2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)

LOCAST: Optimal location casting by crowdsourcing and open data integration

*

Konstantinos Platis
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
platiskp@csd.auth.gr

Ilias Dimitriadis
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
idimitriad@csd.auth.gr

Athena Vakali
School of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
avakali@csd.auth.gr

***Abstract*—Social media dominance largely affects multi-store brands success potential. How can a brand choose the optimal place to locate its stores, given the social media pulse? Which are the suitable metrics to guide such challenging decisions? This work addresses such crucial problems by a novel location casting approach which extracts and integrates knowledge from open data and social media, providing specific indicators and a systematic pipeline for effective locations casting. Emphasis is placed on how the derived knowledge will assess the particular characteristics of accessibility, interest, and centrality to identify fine grained urban indicators. The proposed pipeline predicts the success potential of a chain store's location, under individual or combined such indicators individually. The experimentation under qualitative tests, indicates that the proposed approach provides reliable estimations of brand's locations suitability, and also outperforms existing similar state-of-the-art approaches.**

***Index Terms*—Optimal Location, Social Data Mining, Success Indicators, Open Data, Crowdsourcing**

I. Introduction

Detection of an optimal and promising location for a brand's store, is a well studied and complex problem under the area of optimal location detection. Mostly static data sources have been utilized for such detection tasks, when new solutions are required in our era of dynamic data emerging from semantic and social Web sources. The open nature and the abundance of social media threads offers location declaration capabilities which can enhance knowledge extraction features and social phenomena detection. The so called **crowdsourcing data** along with cities' **open data** collections provide valuable sources to allow location's and areas significance detection. Such dynamic data sources can be harvested to identify indicators of success for a brand's location. Up to now, such indicators mostly involve geographical and internal factors and not other,more contextual, urban city characteristics like the competitiveness or the density of stores in a given area, the urban data granularity, etc.

In this paper, the open problem of defining rich and contextualized indicators for location's detection is addressed, while focus is placed on fully exploiting the dynamic nature of open and social media data. Three new urban indicators are proposed with emphasis on dynamic city data, namely *crowdsourcing data* and *open data*, integration. An important contribution of this work is that the proposed next indicators follow a *fine-grained* approach which embeds important contextual characteristics :

- the *Concentration Impact Index*, which outlines the presence of public transport in the near area based on open city data;
- the *Interest Level Index*, which uses data crawled from social media combined with open city data to identify the customers activity nearby;
- the *Route Frequency Index*, an indicator to capture how busy a street on a daily basis, extracted using both Open and Location Based Social Networks data.

Finally, the so called LOCAST framework is proposed which involves a pipeline of different modules and functionalities to support the process of collecting the data for the extraction of the urban indicators which then provide input to the evaluation of beneficial store's locations. In summary, the most suitable store locations are identified based on their potential to attract the highest number of the social media presence declarations (the so called check-ins) among a list of possible candidate locations.

The rest of the paper is structured as follows : Initially, in section II previous related work is overviewed and summarized to highlight the most significant solutions for the optimal store location problem. Then, in section III a detailed description for each proposed indicator is given along with the algorithms used to estimate indicators values . The different functionalities of the overall proposed LOCAST framework are described in section IV. Finally, in section V, the use of LOCAST is demonstrated under a real world dataset and the emerging results are compared with state-of-the-art approaches to showcase LOCAST potential and capacity. Conclusions and future work are finally discussed in VI.

This work has been supported by the European Union Horizon 2020 Research and Innovation Programme under Grant Agreements No. 780121 and No. 710659.

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00-60

II. Related Work

Collection of data (such as venues, check-ins) from Location Based Social Networks (such as Foursquare, Twitter) has supported earlier work in the area such as in "Geospotting". In this approach has utilized city areas' geographic and mobility features (density, popularity of the area etc.) to characterize the region near each chain store [1]. Particular learning models utilized in ranking and regression algorithms have been tested with the popular Normalized Discounted Cumulative Gain (NDCG) metric, to generate a prediction model for chain stores in a specific area.

Wang et al [2] has approached the optimal location problem by focusing just on a specific domain (i.e. restaurants). Instead of taking into account geographic and mobility features, the competitiveness and the attractiveness of the area have been prioritized, by leveraging user-generated reviews on social media. Since [3] uses "private" data from call records in the area, the method proposed in that article extends the approach of [1] and suggests the use of three **more fine-grained** geographical and mobility features that even reach the street level and are not limited to certain categories or type of stores. Similarly, Lin et al in their work "Where is the Goldmine?" [4] have pointed out the best positions to place a new food store by mining data from Facebook pages based on features like the type of neighboring stores, food-related and all hotspots , by a machine learning technique, called Gradient boosting.

Most of the existing research, refers to chain stores of low scale (e.g. up to 2-3 stores), and the capacity of the proposed solutions to handle huge and popular chain stores (like Starbucks or McDonald's) has not been fully showcased. This problem was addressed by Li et al in [3] who have tested multiple scales for chain stores and proposed ChainRec, a framework for chain store placement recommendation based on its scale. They used a large set of features, both of mobility and geographical context. An original concept presented by the same research was the usage of Places of Interest (PoI)[1] as a feature, i.e. the places which mostly attract visitors in a specific area. However, they focused on general categories of PoI and not individual stores, while their data is mostly closed and under commercial licenses and not open or social. Finally, the optimal location problem has also been related to the needs of placing service points like in [5], where new features and algorithms were proposed to find the optimal placement of bike sharing stations in urban areas. This research effort has showcased the need to monitor social activities (trends, demographics and POIs), in the optimal location detection process.

III. Location Casting Indicators

The proposed work is mostly inspired by the ideas of [1] and [3], which follow machine learning techniques based on several features. Here, the **Concentration Impact** considers both mobility and geographical context and it is related to the public transport stations near the location of interest, the **Interest level** is calculated based on POIs that can be individual stores or monuments nearby, while the **Route Frequency Index** is introduced for the first time and can provide specific information on the street level. Such characteristics contribute in defining the next indicators of success.

[1]https://en.wikipedia.org/wiki/Point_of_interest)

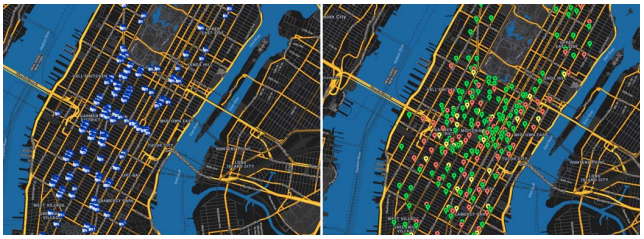

Fig. 1. The NY underground stations as collected in the research from the NY Open Data and the spatial distribution of each Chain Store Brand in the area of Manhattan.

A. Concentration Impact Index

This indicator characterizes the area around a store in terms of mobility contextual information by using city's transportation stations (collected from City Open Databases). For its calculation, transportation stations that are placed within a radius of 200m around every store were considered to set a geographical area boundary, with the inclusion of the next crucial parameters :

1) the number of lines that operate in a transportation (e.g. metro) station, since more lines operating means more passengers;
2) the distance between the station and the store, since when a store is closer to a station, it is more likely to attract passengers which use this station.

As depicted in Figure 1, the position of subway stations is highly related with the position of chain stores. This fact is captured by the Concentration Impact Index (CII) which is increased in closer stations which have more lines operating :

$$CII = \sum_{i=1}^{\text{TH}} \frac{1}{\text{dist}(cs, th_i)} * |L_i| \quad (1)$$

where:

TH	transportation hubs located inside the radius r
L	number of lines operating in th_i
dist(cs,th_i)	distance between a store cs and a transportation hub th_i

Therefore, the CII for a chain store cs_i should be **inversely proportional** to the *distance from the station* and **proportional** to the *number of lines* operating in this station. The CII for a store cs is calculated using the formula in equation 1.

B. Interest Level Index

This index encapsulates the number of Points of interest (POIs) located in the area around a chain store, since the

number of POIs in an area is indicative for its capacity to attracts people. Here, a POI is identified as *every location (venue) that belongs in the 25% most famous stores/venues of the selected city* to preserve the important POIs which attract a big number of customers or tourists. Therefore, a fine-grained estimation is followed to examine any potential case and not to focus only in specific types of stores :

$$ILI = \left(\sum_{i=1}^{\text{POI}} \frac{1}{\text{dist}(cs, poi_i)} \right) * |POI| \quad (2)$$

where:

POI Places of Interest located inside the radius r

dist(cs,poi_i) distance between a store cs and a POI poi_i

The two crucial factors that determine the value of *ILI*, as given in formula (2), for a chain store cs_i are: the number of POIs existing in the area and the distance between each POI and the chain store.

C. Route Frequency Index

Route Frequency Index (RFI) calculates how central the chain store's position is, by measuring how busy its street can be on a daily basis. The rationale behind this index is based on the fact that chain stores that have the most pedestrians passing in front of them, are more likely to gain more visitors and eventually succeed. Therefore, the Route Frequency Index (RFI) is the fraction of routes passing in front of the store to the total count of routes passing in a radius r around the store to normalize its value.

$$RFI = \frac{\text{number of passing routes}}{\text{total number of routes in } r} \quad (3)$$

Both the number of passing routes in front of the store and the total number of routes in the radius r, are statistically estimated based on the shortest paths approximations as in [6] which has estimated that 80% of the commutes whose distance between starting and destination points is less than 1.5 km. Based on that assumption, the RFI exploits data from applications that recommend shortest path between two points, to find whether a user passes in front of a chain store. The information of whether a user moved from point A to point B is extracted from the consecutive check-ins made on Foursquare. More specifically, consecutive check-ins between two different venues of any type have been used and then such pairwise values were sorted to estimate the shortest paths.

To the best of the authors knowledge, the RFI as a feature has not been used before and all the indicators that have been introduced by earlier studies are of lower granularity levels because the they take into consideration only the area around each store. On the other hand, RFI examines the success of a store based on data of finer granularity, since it examines the location of a store at the **street level**. The estimation of the RFI for a chain store is outlined on the algorithm of Figure 2. RFI estimation algorithm receives as input the Chain Stores CS, the consecutive check-ins CCI (as consecutive check-ins are the ones which occur in less than 2 hours intervals),

Algorithm 1: The overview of Route Frequency Index calculation for each chain store

Data: radius r, Chain Store CS, DirectionSteps, Concecutive Check-Ins CCI

Result: The Route Frequency Index for each researched chain store

1 **for** $cci_i \in CCI$ **do**
2 **if** *distance between the 2 venues in* $cci_i <= 1.5km$ **then**
3 Calculate the steps followed between the 2 venues and add them in $DirectionSteps$
4 **end**
5 **for** *Researched Chain Store* $cs_i \in CS$ **do**
6 $routesInFrontOfChainStore = 0$;
7 $routesInRadius = 0$;
8 **for** *direction_step* $ds_i \in DirectionSteps$ **do**
9 **if** $cs_i\_is\_located\_in\_front\_of\_ds_i$ **then**
10 $routesInFrontOfChainStore++$;
11 **if** $ds_i\_is\_in\_radius\_of\_ch_i$ **then**
12 $routesInRadius++$;
13 **end**
14 $route\_frequency\_index$ =routesInFrontOfChainStore / $routesInRadius$;
15 **end**

Fig. 2. The overview of Route Frequency Index calculation for each chain store

and the radius r. Initally, the algorithm calculates the steps of the consecutive check-ins (lines 1 to 4) and if the distance between the 2 venues is less than 1.5km,the steps followed are calculated using the Google Maps API which returns the shortest walking path. Then, after calculating all the steps, for every chain store cs_i of the specific area, the algorithm estimates the number of routes that pass in front of cs_i (lines 5 to 15) and the number of routes that pass within the given radius. These two variables are used to calculate the index for each chain store (line 14). It should be noted that the parameters' values chosen are indicative and based on the similar earlier work thresholds and the proposed approach is fully parameterized.

IV. The LOCAST Framework

A flexible framework is proposed to enable the proposed location detection. As demonstrated in figure 3, the so called LOCAST framework, operates with the implementation of the next three sequential modules :

A. Data Collection Module

The Data Collection Module harvests data from popular social media and open data sources, namely from Foursquare, Twitter and City Open Data. By using the Foursquare API, a recursive process is implemented to find random spots within the borders of a studied city and to spot stores inside a specific radius. Data for stores locations along with their metadata (like the location, id, category of the store etc) are collected. Since Foursquare API does not provide check-in information

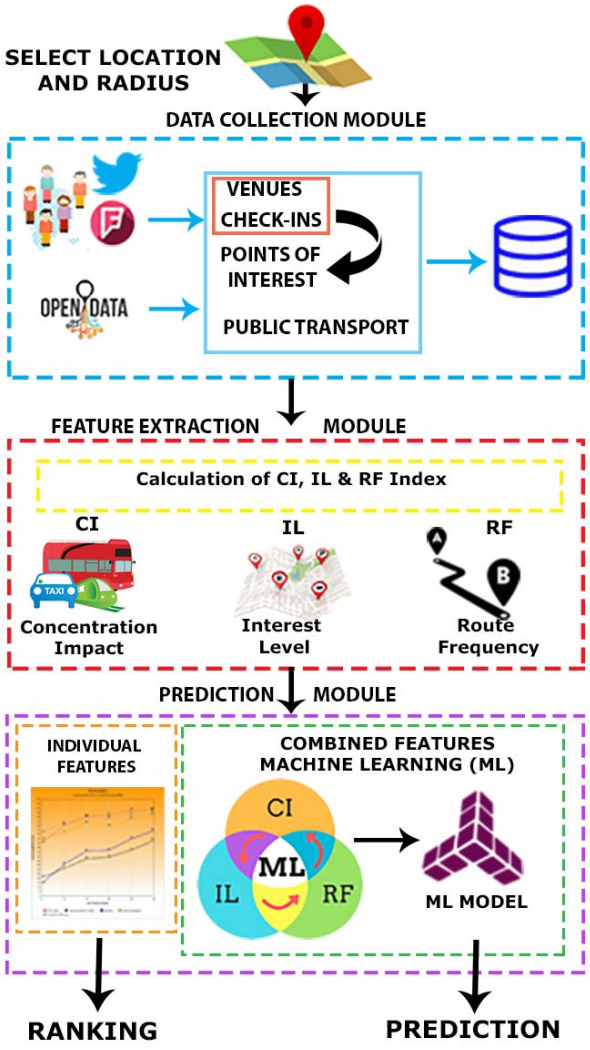

Fig. 3. An overview of the LOCAST framework

of the users, the popular Twitter API [2] is jointly used since Foursquare users tend to publish their check-ins in Twitter. An iterative process is followed to collect Tweets related to Foursquare check-ins and then map each check-in to a store. The required transportation data were provided by City Open Data, i.e. the open datasets provided by the cities for citizens and researchers. As demonstrated in figure 3, the last element of the Data Processing Module are the City Maps, which have been collected by exploiting the open Google Maps datasets.

B. Features Extraction and Prediction Modules

In the Features Extraction Module, the collected data are loaded and the the proposed indicator features are estimated for each store location. This outcome feeds the Prediction Module which measures the accuracy reached under the chosed features. The features validation has been implemented either for *each feature individually* or for *combination of features*. To measure the accuracy of the estimated predictions, the metrics introduced in section V are used. For the validation of combinations of features, several Machine Learning Algorithms were implemented to experiment with various features together and predict the order of the stores. Three mostly relevant regression machine learning algorithms from the related work were used to combine the features and predict their values. In specific, these algorithms are : *Support Vector Regression* [7]; *Linear Regression* [8]; and *M5 Decision Trees* [9]. In both individual and combined features cases, a random sub-sampling method has been used to validate the results. More specifically, the method is repeated 1000 times and in each iteration it selects a subset that includes 33% of the total candidate areas (stores) which are considered as the testing set. The rest of the areas are used to form the training set. The goal is to predict the popularity ranking of the stores, in their potential establishment at each of the considered areas.

[2]https://developer.twitter.com/

V. Experimentation-Results

To measure the accuracy of each index in predicting the best location to open a new store, two metrics were used: **nDCG@k** and **Accuracy@X%**. The main motivation behind their choice is that in LOCAST the position of each store plays a crucial role on whether it will be considered successful or not (section III), thus these two metrics are ideal for this scope as discussed next :

1) nDCG@k metric: The **Normalized Discounted Cumulated Gain** metric measures how relevant a list of ordered stores is to the ground truth list, i.e. nDCG@k value indicates how close the produced list is to the one containing the ground truth values. [10]

$$\sum_{i=1}^{k} \frac{2^{\,\mathrm{rel}(l_i)} - 1}{\log_2(i+1)} \tag{4}$$

$$\text{where } \mathrm{rel}(l_i) = \frac{|L| - \overline{\mathrm{rank}}(l_i) + 1}{|L|} \tag{5}$$

2) Accuracy@X metric: Given two lists, one with items sorted by a relevance factor and another sorted by ground truth factor, the **Accuracy@X%** metric measures the fraction of times that the first item in the relevance list is at the top-X% of the ground truth list, as introduced in [1]. Subsequently, the Accuracy@X% is a binary metric which can be used in cross-validation methods to measure the average occurrence frequency of the highest ranked item in a ground truth list. For example, in a cross-validation method which is repeated 1000 times, the Accuracy@X% should be:

$$Accuracy@X\% = \frac{\sum_{n=1}^{1000} top\_of\_list(RL, GL)}{1000} \tag{6}$$

where:

RL	a list with items sorted by a relevance factor
GL	the ground truth list

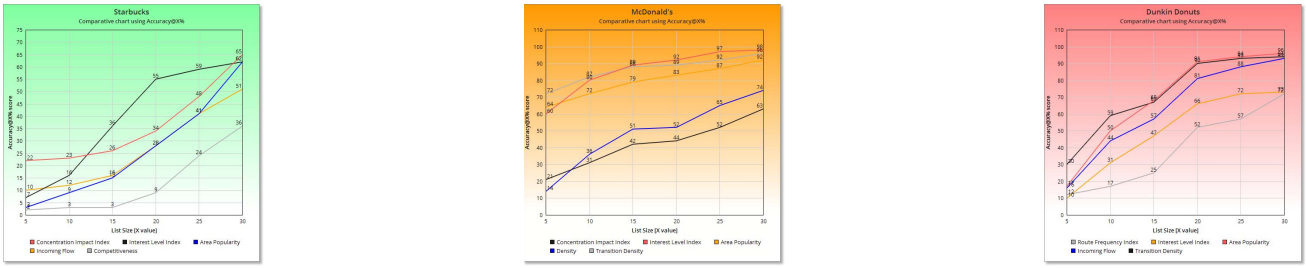

Fig. 4. Comparative results of Individual features for the three introduced features for different values of Accuracy@X% metric.

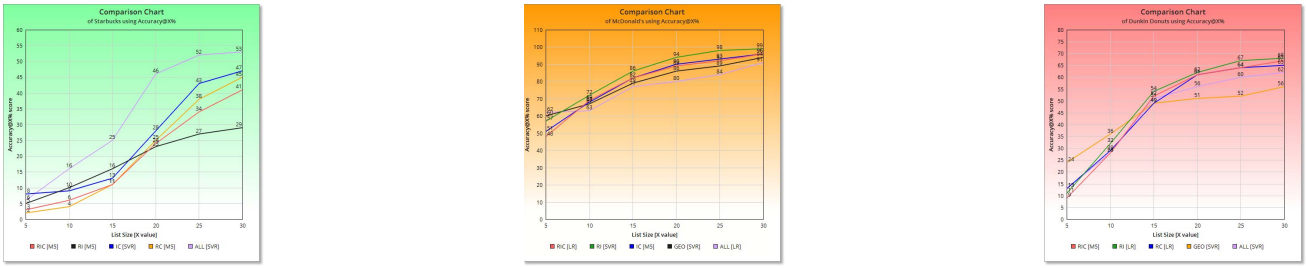

Fig. 5. Comparative results of Combinations of features features for the three introduced features for different values of Accuracy@X% metric.

TABLE I
THE COLLECTED DATA FROM THE CITY OF NEW YORK.

Collection	Size
Venues	55102
Underground Stations	753
Check-ins	176074[3]
Steps	25466
POIs	13772

top_of_list(RL,GL) returns 1 if the highest ranked item in RL, is also in top-X% of GL

A. The New York City Dataset

The city of New York and the geographic area of Manhattan, which attracts many people have been used in the experimentation. Results for three popular American chain stores are presented here to enable comparisons between the stores of each chain. The dataset was created using the methods described in Section IV-A with a set of 203 *Starbucks*, 55 *McDonald's* and 133 *Dunkin Donuts* stores. This dataset has been used for LOCAST experimentation and results are compared with those of the most similar Geospotting work [1].

B. Location Casting Based on Individual Features

As implied by the nDCG@10 metric, the higher the score the closer the relevance list is to the ground truth. In Table II, each chain store is evaluated separately and as indicated in these indexes LOCAST solution overpasses most of the earlier similar work. In specific, the CII in Starbucks' case and ILI in McDonalds' case outperform all the other indices

[3]Within a period from June 2015 to Nov 2015

TABLE II
COMPARATIVE RESULTS OF INDIVIDUAL FEATURES FOR THE THREE INTRODUCED FEATURES USING THE NDCG@10 METRIC. WE HIGHLIGHT WITH BOLD LETTERS THE BEST SCORING INDEX FOR EACH CHAIN STORE BRAND.

Feature	SB	MD	DD
Our research			
Route Frequency Index	0.52	0.7	0.62
Concentration Impact Index	**0.64**	0.85	0.72
Interest Level Index	0.58	**0.93**	0.77
Geospotting			
Density	0.6	0.84	0.75
Neighbors Entropy	0.49	0.71	0.65
Competitiveness	0.48	0.6	0.68
Area Popularity	0.57	0.92	0.81
Transition Density	0.52	0.92	**0.82**
Incoming Flow	0.53	0.9	0.8

when used separately. Furthermore, the charts of figure 4 present the produced results for each individual index using the Accuracy@X% metric. In Starbucks' case both the CII and ILI indexes detect the most successful stores since their scores are better than any other index. ILI is also successful in the McDonald's case (middle chart of figure 4, since for small sets (X=5), it detects 60% of the stores and it constantly rises to 98% for bigger sets (X=30). As a result, ILI index outperforms any other index when it comes to finding the best location for both Starbucks and McDonald's stores while CII is more capable compared to the rest of the indexes in the case of Starbucks.

C. Location Casting based on Combined Features

To combine features and predict check-in values, the three regression machine learning algorithms of *Support Vector Regression*, *Linear Regression* and *M5 Decision Trees* have

TABLE III
COMPARATIVE RESULTS OF COMBINATIONS OF FEATURES FEATURES FOR THE THREE INTRODUCED FEATURES USING THE NDCG@10 METRIC. WE HIGHLIGHT WITH BOLD LETTERS THE BEST SCORING COMBINATION OF INDEXES FOR EACH CHAIN STORE BRAND.

Feature	SB	MD	DD
Our research			
RIC	0.57 [LR]	0.84 [LR]	0.72 [LR]
IC	0.63 [LR]	0.84 [SVR]	**0.96 [LR]**
RC	0.58 [M5]	0.83 [LR]	0.75 [M5]
RI	0.64 [LR]	**0.85 [M5]**	0.73 [M5]
Geospotting			
Geographic Features	**0.66 [LR]**	0.82 [SVR]	0.90 [LR]
All Features	0.61 [LR]	0.80 [SVR]	0.86 [M5]

been implemented with all the possible combinations of the introduced features as next :

- Route Frequency Index, Interest Level Index and Concentration Impact Index which, for brevity reasons, referred as **RIC**
- Route Frequency Index and Concentration Impact Index, referred as **RC**
- Interest Level Index and Concentration Impact Index, referred as **IC**
- Route Frequency Index and Interest Level Index, referred as **RI**

Moreover, two combinations of features introduced by GeoSpotting were used to have a comparison bases and the same as above accuracy metrics were used. The results for each chain store are presented in table III and figure 5 respectively for nDCG and Accuracy@X%.

In McDonald's case the RI combination succeeds to predict the overall success of a new McDonalds store by scoring better results than any other combination. The same applies when it comes to predicting the best location (middle graph of figure 5), where RI dominates over all other indices. In a similar way, in the case of Dunkin Donuts, the best scoring combination are the IC,which achieves a score of 0.96 and outperforms all other combinations and RI combination which produces higher scores than the rest of the combinations indicating that it is more capable in predicting the best location of a new Dunkin Donuts store. In total, the combinations of RI and IC achieve the best performances in the respective cases of McDonald's and Dunkin Donuts.

To summarize the results of the two metrics for both individual and combinations of indices, the *CII* and *ILI* combinations provide the best results for spotting the best location of a new **Starbucks** store using the Accuracy@X% metric. In **McDonalds'** case, *ILI* is the best scoring indicator of success for finding the general success of stores in the area (NDCG metric) as it exceeds all the indexes of the bibliography and - in similar way - the combination of Route and Interest Level Indexes provide the best results using the Accuracy@X% metric scoring an overall score of almost 99% for long lists. The latter eventually means that this combination is the strongest indicator in finding the best area to open a new McDonald's store. Finally, in **Dunkin Donuts'** case, the *IC* combination is more capable than any other index in finding the overall success of Dunkin Donuts store in the metropolitan area of Manhattan(0.96 score in NDCG@10).

VI. CONCLUSIONS-FUTURE WORK

In this paper three new fine-grained indicators of success based on urban characteristics of an area are introduce towards detection the best location for stores. The introduced Concentration Impact Index, Interest Level Index and Route Frequency Index, have been tested and validated under experiments which have implemented thel individually or in combination. The three indexes contribution in the Optimal Location Problem is highly beneficial, since they outperform earlier similar approaches. The proposed indicators can be extended in future implementations as a proof of concept, to predict which area has the most potential when opening of a new chain store is planned. Transportation and Interest Level Indexes are strong indicators in all of the cases under experimentation, either used individually or combined. Route Frequency Index, on the other hand, does not provide satisfying results when used individually but it is proven to be more effective when combined with the other two indices. Future studies can use these indicators combined with other qualitative ones as well, under different cities scales and/or under different chain stores capacities, to further validate the most promising considered contextual features.

REFERENCES

[1] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, "Geo-spotting: Mining online location-based services for optimal retail store placement," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.

[2] F. Wang, L. Chen, and W. Pan, "Where to place your next restaurant?: Optimal restaurant placement via leveraging user-generated reviews," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 2371–2376.

[3] J. Li, B. Guo, Z. Wang, M. Li, and Z. Yu, "Where to place the next outlet? harnessing cross-space urban data for multi-scale chain store recommendation," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16. New York, NY, USA: ACM, 2016, pp. 149–152.

[4] J. Lin, R. Oentaryo, E.-P. Lim, C. Vu, A. Vu, and A. Kwee, "Where is the goldmine?: Finding promising business locations through facebook data analytics," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. New York, NY, USA: ACM, 2016, pp. 93–102.

[5] L. Chen, D. Zhang, G. Pan, X. Ma, D. Yang, K. Kushlev, W. Zhang, and S. Li, "Bike sharing station placement leveraging heterogeneous urban open data," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15. New York, NY, USA: ACM, 2015, pp. 571–575.

[6] S. Zhu and D. Levinson, "Do people use the shortest path? an empirical test of wardrops first principle," *PLOS ONE*, vol. 10, no. 8, pp. 1–18, 08 2015.

[7] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[8] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 2014, vol. 326.

[9] J. R. Quinlan *et al.*, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, vol. 92. Singapore, 1992, pp. 343–348.

[10] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.