Finding the best ranking model for spatial objects

Hadi Fanaee Tork[1]

Abstract. Top-k spatial preference queries has a wide range of applications in service recommendation and decision support systems. In this work we first introduce three state of the art algorithms and apply them on a real data set which includes geographic coordinates and quality data of over 355 hotels, 276 point of interests and 563 restaurants in Lisbon, Portugal extracted from well-known TripAdvisor[2]. This is the first time that mentioned algorithms are evaluated on a real data set. We also use some optimization tasks for the estimation of algorithms parameters. Finally we rank the hotels using the best obtained ranking model. Result reveals that influence score with a particular radius is able to rank spatial objects very near to the real rankings.

1 INTRODUCTION

There exists an wide range of location-based applications that rely on spatial preference queries. For instance, the tourist species a spatial constraint (for instance the range around a hotel) to retrieve the facilities around the hotel. Then, if the eligible facilities are rated, the result of the query might be the top-k hotels which have the best ranked facilities [3]. Top-k spatial preference query answers such kind of questions. It returns a ranked set of the k best data objects based on the non-spatial score (quality) of feature objects and spatial score (distance) in its spatial neighborhood [1,2]. Several approaches have been proposed for ranking spatial data objects based on defining the score of a spatial data object p based on the scores of feature objects that have p as their nearest neighbor. In the rest of the paper we first introduce a general framework of three algorithms entitled Range Score, Nearest neighbor (NN) and Influence Score. Then in section 3 we present the data set used in the paper. In the section 4 we explain our performed experiments. Later in section 5 we express the results. in section 6 we show how we rank hotels of Lisbon based on the best ranking model obtained and finally in section 7 we discuss the results and bring the conclusion of the paper.

2 TOP-K SPATIAL FRAMEWORK

A Spatial preference query, ranks the spatial objects based on quality of its neighbor facilities. For instance a tourist might retrieve a sorted list of hotels based on the facilities around that (e.g. restaurant, hospital , market, etc.). Assume that p is our point of interest (e.g. a hotel) and we have *m* type of facilities(e.g. restaurant means m=1 and park means m=2). Then assume that f_m^n is *n-th* facility from type *m (e.g.* Restaurant *A).* First we retrieve a list of candidates for *P* according to Table 1. Table 1 shows how one of the methods choose the primary candidates.

Table 1 Candidate Selection Criteria

Method	
Nearest Neighbor	$\min(d(p, f_m^n))$
Range Score	$d(p, f_m^n) < R$
Influence Score	All

As we can see, Nearest Neighbor, from each type *m* retrieves *n-th* element of that (f_m^n) which has the minimum distance with p. Range score retrieves a list of items which have at least distance(*d*) of pre-defined *R* with *P*. Influence score retrieves all the items for further computation. Afterwards, We define Score of point *P* according to the following equation:

$$S_p = \sum_{1}^{m} Agg\{w_{Ci}^m \times \alpha_{Ci}^m\} \quad (1)$$

Where, *Agg* denotes the aggregation function which can be maximum or sum. *w* is equal to the weight or quality of item(e.g. hotel with 5 star can have weight of 5 and hotel with one star can have weight of 1) and *i* is an index of retrieved candidates. α is influence function which is equal to 1 for Nearest Neighbor and Range score and is equal to the equation 2 for Influence score.

$$\alpha = 2^{-\frac{d(p, f_m^i)}{R}} \quad (2)$$

Where *d* denotes the distance between point P and facility *i* of category *m*. and *R* is a pre-defined radius.

Then the result of Top-K spatial preference query is a sorted list of Sp for all point of interests (*P*).

3 DATA SET

Data set is extracted from a well-known online tourism information source *TripAdvisor* which is the most biggest and richest source for travelers around the world to find the relevant information and other user feedbacks about hotels, restaurants and point of interests. One of interesting service of *TripAdvisor* is providing a raking of all tourism locations. The ranking criteria are not visible to the users but in general is a combination of on users opinions and ratings and other sources. Nowadays many users around the world choose their destination, hotels and places to visit based on this ranking.

We extracted all hotels and all near restaurants and point of interests(POI) corresponding to city of Lisbon, Portugal. All GPS

[1] LIAAD-INESC Porto, University of Porto , hadi.fanaee@fe.up.pt

[2] http://www.tripadvisor.com

coordinates and quality factors were extracted from the Raw crawled HTML pages

We then transferred extracted records to the MySQL databases for further process. Finally we had three tables hotels, restaurants and attractions with 355, 563 and 276 records respectively.

Since for some locations , the GPS coordinates were not available, we employed Google Map API[5] and Yahoo Map API[6] Geocoding service to fetch GPS coordinates. Then we removed the places which their coordinate was not available after the Geocoding step. We also removed those hotels which for them ranking was not available in TripAdvisor.

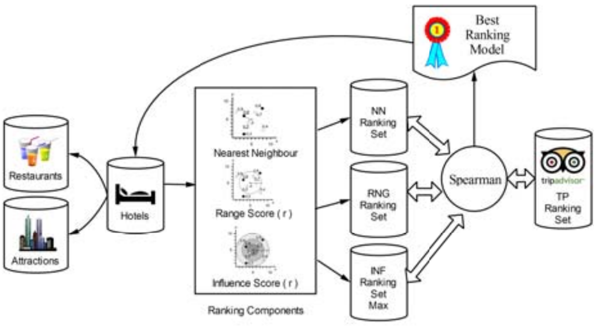


Figure 1. Experiment overview

4 EXPERIMENTS

Two significant problems regarding the Top-k spatial preference query is that first no evaluation on the ranking results is presented yet and second there is not any solution for estimating the radius value in two of algorithms range score and influence score. In other words, when a ranking is made how we can make sure about the correctness of that, or better say how the ranking model correctly assign the spatial objects to the true ranks.

Solving this problem is impossible unless we could compare two generated and real ranking sets together. TripAdvisor real ranking set enable us to perform such comparison and measurement. Our performed experiments are illustrated in figure 1. we first apply Top-K spatial preference query algorithms on the data set and generate three ranking set namely *NN*, *RNG* and *INF* which stands for Nearest Neighbor , Range Score and Influence score respectively. Then in order to evaluate the ranking model we benefit from Spearman's rank correlation coefficient[7]. After this step we find out that which model with which parameters is the best model for predicting the ranking of a hotel. Thus in the next step we employ our best model to rank all the hotels in Lisbon.

As mentioned in the section 2, *Nearest neighbor* is not dependant on radius *R*, so this algorithm doesn't have any input parameters, instead, two other algorithms *Range score* and *Influence score* has radius *R* as their input. In order to study the impact of quality weight on *Influence Score* method, we defined two kind of Influence score, *INFMAX0* and *INFMAX1* so that in the latter one, w is considered to be equal to 1. it means INFMAX1 just consider the spatial property of place and ignores the weight(*w*).

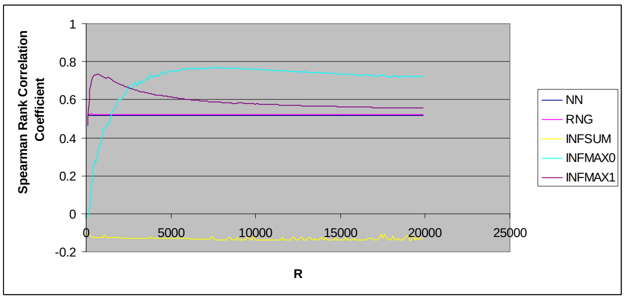
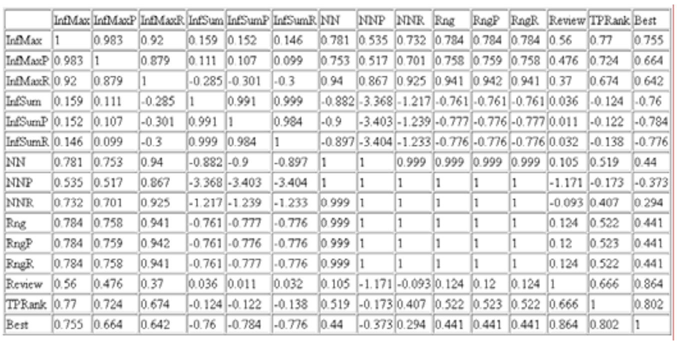


Figure 2. Spearman's rank correlation for different R from 100m to 20000m for 4 rankers NN (nearest neighbor), range score(RNG), influence score with sum module(INFSUM), influence score with max module with considering the rating of attrac-tions(INFMAX0), influence score with max module and not considering the rating of attractions(INFMAX1)

On of the important problem regarding the *Influence Score* approach is determination of R. In order to estimate the best *R* we generated 5 ranking set for R from 100 to 19900 meter by granularity of 100 meter. For both RNG and NN we used maximum aggregation while for INF we tested both maximum(*INFMAX0* and *INFMAX1*) and sum function (*INFSUM*). Then we compute spearman rank correlation coefficient for each 5 generated rankings sets to the TripAdvisor Real ranking set.

Results are shown in figure 2. The vertical axis represents the spearman rank correlation coefficient and the horizontal axis shows the *R* value. The best rankers are those that have the biggest area under their curve. Therefore green curve which is related to the *INFMAX0* would be identified as the best model. INFMAX1 which do not consider the facilities quality is also placed at the second place. The maximum correlation (73.4%) is obtained at R=700m for *INFMAX0* ranker and for INFMAX1 77% correlation is obtained at R=7500m. In terms of *RNG* have a constant behavior between 0.522 and 0.526 very near to *NN* which is always equal to 0.519 and doesn't change by the increasing of *R*.



	InfMax	InfMaxP	InfMaxR	InfSum	InfSumP	InfSumR	NN	NNP	NNR	Rng	RngP	RngR	Review	TPRank	Best
InfMax	1	0.983	0.92	0.159	0.152	0.146	0.781	0.535	0.732	0.784	0.784	0.784	0.56	0.77	0.755
InfMaxP	0.983	1	0.879	0.111	0.107	0.099	0.753	0.517	0.701	0.758	0.759	0.758	0.476	0.724	0.664
InfMaxR	0.92	0.879	1	-0.285	-0.301	-0.3	0.94	0.867	0.925	0.941	0.942	0.941	0.37	0.674	0.642
InfSum	0.159	0.111	-0.285	1	0.991	0.999	-0.882	-3.368	-1.217	-0.761	-0.761	-0.761	0.036	-0.124	-0.76
InfSumP	0.152	0.107	-0.301	0.991	1	0.984	-0.9	-3.403	-1.239	-0.777	-0.776	-0.777	0.011	-0.122	-0.784
InfSumR	0.146	0.099	-0.3	0.999	0.984	1	-0.897	-3.404	-1.233	-0.776	-0.776	-0.776	0.032	-0.138	-0.776
NN	0.781	0.753	0.94	-0.882	-0.9	-0.897	1	1	0.999	0.999	0.999	0.999	0.105	0.519	0.44
NNP	0.535	0.517	0.867	-3.368	-3.403	-3.404	1	1	1	1	1	1	-1.171	-0.173	-0.373
NNR	0.732	0.701	0.925	-1.217	-1.239	-1.233	0.999	1	1	1	1	1	-0.093	0.407	0.294
Rng	0.784	0.758	0.941	-0.761	-0.777	-0.776	0.999	1	1	1	1	1	0.124	0.522	0.441
RngP	0.784	0.759	0.942	-0.761	-0.776	-0.776	0.999	1	1	1	1	1	0.12	0.523	0.441
RngR	0.784	0.758	0.941	-0.761	-0.777	-0.776	0.999	1	1	1	1	1	0.124	0.522	0.441
Review	0.56	0.476	0.37	0.036	0.011	0.032	0.105	-1.171	-0.093	0.124	0.12	0.124	1	0.666	0.864
TPRank	0.77	0.724	0.674	-0.124	-0.122	-0.138	0.519	-0.173	0.407	0.522	0.523	0.522	0.666	1	0.802
Best	0.755	0.664	0.642	-0.76	-0.784	-0.776	0.44	-0.373	0.294	0.441	0.441	0.441	0.864	0.802	1

Figure 3. Spearman's rank correlation for R=7500m and Influence Score with Max module and considering the attractions rating

5 Ranking of Lisbon Hotels

We used our best ranking model (*INFMAX0* with R=7500m) to rank all hotels in Lisbon. Figure 6 shows Spearman's rank correlation computed between all generated rankings sets. P and R at the end of titles stands for attractions and restaurants respectively. For instance *InfMaxR* means that the corresponding generated ranking set is obtained by just taking into account the restaurants and by using Influence score method. *Best* column represents our best ranking model. The columns that doesn't have any R or P at the end of their title are those which both restaurants and attractions are considered in the ranking generation. Also another two columns review and TPrank denote the number of reviews done for that item in TripAdvisor and the corresponding rank in TripAdvisor respectively.

Some interesting facts can be extracted from this table. For instance intersection of *InfMax* and *TPrank* shows that generated ranking set by InfMax has +0.77 correlation to the real ranking provided by TripAdvisor. Also some other interesting results can be obtained from this table. For example we can realize that Influence score with max aggregation if applied on just restaurant data set has +0.94 correlation with ranking set generated with Nearest Neighbor. We also understand that influence score with sum aggregation never performs good and always show a negative correlation to *TPrank*. If we look the correlation between *NN* and *RNG* we discover an interesting fact. It reveals that by using R=7500m ranking set get highly correlated to nearest neighbor ranker with 99.9% confidence.

6 DISCUSSIONS & CONCLUSION

In this paper we presented a new method for evaluation of Top-k spatial preference query. One of the direct result we obtained was the high performance of original influence score ranker with max aggregation function that shows 77% correlation to real ranking of TripAdvisor. It means that when there is no ranking set available, this method can be a good alternative since it generates close ranking set. Second we proved that despite by a first glance, influence score with sum aggregation could have a wide cover on all attractions and thus could have a better ranking result, the opposite happened and it generally didn't provide a good result.

When we are dealing with very large data set, the computation cost will be the most important factor to choose a solution. Nearest neighbor and Range score can be a good choice since provide constant correlation of approximately 50%.

As we also observed there is not considerable difference between *INFMAX0* and *INFMAX1* them. Even in R<700m not considering INFMAX1 that doesn't consider the quality of facilities performs better. It reveal an important fact. Tourist usually use to visits close attractions to their hotel without considering the quality of them. However when distance goes upper than 700m the quality of that attraction gets important and they pay attention to the rating of that place with the goal of not wasting their time and money in transfer. In other words, tolerance threshold of travelers is the intersection of two curves InfMax0 and InfMax1 which is 2700m. It means that by increasing the distance from 700m to 2700m from the hotel, the motivation of travelers to look for rating of the attractions is increased.

The reason why *RNG* and *NN* show a constant value is this fact that most hotel owners establish their hotel in a place that is close to at least some attractions. Except some minor cases, no hotel company invests on a place that is very far from all attractions. So when there is for example 4-5 attractions near to the hotels, their *NN* and *RNG* is affected by the rating of them and thus doesn't change a lot. Because always it is possible to find one high quality attraction near to the hotel.

The reason why influence score with sum aggregation gets negative correlation is this fact that it counts all attractions and thus consider very far attractions and thus distance in equation 2 goes upper and deduct the overall score.

REFERENCES

[1] Man Lung Yiu; Hua Lu; Mamoulis, N.; Vaitis, M.; , "Ranking Spatial Data by Quality Preferences", IEEE Transactions on Knowledge and Data Engineering, vol.23, no.3, pp.433-446, March 2011

[2] J.B. Rocha-Junior , A.Vlachou , C.Doulkeridis , K.Nørvåg , Efficient processing of top-k spatial preference queries , Journal Proceedings of the VLDB Endowment ,Volume 4 Issue 2, November 2010

[3] Hauke J., Kossowski T., Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. Quaestiones Geographicae 30(2), Bogucki Wy-dawnictwo Naukowe,

[4] 4. Barcelona Field Studies Centre S.L. Spearman's Rank Correlation Coefficient http://geographyfieldwork.com/SpearmansRank.htm

[5] https://developers.google.com/maps/

[6] http://developer.yahoo.com/maps/

[7] C. Spearman, The Proof and Measurement of Association between Two Things, *The American Journal of Psychology*, Vol. 15, No. 1 (Jan., 1904), pp. 72-101