with a wide variety of definitions of site recommendation, including the association analysis [25] for feature selection, the rating prediction [12], the consumption prediction [36, 37] and the top-N recommendation [31, 34, 63]. While they share the idea of treating the site recommendation problem as a ranking task, their significantly different definitions make it difficult for them to be compared directly. Thus, all these works are independent of each other and they have to undesirably start from scratch instead of making continuous improvements, which is detrimental to the subsequent research on site recommendation.

[*]This work was done during the visiting at Nanyang Technological University when the author studied at Tsinghua University.
[†]Cheng Long and Yong Zhang are the corresponding authors.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657875

Table 1: Comparison with the datasets used by other site recommendation approaches. Specifically, '/' between numbers denotes the split w.r.t. different brands, '*' denotes the approximate number w.r.t. the setting in the corresponding paper.

Approach	Venue	Year	City	Brand	Region	Site
Geo-Spotting [25]	KDD	2013	New York City	3	32*	186/104/66
ANNRR [5]	UbiComp	2015	Washington, D.C. Hangzhou	1 1	181 882	203 2,115
PAM [31]	SIGSPATIAL	2015	Tianjin	1	99,007	34
BL-G-CoSVD [63]	TKDD	2016	Shanghai	5	17,435	17,435
DD3S [59]	SIGSPATIAL	2016	Beijing	2/2	10*	1,882/1,343
CityTransfer [12]	UbiComp	2017	Beijing Shanghai Xi'an Nanjing	3	1,000-3,000*	123/160/179 147/46/156 69/59/189 73/46/57
DeepStore [36]	IOT	2019	13 Cities	49 in Total	300*	49 in Total
WANT [37]	TKDD	2021	6 Cities	100* in Total	300*	100* in Total
UrbanKG [34]	ArXiv	2021	Beijing Shanghai	398 441	528 2,042	22,468 38,394
O^2-SiteRec [60]	ICDE	2022	Shanghai	122	2,000*	39,465
UUKG [41]	ArXiv	2023	New York City Chicago	15 15	260 77	62,450 31,573
OpenSiteRec	/	/	Chicago New York City Singapore Tokyo	969 2,702 1,922 4,861	801 2,325 2,043 3,036	8,044 14,189 9,912 26,765

Meanwhile, their datasets only cover very small yet different scopes in site recommendation, such as bike sharing station [5], chain hotel [12] and online stores with courier capacity [60]. This leads to failure in utilizing comprehensive information across scopes and severe data sparsity problems in site recommendation. Even more unfortunately, none of them has released their datasets so far. In most cases, collecting data and creating dataset are necessary in research but low-yielding since the dataset is typically a fundamental part in a scientific thesis. Therefore, the lack of publicly available dataset brings inconvenience to the researchers and forces them to spend long time dealing with the dataset construction, which even hinders the development of site recommendation solutions. According to the above problems, we believe a unified definition and a comprehensive open dataset of site recommendation are necessary and crucial for the benign development of the following research on site recommendation.

To this end, we propose a formal problem definition of site recommendation by jointly considering and summarizing the definitions of existing studies. Based on this problem definition, we collect, construct and release an **Open** benchmarking dataset for **Site Rec**ommendation, namely **OpenSiteRec**. Specifically, OpenSiteRec consists of four international metropolises, including *Chicago*, *New York City*, *Singapore* and *Tokyo*. Different from the datasets used by the existing approaches, our proposed OpenSiteRec contains all the brands and regions from all the scopes and types in the whole cities and thus yields a wide-range, much larger and more comprehensive dataset. Meanwhile, OpenSiteRec provides sufficient trustworthy commercial relationships and organizes the different types of real-world concepts into a heterogeneous graph to offer more comprehensive information. Furthermore, we also conduct benchmarking experiments of several representative baselines on OpenSiteRec to facilitate future research. Some discussions of the potential application directions of OpenSiteRec in other research areas, including brand entry forecasting and business area planning, and also the limitations of OpenSiteRec are presented to give a broader view of OpenSiteRec.

The contributions of this paper are summarized as follows:

- We introduce a formal definition of site recommendation by summarizing the task definitions of existing works, which unifies them to provide open benchmarks for the following research.
- We collect, construct and release an open comprehensive dataset of four international metropolises, namely **OpenSiteRec**, to facilitate the subsequent research on site recommendation. To the best of our knowledge, OpenSiteRec is the first publicly available dataset for site recommendation.
- We conduct benchmarking experiments of 16 widely-used baseline models in recommendation on OpenSiteRec, to verify their effectiveness in site recommendation and to facilitate future research for comparison.

- Besides site recommendation, various other research areas such as brand expansion, urban planning and facility location can benefit from OpenSiteRec given that it embeds rich information on both commercial and geographical aspects of urban spaces.

2 RELATED WORKS

Site recommendation for store brands and public facilities is a widely-studied problem with strong practical significance in modern business [1] and urban planning [48]. The most seminal attempt [22, 23] proposes to investigate the potential effects of different features for retail store site identification. Following it, some data mining approaches are proposed in evaluating the correlations between the street centrality and the geographical distribution of activities in Bologna [47] and Barcelona [46].

With the booming development of machine learning algorithms, applying them for site recommendation becomes an effective and efficient solution, which has attracted increasing interest and thus yields many approaches. Unfortunately, not only their datasets are limited to small scopes but also none of them has released their datasets. The detailed comparison between the datasets of these approaches and our OpenSiteRec is shown in Table 1.

Geo-Spotting [25] first extends the early data mining approaches with machine learning algorithms to better analyze the effectiveness of geographical and mobility features in site recommendation. This approach focuses on 3 fast food brands in New York City and analyzes the key factors of site recommendation separately for each brand with sufficient samples. ANNRR [5] proposes a semi-supervised feature selection method on heterogeneous urban data to predict bike trip demand for the bike sharing station recommendation in two cities. In this work, the site recommendation is limited to bike sharing station. PAM [31] utilizes the traffic information with partitioning around medoids algorithm to determine the optimal location of new ambulance station. Although there are a great number of candidate regions, there are only 34 existing stations available for training that may do harm to the credibility of results. BL-G-CoSVD [63] introduces bias learning and integrates both location and commercial features into SVD to recommend the suitable shop-type for each site. Contrary to the other approaches that predict the optimal region for a given brand, this approach predicts the shop-type for a given region with only 5 candidate shop-types so that the difficulty of this task is much lower. DD3S [59] learns to rank the candidate demand centers for two coffee shop brands and two chain hotel brands by predicting the number of customers at the given location with multiple spatial-temporal data sources. Since the candidate demand centers are predetermined by clustering the demand and supply gaps, there are only about 10 regions being determined as candidates, which is a small number. DeepStore [36] leverages a deep neural network on both dense and sparse features for predicting the consumption level of 49 stores in their surrounding areas of 13 cities to recommend the optimal site. However, the limited amount of data significantly increases the dependence of this task to the quality of features.

While these approaches consider the site recommendation of different cities individually, there are also some works focusing on knowledge transfer across different cities. CityTransfer [12] transfers knowledge from a source city to a target city via both inter- and intra-city views for site recommendation of 3 chain hotel brands in a new city. Specifically, it chooses Beijing or Shanghai as the source city and Xi'an or Nanjing as the target city to improve the performance on the cities with less stores. WANT [37] employs adversarial learning to diminish the distribution discrepancy between the source city and target city for predicting the consumption of stores in given areas. Then, it ranks the candidate areas in the target city according to the consumption for site recommendation. Similar to DeepStore, the dataset used by WANT is also small in size.

Despite their success, all these traditional site recommendation approaches rely heavily on feature engineering with simple model structures, especially the fine-grained manually-crafted features [12, 25, 36, 37, 59, 63], which are hard to design and may introduce human biases. Different from them, the recent approaches pay efforts in utilizing complex models to automatically capture the latent features from multi-source data for site recommendation, which is a new research trend of site recommendation. UrbanKG [34] constructs a knowledge graph from urban data, built upon which a relational graph neural network model is designed for efficient and effective site recommendation. UrbanKG utilizes a comprehensive dataset and the well-defined data structure is suitable for the research of site recommendation. However, this dataset focuses on the giant brands and has not been released for publicly available. O^2-SiteRec [60] conducts site recommendation by ranking the candidate regions with order number and delivery time from the courier capacity perspective in Online-to-Offline (O2O) stores of delivery platforms. This approach only uses the data of a small scope of brands from a single city, which is not sufficient. UUKG [41] proposes an urban knowledge graph as the foundation for downstream spatial-temporal prediction tasks. Although UUKG is not designed for site recommendation, the link prediction task on it shares many similarities with site recommendation. Differently, it only have 15 broad categories to distinguish the POIs rather than fine-grained brands and more than 80% of the POIs are from merely 2 categories. Therefore, the vanilla UUKG is not sufficient to well support site recommendation.

Compared with their datasets, our proposed OpenSiteRec has outstanding superiorities from four perspectives:

- **Wide-range and Large-scale**: Existing site recommendation approaches mainly consider a small range of brands or focus on a specific scenario, such as chain hotel and courier capacity. Therefore, they usually collect data only for the specific demand, which yields a small size of data in the final task. In contrast, OpenSiteRec takes all the brands and regions in the city into account and thus yields much larger data size.
- **Rich Commercial Features**: Existing approaches focus more on leveraging the features of regions for prediction while paying less attention to the features of brands. In contrast, OpenSiteRec provides more commercial features of brands and institutions via various types of relations among them.
- **Comprehensive Information**: Most of the existing approaches manually define the fine-grained features using original geographical and demographical information, which may lead to human bias and information loss. In contrast, OpenSiteRec models the data as a heterogeneous graph with different types of

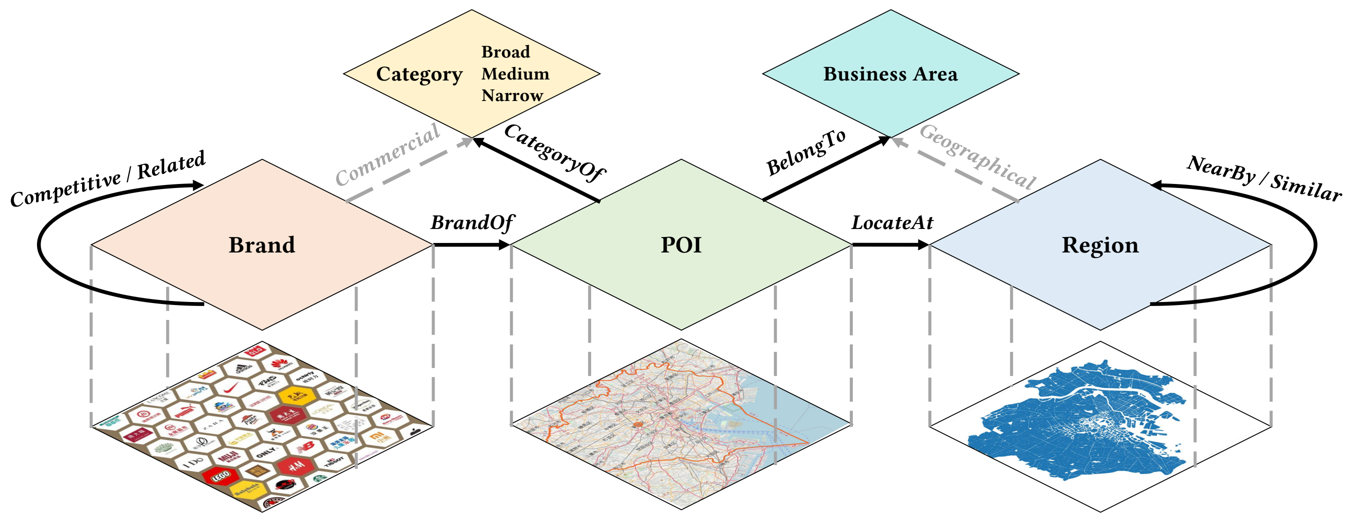


Figure 1: Schema of OpenSiteRec. The different types of entities and relations are represented by the polygons and arrows respectively. The solid arrows denote the definite relations between different entity types while the dotted arrows denote the indefinite relations that are not exactly defined but can be derived from existing information.

nodes and edges, which provides more comprehensive information, such as the competitive relations between brands, along with the original features.

- **Publicly Available**: None of the existing studies have released their datasets. To the best of our knowledge, OpenSiteRec is the first publicly available dataset for site recommendation. The released OpenSiteRec dataset will encourage continuous research in site recommendation.

3 DATA DESCRIPTION

3.1 Problem Definition

While most of the existing works formulate site recommendation as a ranking task ultimately, their data is usually used for a specific purpose and thus have different task definitions, such as store-type recommendation [63], consumption level prediction [36] and knowledge graph link prediction [34]. However, no matter what kinds of aspects of information they use or forms of task they apply, their ultimate objectives are the same to obtain the ranking list of sites. Since there is no precise consentient definition of site recommendation, we propose a formal definition as follows:

Definition 3.1. Let $\mathcal{B} = \{b_1, b_2, ..., b_M\}$ denote the brand set with M brands and $\mathcal{R} = \{r_1, r_2, ..., r_N\}$ denote the region set with N regions. Each POI that belongs to brand b_i and locates at region r_j contributes to a value $P_{ij} = 1$ in the matrix $\mathbf{P} = \{0, 1\} \in \mathbb{R}^{M \times N}$. Therefore, the site recommendation task aims at predicting a ranking list of candidate regions for each given brand.

Different from the definitions of other studies, our definition is more general that unifies the different concepts, such as brand, store and shop type, and requires no additional definition, e.g. the relation in knowledge graph link prediction. Moreover, such a definition is more straightforward that directly considers the ranking objective.

3.2 Schema Definition

In order to clarify the goal of data to collect and the final structure of dataset to construct, we deliver a schema as shown in Figure 1 for illustration. Such an overall schema has a graph structure, which consists of different types of entities to represent the real-world concepts and different types of edges between entities to indicate the commercial or geographical relations. Based on this schema, we elaborate the definition of each kind of entities and relations in details in the following paragraphs.

To be in line with the aforementioned problem and schema definition, we first define five types of entities to denote the real-world concepts following UrbanKG [34]:

- *Brand*: Brands denote either commercial brands in business that own multiple branches, e.g., Starbucks and Apple, or institutions that refer to special functions, e.g., Columbia University and United States Postal Service.
- *Category*: Categories represent the functions of the venues. Due to the significant functional differences between venues, we define three levels of categories (broad, medium and narrow) for classification. For example, a Starbucks store has the categories of 'Food and Beverage', 'Beverage Shop' and 'Coffee and Tea Shop' for broad, medium and narrow respectively.
- *POI*: POIs are the basic functional venues in a city, such as shops, restaurants and schools. Each POI has sufficient commercial and geographical information.
- *Business Area*: Business areas denote the special planned areas for business, where the venues are usually very dense.
- *Region*: Regions refer to the geographical divisions planned by the city governments. The principle and granularity of region division vary depending on the governments. For example, the regions in Chicago are large with only IDs while the regions in Tokyo are smaller with specific names like 'Ginza 1 Chōme'.

Table 2: Statistics of OpenSiteRec.

City	Geographic Unit	POI	Business Area	Region	Site (POI with Brand)	Brand	Category		
							Broad	Medium	Narrow
Chicago	1,471,416	16,154	77	801	8,044	969	10	39	107
New York City	2,678,932	41,403	71	2,325	14,189	2,702	10	43	129
Singapore	592,663	18,580	387	2,043	9,912	1,922	10	39	116
Tokyo	1,427,914	60,042	/	3,036	26,765	4,861	10	37	98

Based on the definitions, we define the relation types between these five types of entities. Each *POI* can be mapped to a *Brand* and a *Region* from the commercial and geographical aspects of it. Therefore, the *POI* serves as a bridge to connect *Brand* and *Region* to construct the dataset for site recommendation. Meanwhile, the *Category* and *Business Area* of a *POI* are uniquely determined. However, according to the real-world situation, a *Brand* may have multiple relations of *Category* and a *Business Area* may consist of several parts from multiple instances of *Region*. Moreover, there are also plenty of relations within *Brand* (*Competitive* and *Related*) and relations within *Region* (*NearBy* and *Similar*) that may be useful for site recommendation as follows:

- *Competitive*: This commercial relation means that two brands are competitive, e.g., KFC and McDonald's.
- *Related*: This commercial relation means that two brands belong to the same company or group, e.g., KFC and Pizza Hut both belong to Yum! Brands.
- *NearBy*: This geographical relation means that two regions are close in physical space, e.g., 'Ginza 1 Chōme' and 'Ginza 2 Chōme'.
- *Similar*: This commercial relation means that two different regions have similar distributions of POI category, e.g., 'Ginza 1 Chōme' and 'Toranomon 4 Chōme'.

3.3 Data Construction

Due to the requirements of data quality to produce reliable dataset, we choose four international metropolises for OpenSiteRec at present considering their information comprehensiveness and data integrity, which are *Chicago*, *New York City*, *Singapore* and *Tokyo*. On the basis of the schema, we collect data separately from open-source data sources for the aforementioned three aspects following ethical regulations.

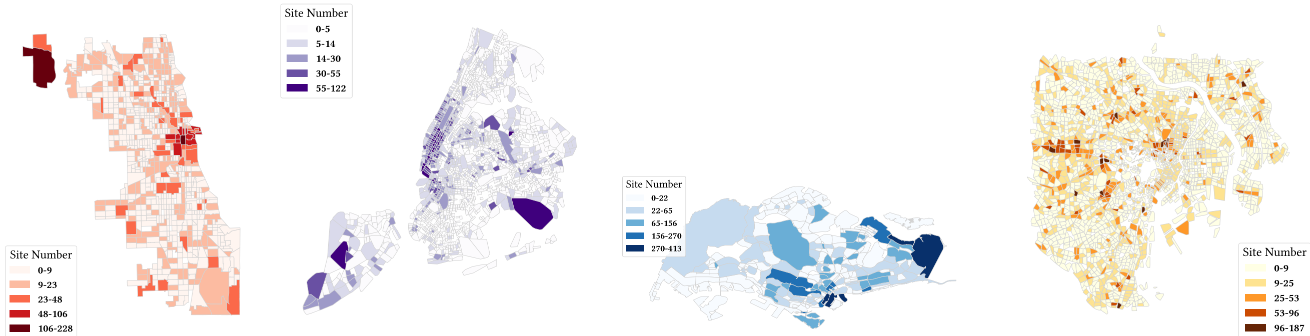
First, the site data, which is the core to connect commercial data and geographical data, is obtained by extracting the POIs from OpenStreetMap[1], licensed under the Open Data Commons Open Database License (ODbL) [2]. OpenStreetMap [14] is an open-source community-built map service that consists of three types of geographic unit, including nodes, ways and relations. Specifically, we extract the data from the data distribution service[3] on December 1st, 2022. For each geographic unit, OpenStreetMap provides a series of tags to describe its characteristics, such as name, brand and amenity. Here we filter the geographic units without name tag to identify POIs since nameless objects are mostly meaningless either and convert the original geographical location of POIs into their centroids' coordinates.

Then, we obtain the commercial data by assigning the POIs to the brands from Wikidata[4]. Unfortunately, the majority of POIs don't have any brand tags and the brand information is contained in their names. Meanwhile, there may be multiple brand names that correspond to the same brand. Therefore, it is essential to design an effective method for brand matching. To achieve more accurate and reliable brand matching, we apply a combination of phonetic matching (soundex [74]) and text matching (Jaro edit distance [8, 21]) algorithms. After the brand matching, the brands are grouped together into several disjoint sets and each set will choose one specific brand name which has record in Wikidata to represent the whole set. Through this process, we can successfully obtain the precise brand of POIs without ambiguity. For each POI, the hierarchical categories are determined based on its functional tags and the brand information from Wikidata. Specifically, the broad category and medium category types are manually defined and the narrow category types are directly extracted from Wikidata. Take a branch store of Starbucks as example, the broad category 'Food and Beverage' is determined by its functional tags 'amenity=restaurant' and the narrow category 'Cafe and Tea Shop' is obtained from the brand information of Starbucks on Wikidata while the medium category is then defined to be 'Beverage Shop'.

Next, the geographical data is collected from the free public data on the data portal of Chicago[5], New York City[6], Singapore[7] and Tokyo[8] governments. For the officially defined regions and business areas, we formulate them into polygon boundaries which consist of a series of coordinates. Subsequently, we determine the geographical assignment of POIs by computing the inclusion relationships between the boundaries of the regions or business areas and the centroids of the POIs.

After obtaining all the entities, we further identify the intra-relations within *Brand* and *Region* to provide more comprehensive commercial and geographical information. For the relations within *Brand*, i.e. *Competitive* and *Related*, we directly extract them by the provided statements from Wikidata. Specifically, we define all the other brands in the same lowest category as *Competitive* and define the brands in 'See Also' section as *Related*. For the relations within *Region*, we calculate the shortest distances between pairs of regions to identify *NearBy* relations with a maximum threshold 0.5km and

[1]https://www.openstreetmap.org/
[2]https://opendatacommons.org/licenses/odbl/
[3]https://download.bbbike.org/osm/bbbike/
[4]https://www.wikidata.org/
[5]https://data.cityofchicago.org/
[6]https://opendata.cityofnewyork.us/
[7]https://data.gov.sg/
[8]https://www.data.go.jp/



(a) **Chicago** (b) **New York City** (c) **Singapore** (d) **Tokyo**

Figure 2: Site distributions of regions in different cities. The uneven color distributions indicate that the site distributions are quite imbalanced on the region side.

Table 3: An example of an instance in OpenSiteRec.

ID 259964052	**Name** Starbucks	**City** Tokyo
Category B. Food and Beverage	**Category M.** Beverage Shop	**Category N.** Coffee and Tea Shop
Brand Starbucks	**Competitive** Tully's Coffee, ...	**Related** McDonald's, ...
Longitude 139.7390476	**Latitude** 35.684236	**District** Chiyoda City
Region Kōjimachi 3 Chōme	**NearBy** Kioichō, ...	**Similar** Kōjimachi 4 Chōme, ...

we calculate the cosine similarity of narrow category distributions of POIs to identify *Similar* relations with a minimum threshold 0.9.

3.4 Statistics and Usage

Through the whole process of data collection and processing, we finally obtain the well-formulated and comprehensive OpenSiteRec dataset. The detailed statistics are shown in Table 2. Here we provide an example of a branch store of Starbucks in Tokyo as illustrated in Table 3.

To better facilitate the usage of OpenSiteRec, we further explore the characteristics of OpenSiteRec by analyzing the dataset distributions. Due to the urban planning and lifestyles of people, the distributions of site counts of categories are different in each city. From the comparison of distributions in Figure 3, we can explore the highlights in these four metropolises. However, on the perspectives of brands and regions, the distributions are much more imbalanced. The site counts of different brands are extremely imbalanced that the top 10% of brands occupy over 50% of sites. This phenomenon indicates that the giant brands are actually dominating the commerce in every city. Similarly, the uneven color distributions in Figure 2 show that the vast majority of POIs (up to 80%) locate in several centres or sub-centres composed of a few regions while most regions

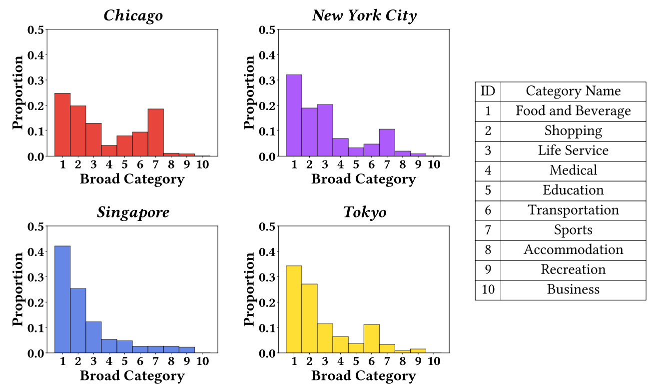


Figure 3: Distributions of site counts of categories in different cities. There are slight differences on the distributions among these four cities, which indicates the differences in urban planning and lifestyles of people.

have only a few POIs. On the basis of these statistics, we conclude five unique characteristics of site recommendation compared with other top-N recommendation tasks: the data is quite **sparse**; the **city-specific** characteristics are significant; the data distribution is extremely **imbalanced**; the **domain-specific features** are very crucial; the correlations are highly **complex**.

4 BENCHMARKING EXPERIMENTS

4.1 Experimental Settings

4.1.1 Dataset Split & Evaluation Metrics. Since the original OpenSiteRec is extremely imbalanced on both brands and regions, we filter the dataset by 5-core setting on the brands (all the brands have possessed at least 5 POIs). Specifically, we randomly split the POIs of each brand with 70%, 10% and 20% as training, validation and test sets, respectively. To assess the model performance, we choose the widely-used evaluation metrics **Recall@20** and **nDCG@20** and regard all regions as candidates, i.e. all-ranking protocol.

4.1.2 Baselines. To explore the effectiveness of proposed OpenSiteRec in practical application, we conduct experiments of site recommendation with several representative recommendation models as the benchmarks, including machine learning models, collaborative filtering [52] models, click-through rate (CTR) prediction [51] models and graph-based models.

For machine learning models, we choose:

- **LR** [19], namely logistic regression, is a simple yet effective model in classification.
- **GBDT** [11], namely gradient boosting decision tree, is an ensemble model with decision tree model as the backbone.
- **SVC** [2] employs support vector machine (SVM) to tackle the classification problem.
- **RankNet** [3] is a famous learning to rank architecture in recommendation. Specifically, we utilize a two-layer neural network as the backbone here.

For collaborative filtering models, we choose:

- **MF-BPR** [49] is a variant of Matrix Factorization (MF) [27] optimized by the Bayesian personalized ranking (BPR) loss.
- **NeuMF** [17] combines neural network and Matrix Factorization (MF) for collaborative filtering with point-wise loss.
- **FISM** [24] extends MF by aggregating the item embeddings of interacted items to represent the user via item similarity.
- **NAIS** [16] introduces attention mechanism onto FISM to conduct weighted aggregation of items.

For CTR prediction models, we choose:

- **DNN** [9] applies deep neural network to capture the complex interaction between features for CTR prediction.
- **Wide&Deep** [6] jointly utilizes linear transformation and DNN for CTR prediction.
- **DeepFM** [13] combines the factorization machine (FM) and DNN to model the first-, second- and high-order feature interactions.
- **xDeepFM** [32] leverages the compressed interaction network (CIN) to achieve vector-wise feature interactions.

For graph-based models, we choose:

- **GC-MC** [55] is a general graph neural network architecture for recommendation.
- **GraphRec** [10] introduces graph neural network into social recommendation by aggregating the embeddings with social relationships. Here we remove the social aggregation component.
- **NGCF** [56] namely Neural Graph Collaborative Filtering, conducts graph message passing on user-item interaction graph for recommendation.
- **LightGCN** [15] simplifies the graph convolutional network with only neighborhood aggregation for collaborative filtering.

4.1.3 Implementation Details. The experiments are implemented with two NVIDIA GeForce RTX 2080Ti GPUs. Specifically, for LR, GDBT and SVC, we implement them with scikit-learn [45] 1.0.2. For other models that involve low-dimensional embeddings, we implement them with PyTorch [43] 1.12.1 and set the embedding dimension as 100. The model parameters are initialized with Xavier initialization and optimized by Adam [26]. For all models, we tune hyper-parameters with the performance on the validation set via grid search. The detailed implementation codes could be found at https://github.com/HestiaSky/OpenSiteRec/.

4.2 Benchmark Results

The benchmark results of the baselines on OpenSiteRec are shown in Table 4. In order to deliver our insights for future research, we analyze the experimental results and summarize the following points:

- Traditional machine learning methods are not capable to handle the complex scenario of site recommendation. Although LR, GBDT and SVC are essentially different, they all converge to the same local optimal point and thus have the same performance. Given the situation with highly limited data but highly rich features, it is extremely difficult for them not to over-fit to the training set. Thus, they fail to generalize to new POIs and are not suitable for site recommendation.
- The pair-wise loss is significantly better than the point-wise loss. As shown in the results, the models with pair-wise loss (i.e. BPR loss) including RankNet, MF-BPR, FISM, NAIS, NGCF, LightGCN, all outperform the models with point-wise loss (i.e. BCE loss) including NeuMF, DNN, Wide&Deep, DeepFM, xDeepFM, GC-MC, GraphRec. This phenomenon indicates that the pair-wise loss is more suitable than the point-wise loss in site recommendation.
- The feature interaction has marginal effects on performance. From DNN to xDeepFM, the degree of feature interaction is increasing but the performance improvement is not significant. This may be because the correlations of features are either fully dependent (broad category and narrow category) or fully independent (brand category and geographical coordinate).
- The high-order interactions between brands and regions are crucial. Under the same condition of other factors, the graph-based models are substantially better than others. Since the data are highly sparse in site recommendation, high-order interactions are important to make correct predictions. Therefore, exploiting the graph representation learning techniques to better model the high-order interactions between brands and regions, especially the explicitly defined relations like *Competitive*, are beneficial to obtain high performance in site recommendation.

4.3 Long-tail Scenario

Since the imbalance problem is severe in site recommendation, we conduct an additional experiment to evaluate some representative baseline models under the long-tail scenario. Specifically, we regard the bottom 90% of regions with fewer POIs as the long-tailed regions which totally occupy less than 50% of POIs and are only for testing. As shown in Figure 4, the performances of the baseline models drop dramatically on these long-tailed regions, which implies a big challenge in site recommendation. From the perspective of the cities, a greater drop in performance indicates a higher degree of centrality in urban planning. According to the results, *Chicago* and *Singapore* have a higher degree of centrality than *New York City* and *Tokyo*, which means the giant brands have more dominant positions in *Chicago* and *Singapore*. However, to better achieve fairness, which is crucial to provide forward-looking recommendation results, addressing this imbalance issue is really important.

4.4 Potential Applications

Besides site recommendation, OpenSiteRec also support many other potential applications [68–73], including transportation demand prediction [33, 61], electric vehicle charging recommendation [66,

Table 4: Benchmarking experimental results on OpenSiteRec. Best results in each category are highlighted in bold.

Method	Chicago		New York City		Singapore		Tokyo	
	Rec@20	nDCG@20	Rec@20	nDCG@20	Rec@20	nDCG@20	Rec@20	nDCG@20
LR	0.1203	0.0868	0.0886	0.0655	0.1784	0.1336	0.0795	0.0594
GBDT	0.1203	0.0868	0.0886	0.0655	0.1784	0.1336	0.0795	0.0594
SVC	0.1203	0.0868	0.0886	0.0655	0.1784	0.1336	0.0795	0.0594
RankNet	**0.2269**	**0.1427**	**0.1224**	**0.0654**	**0.4297**	**0.2271**	**0.1213**	**0.0667**
MF-BPR	0.2494	0.1465	0.1702	0.0917	0.4430	0.2351	0.1323	**0.0781**
NeuMF	0.1942	0.1293	0.1200	0.0576	0.4289	0.2236	0.1225	0.0639
FISM	**0.2547**	**0.1468**	**0.1745**	0.0928	**0.4583**	**0.2382**	**0.1343**	0.0734
NAIS	0.2534	0.1432	0.1743	**0.0964**	0.4557	0.2380	0.1328	0.0774
DNN	**0.1927**	**0.1311**	**0.1215**	**0.0568**	**0.4268**	**0.2204**	**0.1284**	**0.0648**
Wide&Deep	0.1910	0.1284	0.1193	0.0526	0.4202	0.2174	0.1225	0.0647
DeepFM	0.1898	0.1254	0.1184	0.0531	0.4177	0.2182	0.1225	**0.0648**
xDeepFM	0.1886	0.1225	0.1157	0.0515	0.4178	0.2180	0.1237	0.0647
GC-MC	0.2332	0.1657	0.1514	0.0513	0.4685	0.2317	0.1558	0.0884
GraphRec	0.2365	0.1640	0.1538	0.0550	0.4697	0.2293	0.1594	0.0905
NGCF	0.2866	0.1838	0.1920	**0.1102**	0.4929	0.2674	0.1619	0.1012
LightGCN	**0.2875**	**0.1902**	**0.2087**	0.1088	**0.5013**	**0.2745**	**0.1751**	**0.1068**

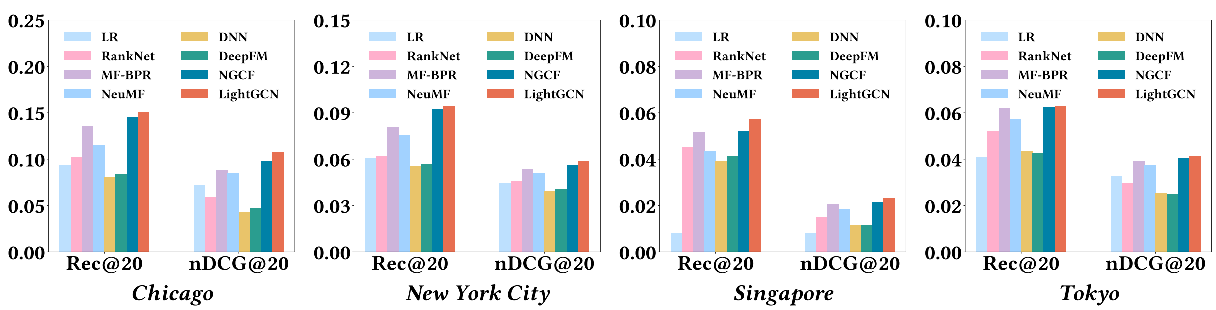


Figure 4: Experimental results of long-tailed regions on OpenSiteRec.

67] and high-potential startup detection [64]. Here, we will discuss the significance and the feasibility of two potential tasks: brand entry forecasting [38, 53] and business area planning [62].

4.4.1 Brand Entry Forecasting. People in different cities have different preferences for brands due to culture, history or lifestyle [29, 42]. These factors along with the commercial strategies of brands result in the different brand distributions in different cities [44]. Typically, some brands are more international that have been spread all over the world while some brands are more local at present but also have a great potential to expand to other cities [4]. A feasible way is to exploit the existing brands in different cities to mine the brands with a high probability to succeed in other cities [64, 65], i.e. brand entry forecasting. Since OpenSiteRec provides plenty of brands in four metropolises, it is credible to carry out such research.

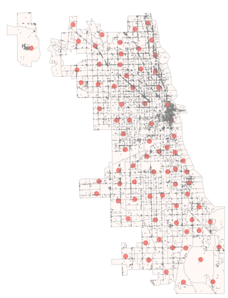
To demonstrate the applicability of our OpenSiteRec to support brand entry forecasting, we present a case study by ranking the brands of *Fast Food* and *Cafe & Dessert* in different cities. As illustrated in Table 5, many popular international brands like *Starbucks* have multiple similar brands that are popular in local, such as *Coffee Bean* in *Singapore* and *Doutor* in *Tokyo*. The fact that these local brands have the ability to pose a position in the fierce commercial competition with the international giant brands not only demonstrates their success in commerce so far, but also indicates their adequate competitiveness to success in other places in the future.

4.4.2 Business Area Planning. To form a strong scale effect, planning business area [62] (i.e. central business district) is crucial in the city development. Generally, business area planning takes many factors into account, such as population, traffic convenience and surrounding environments. Since POI distributions and geographical information implicitly indicate these factors [25], we deem that OpenSiteRec is beneficial to support business area planning.

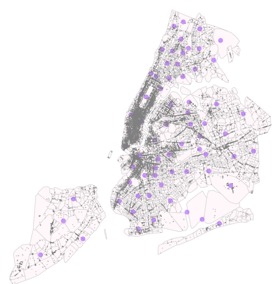
To verify this idea, we visualize the planned business area and POI distributions for comparison. From Figure 5, we can see that the distributions of business areas are relatively uniform in the cities and have high coincidence with the traffic hubs. Such a regular pattern is consistent with the POI distributions and it is also true that the more concentrated the POIs are, the more business districts are planned. Our OpenSiteRec provides tens of thousands of POIs

Table 5: Comparison of the top brands of *Fast food* (in red) and *Cafe & Dessert* (in blue) in different cities. The dark color denotes the more local special brand that hasn't entered all the other cities while the light color denotes the more international chain brand that has opened stores in every city.

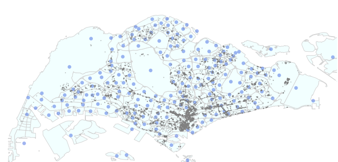
Chicago		New York City		Singapore		Tokyo	
Brand	**Portion**	**Brand**	**Portion**	**Brand**	**Portion**	**Brand**	**Portion**
Dunkin'	6.15%	Dunkin'	4.24%	Starbucks	3.33%	**Doutor**	2.65%
Subway	5.32%	Starbucks	3.17%	McDonald's	3.01%	Starbucks	1.98%
Starbucks	5.17%	McDonald's	1.98%	**Toast Box**	1.98%	McDonald's	1.80%
McDonald's	4.26%	Subway	1.63%	KFC	1.79%	**Matsuya**	1.77%
Burger King	1.85%	Burger King	0.98%	Subway	1.77%	**Sukiya**	1.18%
Baskin-Robbins	1.55%	Popeyes	0.83%	**Coffee Bean**	1.77%	**MOS Burger**	1.17%
Popeyes	1.36%	**Chipotle**	0.83%	**Ya Kun Kaya Toast**	1.54%	**Tully's Coffee**	1.17%
Potbelly	1.36%	Domino's	0.54%	**Kopitiam**	1.10%	**Yoshinoya**	1.02%



(a) **Chicago**



(b) **New York City**



(c) **Singapore**

Figure 5: Visualization of the planned business area of different cities with POI distributions. Specifically, the government of Tokyo has not released the official business area planning so Tokyo is excluded.

along with their categories and other characteristics. By analyzing these characteristics of all the POIs in an area, it is able to effectively grasp the essential factors for business area planning.

5 DISCUSSION CONCLUSION

Site recommendation is an important and beneficial information retrieval task for the development of brands or institutions in real-world applications, which has been established for decades. Despite the success of recent approaches in site recommendation, none of them have released their datasets, which brings inconvenience for the researchers and is harmful to the research in this area, and most of them only focus on a small scope of site recommendation, yielding limited significance and impact. Therefore, we collect, construct and release the first open dataset OpenSiteRec for site recommendation, which consists of multi-source data from four international metropolises and is comprehensive to support the following research. To verify the applicability of our OpenSiteRec and the suitability of the existing recommendation models on site recommendation task, we conduct benchmarking experiments of totally 16 representative baseline models on OpenSiteRec.

The greatest limitation is the lack of temporal dimension, which means OpenSiteRec collects the data at a specific time without variation from urban development. Therefore, we are considering to add the temporal dimension to deliver information at multiple time points. In addition, we have also planned to expand OpenSiteRec by adding more cities to better support following research.

In conclusion, site recommendation is is still an underestimated topic considering its significance in modern business and its predicted impacts. Therefore, we believe that the emergence of our OpenSiteRec dataset will strongly promote the research on it and help the development of business intelligence in the next few years.

ACKNOWLEDGEMENTS

This research was partially supported by Research Impact Fund (No.R1015-23), APRC - CityU New Research Initiatives (No.9610565, Start-up Grant for New Faculty of CityU), CityU - HKIDS Early Career Research Grant (No.9360163), Hong Kong ITC Innovation and Technology Fund Midstream Research Programme for Universities Project (No.ITS/034/22MS), Hong Kong Environmental and Conservation Fund (No.88/2022), and SIRG - CityU Strategic Interdisciplinary Research Grant (No.7020046, No.7020074), Ant Group (CCF-Ant Research Fund, Ant Group Research Fund), Huawei (Huawei Innovation Research Program), Tencent (CCF-Tencent Open Fund, Tencent Rhino-Bird Focused Research Program), CCF-BaiChuan-Ebtech Foundation Model Fund, and Kuaishou. This research was also supported by the Ministry of Education, Singapore, under its Academic Research Fund (Tier 2 Awards MOE-T2EP20221-0013 and MOE-T2EP20220-0011). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

REFERENCES

[1] Oded Berman and Dmitry Krass. 2002. The generalized maximal covering location problem. *Comput. Oper. Res.* 29, 6 (2002), 563–581.

[2] Erin J. Bredensteiner and Kristin P. Bennett. 1999. Multicategory Classification by Support Vector Machines. *Comput. Optim. Appl.* 12, 1-3 (1999), 53–79.

[3] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. 2005. Learning to rank using gradient descent. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005 (ACM International Conference Proceeding Series)*, Vol. 119. ACM, 89–96.

[4] Gregory S. Carpenter and Kent Nakamoto. 1990. Competitive Strategies for Late Entry into a Market with a Dominant Brand. *Management Science* 36 (1990), 1268–1278.

[5] Longbiao Chen, Daqing Zhang, Gang Pan, Xiaojuan Ma, Dingqi Yang, Kostadin Kushlev, Wangsheng Zhang, and Shijian Li. 2015. Bike sharing station placement leveraging heterogeneous urban open data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*. ACM, 571–575.

[6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2016, Boston, MA, USA, September 15, 2016*. ACM, 7–10.

[7] Richard L. Church and Alan T. Murray. 2008. Business Site Selection, Location Analysis and GIS.

[8] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), August 9-10, 2003, Acapulco, Mexico*. 73–78.

[9] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*. ACM, 191–198.

[10] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Yihong Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph Neural Networks for Social Recommendation. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 417–426.

[11] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29 (2001), 1189–1232.

[12] Bin Guo, Jing Li, Vincent W. Zheng, Zhu Wang, and Zhiwen Yu. 2017. CityTransfer: Transferring Inter- and Intra-City Knowledge for Chain Store Site Recommendation based on Multi-Source Urban Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (2017), 135:1–135:23.

[13] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. ijcai.org, 1725–1731.

[14] Mordechai (Muki) Haklay and Patrick Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Comput.* 7, 4 (2008), 12–18.

[15] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. ACM, 639–648.

[16] Xiangnan He, Zhankui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. 2018. NAIS: Neural Attentive Item Similarity Model for Recommendation. *IEEE Trans. Knowl. Data Eng.* 30, 12 (2018), 2354–2366.

[17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, 173–182.

[18] Tony Hernández and David J. Bennison. 2000. The art and science of retail location decisions. *International Journal of Retail & Distribution Management* 28 (2000), 357–367.

[19] David W. Hosmer and Stanley Lemeshow. 1991. Applied Logistic Regression.

[20] David L. Huff. 1966. A Programmed Solution for Approximating an Optimum Retail Location. *Land Economics* 42 (1966), 293–303.

[21] Matthew A. Jaro. 1989. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *J. Amer. Statist. Assoc.* 84 (1989), 414–420.

[22] Pablo Jensen. 2006. Network-based predictions of retail store commercial categories and optimal locations. *Physical review. E, Statistical, nonlinear, and soft matter physics* 74 3 Pt 2 (2006), 035101.

[23] Pablo Jensen. 2009. Analyzing the Localization of Retail Stores with Complex Systems Tools. In *Advances in Intelligent Data Analysis VIII, 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31 - September 2, 2009. Proceedings (Lecture Notes in Computer Science)*, Vol. 5772. Springer, 10–20.

[24] Santosh Kabbur, Xia Ning, and George Karypis. 2013. FISM: factored item similarity models for top-N recommender systems. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. ACM, 659–667.

[25] 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 *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. ACM, 793–801.

[26] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[27] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

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

[29] Donald R. Lehmann and Yigang Pan. 1994. Context Effects, New Brand Entry, and Consideration Sets. *Journal of Marketing Research* 31 (1994), 364 – 374.

[30] Xinhang Li, Zhaopeng Qiu, Jiacheng Jiang, Yong Zhang, Chunxiao Xing, and Xian Wu. 2024. Conditional Cross-Platform User Engagement Prediction. *ACM Trans. Inf. Syst.* 42, 1 (2024), 6:1–6:28.

[31] Yuhong Li, Yu Zheng, Shenggong Ji, Wenjun Wang, Leong Hou U, and Zhiguo Gong. 2015. Location selection for ambulance stations: a data-driven approach. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, November 3-6, 2015*. ACM, 85:1–85:4.

[32] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. ACM, 1754–1763.

[33] Hao Liu, Qiyu Wu, Fuzhen Zhuang, Xinjiang Lu, Dejing Dou, and Hui Xiong. 2021. Community-Aware Multi-Task Transportation Demand Prediction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 320–327.

[34] Yu Liu, Jingtao Ding, and Yong Li. 2021. Knowledge-driven Site Selection via Urban Knowledge Graph. *CoRR* abs/2111.00787 (2021).

[35] Yu Liu, Jingtao Ding, and Yong Li. 2022. Developing knowledge graph based system for urban computing. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Knowledge Graphs* (2022).

[36] 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.

[37] Yan Liu, Bin Guo, Daqing Zhang, Djamal Zeghlache, Jingmin Chen, Ke Hu, Sizhe Zhang, Dan Zhou, and Zhiwen Yu. 2021. Knowledge Transfer with Weighted Adversarial Network for Cold-Start Store Site Recommendation. *ACM Trans. Knowl. Discov. Data* 15, 3 (2021), 47:1–47:27.

[38] Stijn Maesen and Lien Lamey. 2022. The Impact of Organic Specialist Store Entry on Category Performance at Incumbent Stores. *Journal of Marketing* 87 (2022), 97 – 113.

[39] Michael T. Marsh and David A. Schilling. 1994. Equity measurement in facility location analysis: A review and framework. *European Journal of Operational Research* 74 (1994), 1–17.

[40] Daniel McFadden. 1977. Modelling the Choice of Residential Location. *Transportation Research Record* (1977).

[41] Yansong Ning, Hao Liu, Hao Wang, Zhenyu Zeng, and Hui Xiong. 2023. UUKG: Unified Urban Knowledge Graph Dataset for Urban Spatiotemporal Prediction. *CoRR* abs/2306.11443 (2023).

[42] Yigang Pan and Donald R. Lehmann. 1993. The Influence of New Brand Entry on Subjective Brand Judgments. *Journal of Consumer Research* 20 (1993), 76–86.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 8024–8035.

[44] Koen H. Pauwels and Shuba Srinivasan. 2004. Who Benefits from Store Brand Entry. *Marketing Science* 23 (2004), 364–390.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[46] Sergio Porta, Vito Latora, Fahui Wang, Salvador Rueda, Emanuele Strano, Salvatore Scellato, Allessio Cardillo, Eugenio Belli, Francisco Cardenas, Berta Cormenzana, and Laura Latora. 2012. Street Centrality and the Location of Economic Activities in Barcelona. *Urban Studies* 49 (2012), 1471 – 1488.

[47] Sergio Porta, Emanuele Strano, Valentino Iacoviello, Roberto Messora, Vito Latora, Alessio Cardillo, Fahui Wang, and Salvatore Scellato. 2009. Street Centrality and Densities of Retail and Services in Bologna, Italy. *Environment and Planning B: Planning and Design* 36 (2009), 450 – 465.

[48] M. Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho. 2016. Urban planning and building smart cities based on the Internet of Things using Big Data analytics. *Comput. Networks* 101 (2016), 63–80.

[49] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*. AUAI Press, 452–461.

[50] Charles S. Revelle and Horst A. Eiselt. 2005. Location analysis: A synthesis and survey. *Eur. J. Oper. Res.* 165 (2005), 1–19.

[51] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*. ACM, 521–530.

[52] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*. ACM, 285–295.

[53] Satheesh Seenivasan and Debabrata Talukdar. 2016. Competitive Effects of Wal-Mart Supercenter Entry: Moderating Roles of Category and Brand Characteristics. *Journal of Retailing* 92 (2016), 218–225.

[54] Meng Shao, Zhi Han, Jin Wei Sun, Chengsi Xiao, Shu lei Zhang, and Yuan xu Zhao. 2020. A review of multi-criteria decision making applications for renewable energy site selection. *Renewable Energy* 157 (2020), 377–403.

[55] Rianne van den Berg, Thomas N. Kipf, and Max Welling. 2017. Graph Convolutional Matrix Completion. *CoRR* abs/1706.02263 (2017).

[56] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 165–174.

[57] Yunna Wu, Chao Xie, Chuanbo Xu, and Fang Li. 2017. A Decision Framework for Electric Vehicle Charging Station Site Selection for Residential Communities under an Intuitionistic Fuzzy Environment: A Case of Beijing. *Energies* 10 (2017), 1270.

[58] Yunna Wu, Meng Yang, Haobo Zhang, Kaifeng Chen, and Yang Wang. 2016. Optimal Site Selection of Electric Vehicle Charging Stations Based on a Cloud Model and the PROMETHEE Method. *Energies* 9 (2016), 1–20.

[59] 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 *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2016, Burlingame, California, USA, October 31 - November 3, 2016*. ACM, 40:1–40:10.

[60] 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 *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*. IEEE, 525–538.

[61] Junchen Ye, Leilei Sun, Bowen Du, Yanjie Fu, and Hui Xiong. 2021. Coupled Layer-wise Graph Convolution for Transportation Demand Prediction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 4617–4625.

[62] Wenhao Yu, Tinghua Ai, and Shiwei Shao. 2015. The analysis and delimitation of Central Business District using network kernel density estimation. *Journal of Transport Geography* 45 (2015), 32–47.

[63] Zhiwen Yu, Miao Tian, Zhu Wang, Bin Guo, and Tao Mei. 2016. Shop-Type Recommendation Leveraging the Data from Social Media and Location-Based Services. *ACM Trans. Knowl. Discov. Data* 11, 1 (2016), 1:1–1:21.

[64] Shengming Zhang, Hao Zhong, Zixuan Yuan, and Hui Xiong. 2021. Scalable Heterogeneous Graph Neural Networks for Predicting High-potential Early-stage Startups. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*. ACM, 2202–2211.

[65] Ting Zhang, Xiaowei Zhu, and Qinglong Gou. 2017. Demand Forecasting and Pricing Decision with the Entry of Store Brand under Various Information Sharing Scenarios. *Asia Pac. J. Oper. Res.* 34 (2017), 1740018:1–1740018:26.

[66] Weijia Zhang, Hao Liu, Jindong Han, Yong Ge, and Hui Xiong. 2022. Multi-Agent Graph Convolutional Reinforcement Learning for Dynamic Electric Vehicle Charging Pricing. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*. ACM, 2471–2481.

[67] Weijia Zhang, Hao Liu, Fan Wang, Tong Xu, Haoran Xin, Dejing Dou, and Hui Xiong. 2021. Intelligent Electric Vehicle Charging Recommendation Based on Multi-Agent Reinforcement Learning. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 1856–1867.

[68] Zijian Zhang, Ze Huang, Zhiwei Hu, Xiangyu Zhao, Wanyu Wang, Zitao Liu, Junbo Zhang, S. Joe Qin, and Hongwei Zhao. 2023. MLPST: MLP is All You Need for Spatio-Temporal Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*. ACM, 3381–3390.

[69] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. 2023. PromptST: Prompt-Enhanced Spatio-Temporal Multi-Attribute Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*. ACM, 3195–3205.

[70] Xiangyu Zhao and Jiliang Tang. 2018. Crime in Urban Areas: : A Data Mining Perspective. *SIGKDD Explor.* 20, 1 (2018), 1–12.

[71] Xiangyu Zhao, Tong Xu, Yanjie Fu, Enhong Chen, and Hao Guo. 2017. Incorporating Spatio-Temporal Smoothness for Air Quality Inference. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*. IEEE Computer Society, 1177–1182.

[72] Xiangyu Zhao, Tong Xu, Qi Liu, and Hao Guo. 2016. Exploring the Choice Under Conflict for Social Event Participation. In *Database Systems for Advanced Applications - 21st International Conference, DASFAA 2016, Dallas, TX, USA, April 16-19, 2016, Proceedings, Part I (Lecture Notes in Computer Science)*, Vol. 9642. Springer, 396–411.

[73] Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Jian Qiao Yu. 2023. SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

[74] Justin Zobel and Philip W. Dart. 1996. Phonetic String Matching: Lessons from Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*. ACM, 166–172.

[75] Tezcan Kaşmer Şahin, Saffet Ocak, and Mehmet Top. 2019. Analytic hierarchy process for hospital site selection. *Health Policy and Technology* (2019).