Discovery of ROI(s) from Context Rich Geo-spacial data linked to Road Networks

Chahat Bansal
Indian Institute of Technology- Delhi
New Delhi, India
chahat.bansal@iitd.ac.in

B Shanker Jaiswal
Indian Institute of Technology- Delhi
New Delhi, India
B.Shanker.Jaiswal@cse.iitd.ac.in

ABSTRACT

The clusters of related Points of Interest (POIs) are called Regions-of-Interest (ROI). ROI-detection algorithms can either be triggered by user queries or can be passively executed for different monitoring and planning purposes. The two major components of such algorithms are the similarity measure between POIs and the clustering algorithm used to detect relevant ROIs. With the proliferation in the availability of Geo-textual data, the feasibility of ROI-detection techniques has grown multi-fold. Since, the state-of-the-art work lays strong foundation for the future innovations, we write this review paper to study and understand five powerful research studies in this field. We further exhibit a critical comparative analysis of these techniques.

KEYWORDS

Regions of Interest (ROI), Geo-Spatial Clustering, Geo-Social Clusters

ACM Reference Format:
Chahat Bansal and B Shanker Jaiswal. 2018. Discovery of ROI(s) from Context Rich Geo-spacial data linked to Road Networks. In *Proceedings of* . IIT Delhi, COL761, 11 pages.

1. INTRODUCTION

"Region of Interest" (ROI) as the name suggests, is a subset within a dataset singled out for sharp-cut purposes. It encapsulates various significant points of interest (POIs) for the end user. Let us try to understand this concept with a simple analogy.

Suppose a person wants to buy a new house. Some of the parameters which may potentially effect this decision are the proximity of a market, gym, school and park to the house. This parameters are individually called points of interest. The geographical area which can do justice to maximum or all of these POIs is called an ROI.

POI detection and recommendation systems witnessed a noteworthy research scrutiny for several years. With time and advancement of ICT systems, the trajectory of people around target POIs started gaining more attention. This is when ROI came into light.

The headway in communication infrastructure (GPS, 4G/5G, IoT devices etc.), social media platforms (Facebook, Twitter, Instagram, Snapchat etc.) and location-based web services (Google Maps, Zomato, Amazon, Uber etc.) has led to the generation of Geo-tagged information in plenitude. The advance computational and data handling platforms like cloud, HPC etc. have empowered the Data Scientists across the world to mine this data to solve different problems and develop more user-specific applications.

The applications of ROI extend from medical imaging to computer vision to Geographical Information Systems (GIS). Through this review paper, we bring into light some significant research works on ROI detection done in the domain of GIS using road networks.

In last few decades, researchers have published a huge amount of research work related to ROI detection based on the road networks, socio-economic and Geo-tagged information (e.g. As per Google Scholar, more than 2.5K research articles have been published in last one decade). Out of this huge corpus, we have selected five recent and significant research articles for the review.

Let us very precisely introduce you to the shortlisted research articles in reverse chronological order-

(1) **Exploring the Urban Region-of-Interest through the analysis of Online Map Search Queries-** It is a 2018 published piece of work from the KDD conference [1]. It computes ROI Detection and Profiling on the basis of travel flow information extracted from map queries. After dividing the urban area into small region grids, it extracts the travel flows among these grids from query data and thus computes query frequency between corresponding grids. Using PageRank algorithm, visiting popularity of each grid is calculated. The grids are then clustered using grid popularity, applying a density-based algorithm for detecting ROIs. ROI is treated as a document and spatial-temporal preferences are regarded as document-elements.
(2) **Density-Based Place Clustering Using Geo-Social Network Data-** This piece of work is an IEEE publication of May/2018 [2]. The authors present an extension of classic DBSCAN clustering algorithm to incorporate social and temporal information using the check-in data fetched from the social network. The proposed algorithm is called DCPGS (Density-based Clustering Places in Geo-Social network). It is a very rich publication which replaces the epsilon distance in DBSCAN with a weighted combination of spatial and social distance between places. It introduces three significant methods to study the impact of temporal information on cluster generation. In addition to these, visualization and social entropy have been used to determine the quality of social clusters thus generated.

, ,

ACM ISBN . . . $15.00

(3) **SkyGraph- Retrieving Regions of Interest using Skyline Subgraph-** This research work got published in last year's VLDB conference [3]. Instead of detecting a single most optimal ROI, the authors compute one or more ROIs, which are maximal with respect to size and relevance without using any Objective function (preference function for ROI) and constraints on the size and relevance of ROIs. Spatial network is modeled as graph and neighborhood as a connected subgraph.
(4) **GeoScop: A Scalable approach to Geo-social clustering of places-** This work comes from an IEEE conference in the year 2015 [4]. This research work aims at mining Geo-social clusters independent of any specific social network. It collects and combines the check-in data from varied sources like social media, food apps etc. and uses them to find the social similarity between two places based on the community of users who visit them. GeoScop applies Expectation Maximization and DBSCAN in an iterative manner to yield the desired Geo-Social ROIs.
(5) **Retrieving Regions of Interest for User Exploration-** This article got published in the year 2015 in a VLDB conference [5]. This article explored ROI Discovery using graph based technique, on the basis of user preferences expressed as Geo-textual and Geo-social data. It uses objective function to return the most relevant and the smallest ROI covering the user's preference using Greedy approach.

Figure-1 classifies our chosen techniques into active ROI detection and passive ROI detection techniques. While active ROI detection techniques are triggered by user query, the passive techniques are used for ROI exploration for different applications like planning, advertising etc. Our paper presents a critical review of all the above mentioned techniques.

The rest of this paper is organized as follows. Section 2 talks about the motivation behind this work. In section 3 we formalize our problem statement. In section 4, we elaborate on the target ROI detection algorithms. The 5th section holds the comparative analysis followed by the future scope in section 6. We finally conclude our work in section 7.

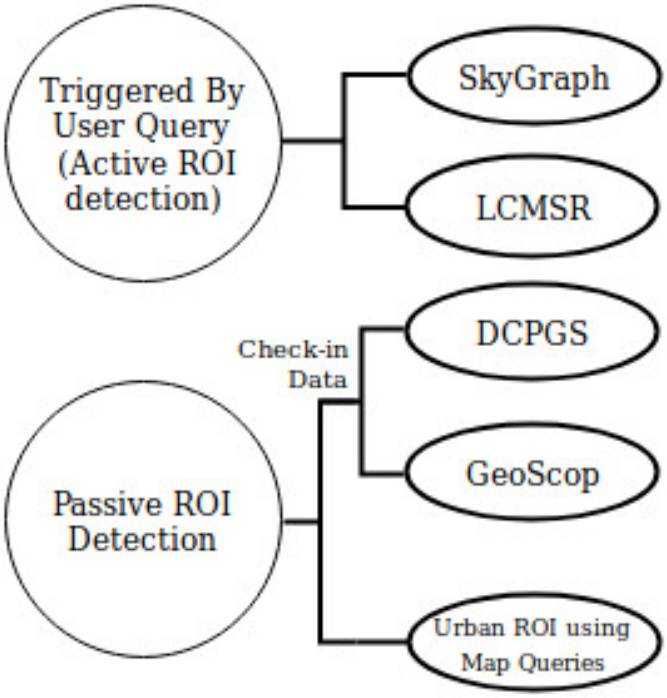
2. MOTIVATION

By virtue of technology advancements, we have access to the Geo-location data of people in great profusion. There have been myriad endeavors towards mining this data for pragmatic applications. One such effort is the detection of Regions-of-interest (ROI) linked to road networks. ROI detection is a significant step towards uplifting marketing strategies, transportation management and development planning.

The initial attempts made for the discovery of ROIs stemmed from the clustering of Geo-spatial dataset. The ROIs emanated from this approach have multitudinous applicability like urban planning, Geo-fencing, socio-economic planning, Intelligent Transportation System (ITS), resource management and disaster management.

With the proliferation in the use of social media networking, Geo-social data came into light. Clustering of this data broadened the horizons of mining higher quality ROIs generating additional semantic information. The extended applications of this approach

Figure 1: Active and Passive ROI detection techniques



are community detection based on ROI, Geo-audiencing, consumer profiling and collaborative campaigns.

These far-reaching benefits of ROI detection call for an understanding of the state-of-the-art techniques, which in turn will lay the foundation for the future advancements in this domain.

3. PROBLEM STATEMENT

The intent of this research paper is to deeply analyze the significant algorithms on ROI detection. We carefully examine the different approaches used by eminent authors and try to classify them on several parameters. We further try to tabulate the differences and correlations among these techniques. Some of the comparison measures are-

- Representation of Road Network (image, graph etc.)
- Distance metrics used (Euclidean, Manhattan etc.)
- Clustering algorithm used
- Scalability of the technique
- Time complexity of the algorithm
- Type of input data fed to the system
- Variable dependencies
- Accuracy of results
- Quality of ROI generated (does it involve semantics or not)

Finally, after a comparative review of tools and techniques used in all the papers, we propose to look into some ideas, with the aim of further optimizing and exploring new areas of applicability. All these will be discussed, hereinafter, in subsequent sections.

4. RELATED WORK

In this section we elaborate on all the target approaches for ROI-detection. For each selected research paper, we talk about its objective, the methodology used, the test dataset used to benchmark its performance, our positive critique about their approach and the limitations identified for that technique. Let us explore each research work one by one.

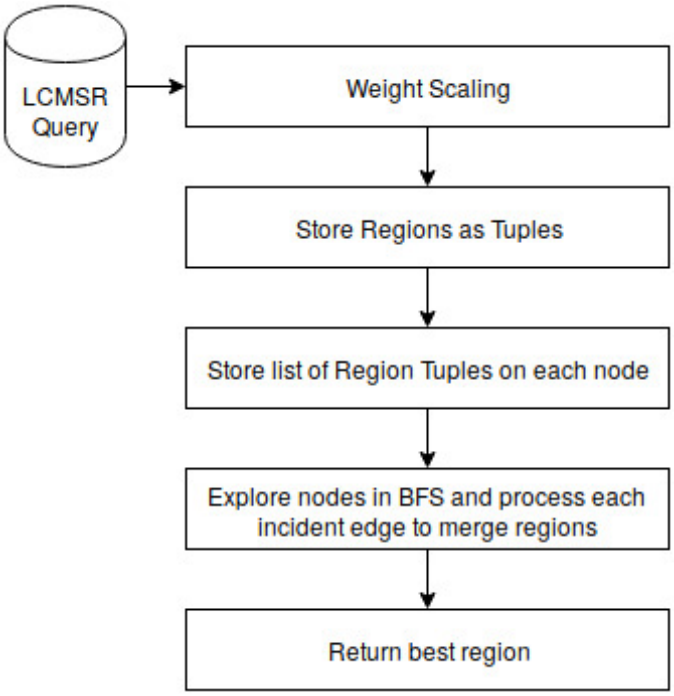
4.1 Retrieving Regions of Interest for User Exploration-

4.1.1 Objective- This piece of research work aims towards finding an ROI which best suits the keywords of the user query. However, in addition to the query keywords, the authors also take the maximum size of the ROI and a general region (on which the search of ROI should be done) as input from the user. This combined user input is called an LCMSR (Length-Constrained Maximum-Sum Region) query. Three ways to answer this query have been presented, namely: Approximation Algorithm (APP), Tuple Generation Algorithm (TGEN) and Greedy Algorithm. The paper further explores the solutions to finding top-k LCMSR query i.e. the top k regions which satisfy the user specified length constraint and lie within the query region.

4.1.2 Methodology- Since the authors claim that TGEN has better accuracy and efficiency over APP, we will consider the improved algorithm i.e. TGEN for our survey. However, even though Greedy algorithm has a greater speed over TGEN, its accuracy is lesser than TGEN. So out of APP, TGEN and Greedy, we find TGEN as the most suitable approach for further discussion.
TGEN is a pseudo-polynomial time algorithm which extends dynamic programming to heuristically find the regions which satisfy the length constraint stated by the user. These regions are called feasible regions. Each node is visited in a breadth-first order and every edge is processed exactly once. Every node is designed to store a list of region tuples to which it belongs, which further helps to create new tuples. When an edge is processed, the regions containing either of the vertices of that edge are combined into a new region connected by this edge. This new region if obeys the length constraint in the query, is used to update the tuple arrays of all the nodes which are contained in it. Finally the region with maximum node weight (called optimal region) is returned as the resultant ROI. This process is depicted by a simple flow diagram in figure-2.

4.1.3 Dataset Used- Two real-life datasets have been used to benchmark the performance of LCMSR. They are shown in Table-1.
4.1.4 Critique- This technique is very flexible in terms of adding different dimensions like popularity, social influence etc. to calculate the weight of the nodes. SkyGraph technique (discussed later) uses the Greedy Algorithm of LCMSR as a touchstone to evaluate its performance. TGEN is suitable for a scenario where the query region specified by the user is small. The LCMSR queries laid the foundation for many enhanced research works in the subsequent years and thus became an important research work to be explored for review.

Figure 2: TGEN for LCMSR queries- algorithm flow



4.1.5 Limitations- LCMSR approach fails the users who lack the knowledge of appropriate ROI size in the query. Therefore, it is less flexible in comparison to the techniques, which find the best possible ROI(s) independent of the size constraint from the user. Despite of being polynomially bounded, TGEN continues to be computationally very expensive. The time complexity of TGEN is $O(|E_Q|T_{max}{}^2)$ where E_Q is the set of edges in the query region and T_{max} is the maximum size of a node's list of region tuples. Since TGEN is a heuristic, it may or may not detect the smallest subgraph which covers all the query keywords. Moreover, the accuracy of this proposed algorithm depends upon the choice of value for α (scaling parameter).

4.2 SkyGraph- Retrieving Regions of Interest using Skyline Subgraph-

4.2.1 Objective- The authors of this paper work towards resolving the trade-off between size (length of minimum spanning tree) and coverage (number of keywords which are part of both the user query and the identified region) of ROIs by putting this decision power in the hands of the end user. SkyGraph is an approximation algorithm for k-skyline subgraph queries which runs in polynomial-time and is presented in this paper to generate smallest size ROIs

Table 1: Dataset used by LCMSR

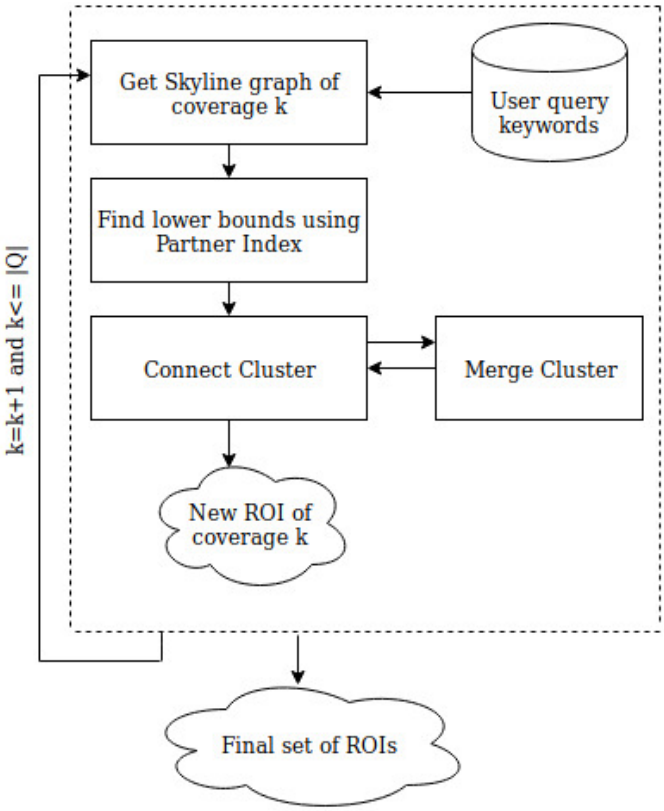
	Dataset 1	Dataset 2
City	New York City	Northwest USA
# Nodes	264,346	1,207,945
# Arcs	733,846	2,840,208
# Objects crawled	500,000	1,000,000
Source of keyword collection	Google Place API	Flickr public API
# Unique Keywords	55,230	107,956

which satisfy the user query at different coverage values. To scale the algorithm further, an index structure called Partner Index is developed which results in a 3-fold speed-up over the basic approach.

4.2.2 Methodology- The ROI detection algorithm is triggered in response to a user query containing different POI keywords called query keywords. In the naive approach, each subgraph of the road network is scored based on a weighted measure of both its size and coverage (number of query keywords contained in that subgraph). A subgraph with the highest score is eventually selected as the target ROI. The subgraph chosen based on an optimal idea of maximum coverage and least size is proven to be a skyline subgraph. Therefore, this approach boils down to finding skyline subgraphs using user query and this naive approach is proven to be NP-Hard and thence Skygraph is proposed.
The baseline algorithm for skygraph tries to find the smallest skyline subgraph which has a coverage of 1,2,3...|Q| query keywords. Each of these skyline subgraph is added in the final set of ROIs and given to the user as the output.
For finding k-skyline subgraph with given coverage k, SkyGraph uses two core algorithms called Connect-cluster and Merge-cluster.
1) Merge-cluster: The aim of this component is to return a cluster with coverage of atleast k/4 query points. This task is achieved by using bottom-up agglomerative clustering. The joining criteria for 2 clusters is chosen to be low single linkage distance, high coverage and low overlap of query points between clusters.
2) Connect-cluster: This function iteratively calls Merge-cluster till the time a region satisfying all k-query points is found. This is done by choosing a root node (preferably the one containing query keyword) and finding its distance from the farthest located query keyword. This distance acts like a radius to the region on which the clustering will be performed by Merge-cluster.

In the non-scalable version of Skygraph Connect-cluster is called once for every node by considering it as a root node. The smallest size subgraph generated would be returned to the user as the final subgraph. However, this approach has a time complexity of $O(n^3)$, which is huge. Therefore, to reduce this time by reducing the number of calls to Connect-cluster, an index structure called Partner Index was presented. Partner Index facilitates the computation of fast lower bounds on the tree size which Connect-cluster returns. This helps us to prune out those calls to Connect-cluster which are certain not to return the k-skyline subgraph. Figure-3 gives the bird-eye view of SkyGraph technique.

Figure 3: SkyGraph- algorithm flow



4.2.3 Dataset Used- The benchmarking of SkyGraph has been established by exploiting the dataset from 3 major metropolitan namely London, Sydney, Dublin and California. Refer Table-2 for the details of this dataset.

4.2.4 Critique- Unlike majority of the ROI detection techniques read so far, Skygraph presents a fresh approach of

Table 2: Dataset used by SkyGraph

City	# nodes	# edges	# keyword nodes	# unique keywords
London	209,407	282,268	87,346	56,648
Sydney	236,041	317,266	23,103	14,063
Dublin	62,975	82,730	1351	170
California	1,752,951	2,157459	172,197	399,238

exploiting k-skyline subgraph queries to determine ROIs based on user query. Moreover, deviating from the density-based clustering algorithms, Skygraph uses agglomerative clustering to merge clusters based on a strong distance measure. The biggest advantage of SkyGraph is that it does not choose any preference function (size or coverage) on behalf of the user. It provides the set of best possible ROIs and the power to choose the optimum lies in the hands of end-user. Even though theoretically the worst-case time complexity of this algorithm is $O(n^3)$, yet practically it turns out to be much less (speed-up of magnitude 1000) when boosted by Partner Index. SkyGraph is proved to be 5 times faster compared to the optimal K-skyline algorithm with similar quality. The authors extensively compare the performance of SkyGraph with LCMSR (explained earlier) which outperforms it in terms of cluster size and user preference dependency. The experiments further prove that SkyGraph is 3-times faster than LCMSR-k and 2-times faster than LCMSR-μ. This technique is best suited for large road networks with densely located huge number of keywords. Another noteworthy advantage of this algorithm is the least number of variable dependencies, unlike all other techniques discussed in this paper.

4.2.5 Limitation- Even though the technique introduced in this research is very unique, it brings in many limitations with it. SkyGraph is less flexible towards adding more than 2 dimensions like social distance, ratings etc. as it will increase the number of skyline neighborhoods generated, which is slow and undesirable. Merge-cluster performs clustering using agglomerative approach which brings in the disadvantages of agglomerative clustering with it. The speed of agglomerative clustering will be directly proportional the size of graph sent to it by the Connect-cluster algorithm, which will be huge if the query keywords are located far apart. The indexing structure consumes a huge memory of O(nW), where W is the number of unique keywords. Section 6.2 of this research paper indicates that SkyGraph in an unsuitable technique for finding ROIs in small sized networks with less number of keywords or having rare keywords in the user query. Under this circumstances, SkyGraph takes more time to execute.

4.3 Exploring the Urban Region-of-Interest through the analysis of Online Map Search Queries-

4.3.1 Objective- This research presents an ROI detection and profiling approach based on mining online map search queries. Post-detecting the intended ROIs, a spatio-temporal latent factor model called Urban ROI Profiling Topic Model (URPTM) is used for ROI profiling to explain the popularity of each ROI.

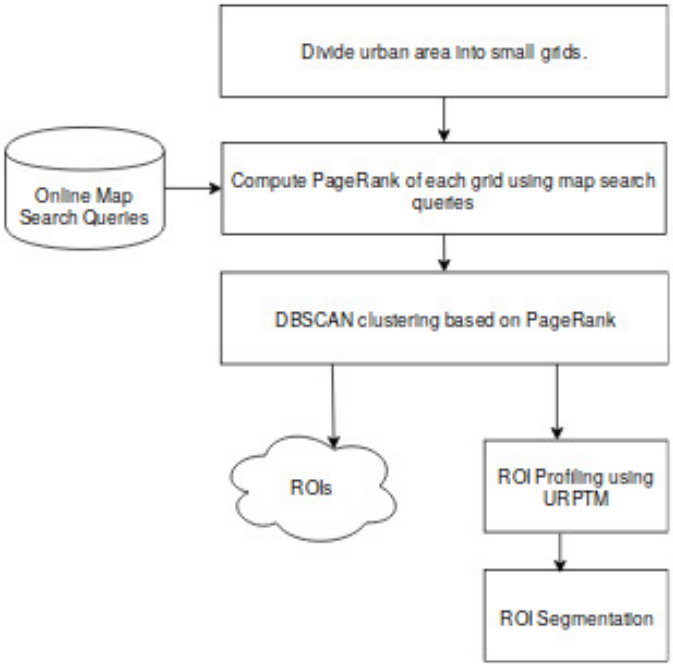
4.3.2 Methodology- This research work begins with the task of ROI detection and then ends up doing ROI profiling to explain the travel patterns of public and reveal interesting features about these ROIs. Let us talk about the ROI detection method used where different nodes are clustered together based on their popularity measure. The travel flows extracted from the map query logs are used to compute the PageRank value of each grid. These PageRank values are further used to compute the Heat Value for each grid which indicates the impact of neighboring grids'popularity on the popularity of the target grid. Once the PageRank and Heat Values are computed, DBSCAN algorithm is used to cluster popular grids together to generate ROIs. Grid with PageRank above a threshold is called a Popular Grid and a grid with Heat value above set threshold is called a Hub Grid. Popular Grids and Hub Grids are together called Active Grids. An active grid with at least some set number of popular grids in its neighborhood are called Core Grids. It is these Core grids which act as a starting point for the DBSCAN clustering algorithm to begin.
An approach similar to classic DBSCAN algorithm is used to find the ROIs. ROI results are checked over different parameter values to finally decide on the desired ROIs.
Once the clusters are detected, an ROI Profiling model called URPTM is initiated to study the travel patterns and POI demands of different people which makes an ROI popular. Temporal data is also taken into account for ROI profiling. Once each ROI profiling is achieved, the ROIs are further segmented based on their profiles for generating final improved clusters. Figure-4 presents a simple flow graph of the overall technique.

4.3.3 Dataset Used- From the city of Beijing, all the search query data from the online maps was extracted and preprocessed for 3 weeks. This dataset can be referred from Table-3.
4.3.4 Critique- This approach is based on solely spatial clustering and does not take into account the semantics of the resultant clusters or temporal information. The major advantage of this technique is exploiting a rich input of online

Table 3: Dataset used by ROI detection algorithm using online map search queries

# Queries	1,500,000
# Destination POIs	120,000
# POI Tags	158

Figure 4: ROI detection using online map search queries- algorithm flow



map search queries. Map queries are a very efficient way of studying people's trajectories and computing popularity of a region. Unlike prominently used check-in data, map queries include both source and destination information which can be further extended to study the communities of people traveling towards popular ROIs. PageRank is a very strong measure to determine the popularity of a region and its influence on the popularity of other regions. DBSCAN was tested with different parameter values to decide on the desired ROIs, which opens up a possibility of OPTICS being a better suited clustering algorithm than DBSCAN in this scenario.

4.3.5 Limitation- Grid-based division of urban area induces major dependency on the size of grid chosen. If the grid size is too small, the complexity of computing PageRank becomes too high, but if it chosen to be large, the clusters within the grid will be missed. Therefore, size of grid should be chosen very carefully. Another drawback of this approach is the shape of ROIs detected which will always be a region bounded by horizontal and vertical boundaries only. These shapes of detected ROIs are non-realistic. Since, this approach uses classic DBSCAN algorithm for clustering, the limitations of DBSCAN get inherited by this approach too. Therefore, this approach is unable to detect sub-clusters and all ROIs generated are of nearly same density. Too many variable dependencies is yet another issue which should be dealt by experts i.e. setting values of Visiting Popularity threshold (PR_th), Heat Value threshold (H_th), Neighborhood size threshold (N_th), Maximum size for ROI (Msize) and grid size.

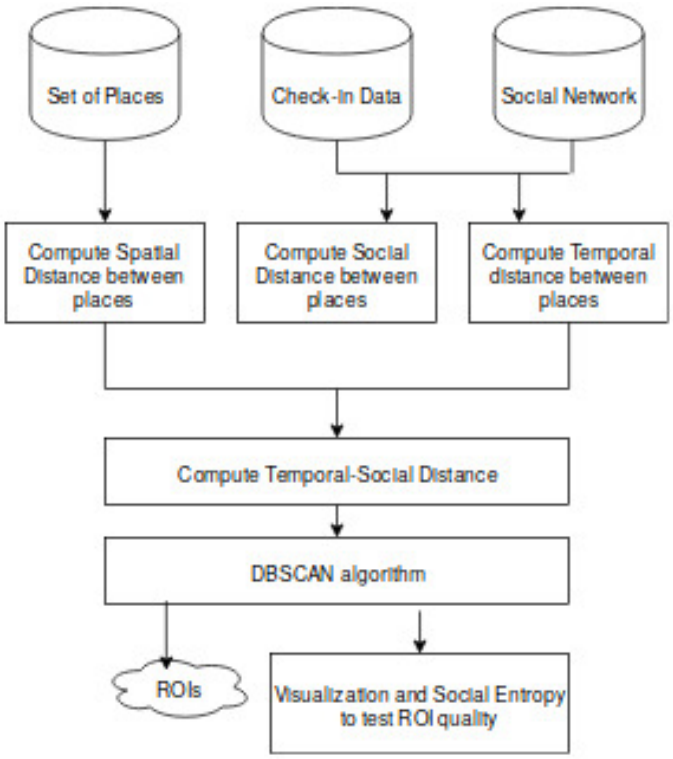
4.4 Density-Based Place Clustering Using Geo-Social Network Data-

4.4.1 Objective- This research work presents a clustering algorithm named Density-based Clustering Places in Geo-Social Networks (DCPGS) which uses spatio-temporal information from check-in data and extracts social relationships from a social network to mine temporal-Geo-social clusters (ROIs) using DBSCAN algorithm. It aggressively compares and provides alternatives for its distance measure and clustering technique using state-of-the-art work. It further presents three interesting ways of incorporating temporal information in the Geo-social clusters which are as follows-
1) History-frame Geo-social clustering- It helps to analyze the evolution of Geo-social clusters by running DCPGS algorithm at different time periods.
2) Damping Window- This approach respects the fact that recent check-ins should be given more weight-age than the historical data.
3) Temporally-contributing Users- It associates the social connectedness of users from their places of visit during a specific time interval.
Once the desired ROIs are generated, the authors test their quality based on Visualization techniques and a new measure called Social Entropy. Social entropy captures the presence of users which are part of the same cluster but belong to different communities. A lower value of social entropy thus indicates a higher quality Geo-social cluster.

4.4.2 Methodology- DCPGS is yet another extension of the classic DBSCAN algorithm which replaces the spatial (Euclidean) distance between places by the Geo-social distance between them. The check-in data and underlying social network is used to compute the social distance between two places. The spatial distance between 2 places continue to be the Euclidean distance between them. The Geo-social distance between two places is a weighted combination of both spatial and social distances. This Geo-social distance can be further strengthened by adding the temporal distance to it calculated in 3 ways i.e. History-frame Geo-social clustering, Damping Window and Temporally contributing users. Once

this amalgamated distance measure between two places is computed, DCPGS moves towards classic DBSCAN algorithm to cluster density-based regions together to produce the desired ROIs. The ROIs generated are further tested for their quality using Visualization and Social Entropy. Figure-5 presents the basic flow of DCPGS algorithm for ROI detection.

Figure 5: DCPGS- algorithm flow



4.4.3 Dataset Used- Gowalla and Brightkite datasets are used to define the efficiency of proposed algorithm. Table-4 summarizes the main properties of both the datasets.

4.4.4 Critique- DCPGS is a unique approach which takes into consideration all spatial, social and temporal aspects for detecting ROIs. Since, the baseline algorithm for clustering is DBSCAN, the running time of this approach is O(nlogn) and can also be brought down to linear time complexity using different extensions of DBSCAN algorithm, which is quite efficient. It compares its algorithm with diverse number of existing techniques like SNN-based clustering, Link Clustering and Metis and successfully proves its efficiency over them using real datasets. It further compares its social distance measure with other prominent techniques like Min-Max version of SimRank, Jacard, Katz and CommuteTime. Their experiments reveal that their chosen social-distance measure has an upper-hand over all these approaches. This paper presents a very strong way of proving the efficacy of their technique using rich comparison-based results which is very convincing and impressive. DCPGS sets the ROI detection bar high by including the temporal aspects in its algorithm. All three techniques i.e. History-frame Geo-social clustering, Damping Window and Temporally contributing users are noteworthy contributions in this domain. All the techniques introduced in this paper help to generate high quality geo-social clusters which can have fuzzy boundaries and are disconnected by geographical barriers like rivers, walls etc. All in all, DCPGS is an avant-garde technique of its time.

4.4.5 Limitation- Just like every rose has thorns, DCPGS has multiple limitations associated with it. A literature named GeoScop (as will be discussed later) elaborates on the major limitations of DCPGS and proves them experimentally. DCPGS collects the check-in data from a single data source along with the underlying social network to compute the social similarity between places. Though, this technique is very simple yet, this dataset can be extremely sparse and user dependent. There are many people who do not post any check-ins on the tested social network and this leads to generation of wrong clusters. In addition to this, the assumption of dealing with single hop friendship edge between people to compute the similarity between them proves to be a weak parameter as it generates incorrect ROIs (as proven by GeoScop). For example, tourist places are visited by different people who do not necessarily have friendship edges between them over the same social network under consideration. Therefore, DCPGS will be unable to identify these clusters. From an algorithm point of view, DCPGS have multiple variable dependencies whose values directly impact the quality of resultant clusters like minimum number of places in an ROI (MinPts), Geo-social distance threshold (ϵ), weighted preference between social and spatial data (ω), social distance constraint (τ) and spatial distance constraint (maxD).

4.5 GeoScop: A scalable approach to Geo-social Clustering of places-

4.5.1 Objective- The basic intention of this research work is to detect ROIs which are not only spatially cohesive but also semantically similar. The user check-in data from multiple sources is recorded and mined to find these Geo-social clusters. Unlike any other state-of-the-art work, the researchers exploit the idea that two places are considered to be close to each other if they are visited by the people of similar community. This paper comes up with a scalable algorithm and extensively compares its performance with DCPGS on real datasets.

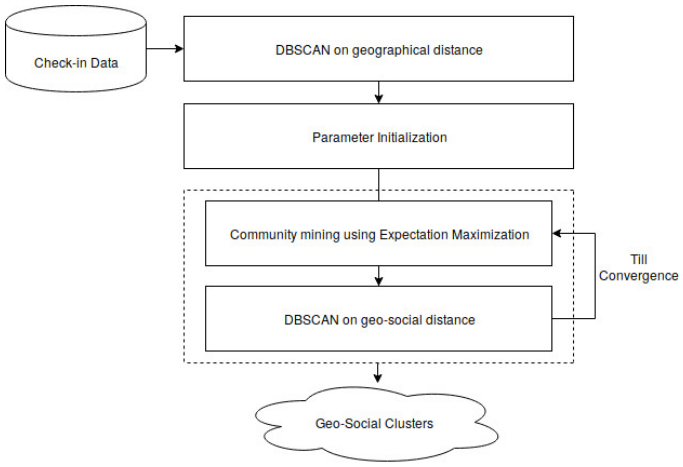
4.5.2 Methodology- The authors lay down a theorem according to which any Geo-social cluster discovered by the GeoScop algorithm will always be a subset of a geography-based cluster. Based on this theorem, an initialization procedure

Table 4: Dataset used by DCPGS

	# users	# check-ins	# places	# edges
Gowalla (Feb/2009 to Oct/2010)	196,591	6,442,892	1,280,969	950,327
Brightkite (Apr/2008 to Oct/2010)	58,228	4,491,143	772,783	214,078

starts by clustering the geographical places based on classic DBSCAN algorithm. Building on this result, each user is initially assigned a community based on the geographical clusters visited by that user. This step is based on an idea that 2 places are considered socially similar to each other if they are visited by the people of similar communities. The Expectation Maximization (EM) model is then initialized and executed till the time its parameter set converges. The temporal information fetched from the check-in dataset is used to study the influence of users on each other. A particular user can influence the visits of other people in its community only within a set time frame from his/her visit. Once the parameter convergence of the EM model is reached, the core intent of clustering Geo-social regions begins by using DBSCAN based on Geo-social distance and EM in an iterative manner till convergence. Figure-6 depicts the basic flow of GeoScop algorithm.

Figure 6: GeoScop- algorithm flow



4.5.3 Dataset Used- This algorithm uses exactly the same dataset as used by DCPGS i.e. open source Gowalla and Brightkite datasets. Refer Table-4.

4.5.4 Critique- GeoScop presents a very strong algorithm for mining Geo-social ROIs from the check-in data of users. The consideration of 2 places to be socially similar if they are visited by the people of similar communities is very practical and produces clusters of higher (6 times better) quality. In addition to this, the algorithm is independent of any social network for computing the social distance between users which gives a major advantage of merging multiple data sources offering user check-in data. It overcomes many limitations of DCPGS which is yet another strong piece of work in finding Geo-social ROIs. It eliminates the dependency on a single social network and the single-hop connection assumption. It outperforms DCPGS in terms of cluster quality and lower number of outliers. We strongly believe that GeoScop is the most effective ROI detection algorithm we have come across till date.

4.5.5 Limitation- GeoScop is vulnerable in terms of variable dependencies and its relatively higher execution time. The GeoScop framework depends upon the density threshold of neighborhood (MinPts), geographical distance threshold (δ_g), social distance threshold (δ_s), minimum similarity constant between users (minSim) and minimum number of similar users in a community (k). All these parameters are very critical and their values highly effects the quality of resultant Geo-social clusters. Due to an iterative implementation of Expectation Maximization and DBSCAN till convergence, GeoScop trade-offs running time with quality. This trade-off is acceptable for applications based on passive ROI detection but not recommended for real-time user query dependent applications. The run time complexity of GeoScop is O(i(nlogn+|D|)) where i is the number of iterations for reaching convergence, n is the number of distinct places and |D| is the size of check-in dataset.

5. COMPARATIVE ANALYSIS

At the outset, we present a tabulated summary of all research papers under consideration in Table-5. This table enumerates the details of each technique on common parameters in reverse chronological order.

From this table we can observe-

- ROI detection techniques started off by exploiting the spatial Geo-textual data and now are progressing towards incorporating social and temporal information too for additional semantics and better results. Out of the explored research works, [2] and [4] are the richest algorithms which exploits all three parameters (spatial, temporal and social) for ROI discovery.
- The ROI-detection techniques which use Euclidean Distance to compute the distance between the POIs, tend to induce error in their computation because the actual road network is not a Euclidean space and different POIs are connected via roads. Therefore, the research papers like [1], [3] and [5] present better distance measures. Out of these three papers, when it comes to active ROI detection technique in response to user query, [3] uses the most appropriate distance measure

Table 5: Summary of ROI Detection Techniques

	Exploring urban ROI through analysis of online map search queries (2018)	Density-based place clustering using geo-social Network data (2018)	Skygraph- retreiving ROI using skyline subgraph queries (2017)	GeoScop: A scalable approach to Geo-social Clustering of places (2015)	Retreiving ROI for user exploration (2014)
Spatial Aspect	✓	✓	✓	✓	✓
Temporal Aspect	✓	✓		✓	
Social Aspect		✓		✓	
Spatial Distance Measure	PageRank and Heat value	Euclidean Distance	Ratio of shortest path in network graph and total coverage of query keywords.	Euclidean Distance	Road Network Distance
Social Distance Measure	N/A	Depends on Visitors of the places and the friendship association between them.	N/A	MinMax Distance	N/A
ROI Detection Algorithm	DBSCAN	Density-based Clustering Places in Geo-Social Networks (DCPGS)	Approximation algorithm for finding k-skyline subgraphs called SkyGraph.	GeoScop- Iterative combination of DBSCAN and Expectation Maximization	Tuple Generation Algorithm (TGEN)
Core clustering approach	DBSCAN	DBSCAN	Agglomerative Clustering	DBSCAN	Length-Constrained Maximum-Sum Region (LCMSR)
Road network representation	Urban region partition on grid basis.	Spatial Network	Spatial Network	Spatial Network	Spatial Network
Time Complexity of algorithm	$O(n^2)$	O(nlogn)	$O(n^3)$	O(i(nlogn+|D|))	$O(\|E_Q\|T_{max}^2)$
Input data to the system	Set of online map search queries	Social network, set of places, Check-in of users	Query keywords from user.	Check-in data	Query keywords, size constraint and rectangular broad region of interest.
Assumptions or dependencies	PR_{th}, H_{th}, N_{th}, Msize & grid size	MinPts, ε, ω, τ, maxD	Radius parameter for Partner Index	MinPts, δ_g, δ_s, minSim, k	α
Quality measure of ROI	Visualization	Visualization, social entropy and community score (based on internal density & conductance)	N/A	Visualization-based and Quantitative Analysis	Annotation Results
Scalability	Scalable because number of grids < number of POI locations	N/A	Structured index called Partner Index.	Grid-based index structure and sliding window partitioning.	Greedy algorithm with accuracy compromised

which includes both shortest distance and total coverage of query keywords.

- We look at two research works [2] and [4] which take into account the social distance as well to find the similarity between the POIs and generate semantically strong ROIs. Out of these two techniques, GeoScop offers a better social distance measure which overcomes the shortcomings of DCPGS and proves the enhancement of the quality of ROIs thus generated.
- Most of the ROI detection techniques studied use DBSCAN to cluster similar POIs into desired ROIs. An interesting diversion from this approach is shown by SkyGraph [3] which uses Agglomerative Clustering.
- Considering the time-complexity, DCPGS outperforms all the algorithms.
- The common input data fed to an ROI detection algorithm are check-in data, online map search queries, user query keywords, additional user constraints (like ROI size etc.) and

Table 6: Advantages and Disadvantages of ROI Detection Techniques

	ADVANTAGES	LIMITATIONS
Exploring urban ROI through analysis of online map search queries (2018)	• Online map search queries is a very profuse dataset to determine PageRank. • ROI profiling provides an informative reason behind the popularity of an ROI.	• Shape of ROI detected is limited to horizontal and vertical boundries. • 5 variable depencies whose value are difficult to set. • ROIs of different densities cannot be determined. (OPTICS should be used instead of DBSCAN)
Density-based place clustering using geo-social Network data (2018)	• Considers spatial, social and temporal aspects to identify clusters. • Barrier/socially separated dense clusters are split. • Allows the clusters to have fuzzy spatial boundries. • Introduces a new & efficient social distance measure.	• Completely dependent on the check-in data of one social network for information. This information can be sparse and highly user dependent. • 5 variable dependencies whose values are difficult to set. • Social similarity based on 1-hop friendship edge between users is a non-realistic assumption.
Skygraph-retreiving ROI using skyline subgraph queries (2017)	• It does not use Euclidean Space to model the road network, which is more practical. • User preference for coverage or size of ROI is not presumed. Therefore more flexible. • Highly suitable for large road networks with many keywords. • Skygraph is 5 times faster than the optimal NP-Hard algorithm.	• Cannot detect overlapping clusters. • Less flexibility in adding more than 2 dimensions like social distance, ratings etc. • Unsuitable for small size networks with less keywords. (High running time) • Agglomerative clustering will be very slow if query keywords are very far apart. • High memory consumption of O(nW).
GeoScop: A scalable approach to Geo-social Clustering of places (2015)	• It supports merging check-in data from multiple data sourses, independent of the social network. • The geo-social clusters are mined independent of any social-network making this approach more flexible and robust.	• Multiple parameter dependencies which are difficult to set. • Communities of users are considered disjoint. Non-realistic assumption as one user can belong to multiple communities.
Retreiving ROI for user exploration (2014)	• TGEN considers the co-location factor of POIs. • TGEN can return abritarily shaped ROIs.	• The TGEN algorithm is slow and computationally expensive. • It does not take into account the coverage of query keywords in the resultant ROI.

data from social network. While check-in data from a single source is highly sparse, using online map search queries and collecting check-in data from diverse sources(and merging them) turn out to be better sources of information to work upon.

- Variable dependencies play a vital role in determining the efficiency of an algorithm. The more difficult it is to set the parameter values, the more overhead gets associated with the technique. We can see that, as the social and temporal considerations get involved, the dependency on different variables increases. This however calls for a technique which can include all these parameters without increasing the dependency overhead. Out of all the techniques studied, SkyGraph turns out to be the most effective in terms of having least variable dependencies.
- Visualization continues to be an important method throughout all the papers to check the quality of the result. Apart from this, social entropy defined by [2] is an interesting measure for determining the social quality of clusters.
- Scalability has not been addressed by DCPGS. All other algorithms have efficient methods for improving their scalability factor. The Greedy approach used by [5] however compromises with the accuracy of the results to gain speed-up, which according to us is not a fair trade off to make.

Keeping these analysis points in mind, we observe that [4] covers the most of the gaps in existing ROI-detection algorithms. It produces highly cohesive ROIs by taking into account all three parameters i.e. spatial, social and temporal. These rich ROIs are generated independent of any social network which is a very fruitful approach. Therefore, we can safely conclude that GeoScop is the best ROI-detection technique out of the considered lot.

6. FUTURE SCOPE

After comparing all the algorithms on some common parameters, we further tabulate the major advantages and disadvantages of all the techniques in Table-6. This will help us understand the strengths of existing technique and the loopholes of the same. The strengths of different techniques can be merged to overcome as many loopholes

as possible. Some challenges for future ROI-detection algorithms to deal with are-

- Minimizing the dependency on variables whose value is difficult to set.
- Reducing the time complexity of algorithms which generate Geo-Social ROIs in response to user queries. The algorithms studied so far involve social and temporal information for passive ROI-detection, which have high latency and are unsuitable for real-time applications.
- Combining heterogeneous datasets to create an exuberant knowledge base upon which mining techniques can be performed. For example, leveraging source location of user from online map search queries to detect their communities and better understand their trajectories on road networks. This feature can be used as an extension to ROI detection based on check-in data.
- Harnessing the potential of Machine Learning and Artificial Intelligence for finding the social similarity between places and generating ROIs which are relevant to the target user.

7. CONCLUSION

This research critically reviews five significant contributions in the field of ROI-detection. Each technique is carefully analyzed and compared on the basis of multiple parameters. Based on these comparisons, we enumerate our observations and conclude that GeoScop is the best technique out of the papers we chose for this review. The last segment examines the advantages and limitations of all these techniques to suggest some future extensions in this domain.

ACKNOWLEDGMENTS

This review work is a product of professor's motivation and institution's support. The authors are grateful to Prof. Sayan Ranu and the Indian Institute of Technology, Delhi.

REFERENCES

[1] Ying Sun, Hengshu Zhu, Fuzhen Zhuang, Jingjing Gu, and Qing He. Exploring the urban region-of-interest through the analysis of online map search queries. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2269–2278. ACM, 2018.

[2] Dingming Wu, Jieming Shi, and Nikos Mamoulis. Density-based place clustering using geo-social network data. *IEEE Transactions on Knowledge and Data Engineering*, 30(5):838–851, 2018.

[3] Shiladitya Pande, Sayan Ranu, and Arnab Bhattacharya. Skygraph: retrieving regions of interest using skyline subgraph queries. *Proceedings of the VLDB Endowment*, 10(11):1382–1393, 2017.

[4] Shivam Srivastava, Shiladitya Pande, and Sayan Ranu. Geo-social clustering of places from check-in data. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 985–990. IEEE, 2015.

[5] Xin Cao, Gao Cong, Christian S Jensen, and Man Lung Yiu. Retrieving regions of interest for user exploration. *Proceedings of the VLDB Endowment*, 7(9):733–744, 2014.