CLUSTER-BASED SMOOTHING OF SPARSE RANKING SIGNALS IN MOBILE LOCAL SEARCH

Yuanhua Lv [†] *Dimitrios Lymberopoulos* [†] *Qiang Wu* [‡] *Jie Liu*[†]

[⋆] Microsoft Research
[‡] Microsoft

ABSTRACT

Users increasingly rely on their mobile devices to search for local entities, typically businesses, while on the go. Recent work has recognized unique ranking signals in mobile local search (e.g., distance, customer rating, and number of reviews), and has proposed various ways of leveraging these signals for ranking. However, these techniques have overlooked a major challenge that is amplified in the case of mobile local search: data sparseness. In this work, we exploit domain knowledge about businesses to cluster them based on either the category of the business or the parent chain store that the business belongs to. We then smooth individual business' sparse ranking signals based on the hypothesis that businesses in the same cluster share similar ranking signals. Our experimental evaluation using 14 months of real mobile local search logs, shows that the proposed cluster-based smoothing of these ranking signals can improve mean average precision by 5%.

Index Terms— Mobile local search, ranking signals, sparseness, cluster-based smoothing

1. INTRODUCTION

The wide availability of internet access on mobile devices, such as phones and personal media players, has allowed users to search and access Web information while on the go. The availability of continuous fine-grained location information on these devices has enabled mobile local search, which employs user location as a key factor to search for local entities, to overtake a significant part of the query volume. Sohn et al. [34] found that 38% of mobile information needs are local. This is also evident by recent reports by BIA/Kelsey which show that 30% of all search volume will be local in nature by 2015, as well as by the rising popularity of location-based search applications such as Google Local, Bing Local, and Yelp.

Mobile local search is similar to general Web search in the sense that they both boil down to a similar problem of relevance/click prediction and result ranking. However, there are three fundamental differences and also challenges in developing effective ranking functions for mobile local search.

First, the ranking signals in mobile local search are quite different from general Web search. On the one hand, Web search handles a wide range of Web objects, particularly webpages, while mobile local search focuses mostly on ranking local businesses (e.g., restaurants). Therefore, special domain knowledge about the ranking objects in mobile local search could be exploited to improve ranking accuracy. For instance, businesses may receive ratings and reviews from their customers thanks to the Web 2.0 services. Recent work has already shown that this information can be useful signals for ranking businesses [1, 2, 3].

Second, the available domain knowledge for businesses, such as customer ratings and reviews, can be rather sparse. On one hand, many businesses might not have any customer ratings or reviews, in particular for new businesses. For instance, in a sample of 5 million businesses that were shown or clicked by real users on a commercially available search engine over a period of 14 months, over 4 million of these businesses did not receive any reviews. Given this level of data sparseness, the process of properly extracting domain knowledge and effectively using it for ranking becomes challenging. On the other hand, large chain businesses, such as Wallmart and Chipotle, tend to be geographically distributed across the whole US spanning hundreds or thousands of stores. Our data analysis shows that, even though the actual brand (Wallmart or Chipotle) might have a large number of customer ratings and reviews, the individual chain stores might not. Even worse, the number and quality of ratings and reviews for each individual store might heavily depend on the geosocial characteristics of the store's location.

Third, the quantitative and qualitative information about businesses might vary a lot across business types. Certain categories of businesses might receive more reviews or higher average ratings than others. For instance, our analysis of 14 months mobile local search logs revealed that businesses in the "computer data recovery" category receive an average customer rating score above 9.0/10.0, while the average rating score for "restaurants" is around 7.0/10.0. At the same time, restaurant businesses receive on average 4.16 reviews, while non-restaurant business only receive 1.07 reviews.

To address these problems, we exploit domain knowledge about businesses to cluster them together. In particular, we

cluster businesses based on either category information or the parent chain store that the business belongs to in the case of chain stores. We then propose different techniques for smoothing each individual business' ranking signals using the information of other businesses in the same cluster. The intuition behind this approach is based on the hypothesis that businesses in the same cluster tend to share more similar signal values than businesses in different clusters.

We evaluate our approach using real mobile local search logs from a commercially available search engine over a period of 14 months. By training click prediction models using multiple additive regression trees, we show that leveraging features generated by the different cluster-based smoothing techniques can improve MAP by 5%.

2. RELATED WORK

There have been several large scale studies in the past on mobile query log analysis for deciphering mobile search query patterns [4, 5, 6, 7, 8]. However, few of these efforts have provided insight about the ranking issue.

Recently, the task of improving the ranking accuracy of mobile local search has also begun to attract efforts [9, 10, 1, 2, 3] which have already observed that distance, customer rating score, clickthrough rate, etc. are effective ranking signals in mobile local search. However, these studies have largely ignored the problems of missing and sparse values in these signals. In contrast to existing studies, our work is a first attempt at smoothing the sparse ranking signals in mobile local search.

Predicting the value for missing attributes is an important data preprocessing problem in data mining and machine learning tasks [11]. Our work can also be regarded as predicting missing ranking signals using cluster-based smoothing methods by exploiting the special domain knowledge of mobile local search, based on the clustering hypothesis that closely related documents should have similar signal values.

Clustering algorithms have been extensively studied [12], which is out of the scope of this paper; our work aims at exploiting domain cluster structures for smoothing ranking signals. The cluster-based smoothing strategies, which have been shown to be quite effective in general information retrieval tasks [13, 14, 15], are more related to our work. However, these techniques have not been applied to smooth sparse ranking signals in mobile local search.

3. CLUSTER-BASED SMOOTHING

The ranking signals in mobile local search are often very sparse. For example, it is often the case that a business does not receive any customer review. In the presence of missing customer reviews, a default value of 0 is often used as the rating score. This, however, may be inaccurate, because (1) a business that does not receive any customer review does not necessarily mean that it should be rated low, and (2) it could be unfair to use the same default rating score for all businesses. Even if a business receives reviews, the rating score may still be inaccurate if the reviews are only contributed by a very small number of customers. Besides metadata like rating and reviews, another useful ranking signal, clickthrough rate, which is mined from history query logs [16], also often suffers from a similar problem. To address these problems and improve ranking accuracy in mobile local search, we propose cluster-based approaches to more accurately estimate these signals based on a smoothing strategy.

3.1. Clusters of Business Entities

We explore two types of clusters by exploiting the domain knowledge of business entities: business category and business chain.

The business category data we adopt contains a set of 2710 categories, which is used by a commercial search engine. In such a business category, each business entity has been manually mapped to one or more categories. To give some examples, "Barbecue Restaurants", "Computer Data Recovery", and "Coffee & Tea" are labels of three different categories. Our data analysis shows that businesses from different categories tend to have different numbers of review, rating scores, and clickthrough rates. Inspired by these observations, we hypothesize that businesses from the same category tend to have ranking signals more similar to each other. This motivates us to use the aggregated signal values of businesses in the same category to smooth the signal value of an individual business. Specifically,

- Customer Rating: One way to aggregate rating scores is to average the rating scores of all businesses in the category. However, averaging by businesses can be questionable, because (1) the rating scores of some businesses may be missing, (2) the rating score of a business would intuitively be more trustable if the business receives more reviews. Thus alternatively, we also propose another way that averages the rating score from every review in the category. In doing this, businesses with a missing rating value will be excluded, and businesses that receive more reviews will play a more important role.

- Number of Reviews: We aggregate the numbers of reviews in the same category by simply averaging them.

- Clickthrough Rate: We calculate the clickthrough rate for a category through dividing the total number of clicks received by businesses in the category by their total number of impressions in the history query logs.

We can see that the business category data has a very good coverage: every business has its corresponding categories.

However, on the other hand, business category as the smoothing unit may still be coarse, in the sense that two businesses in the same category could still have significantly different reputation and popularity; this may raise a concern of using the same category aggregation value to smooth both businesses.

In order to relax this concern, we propose to use "business chain" as an alternative smoothing unit at a finer degree. A business chain is composed of a number of connected business entities that share a brand name and provide similar or the same products and services. The coverage of business chains is presumably not as high as that of the business category, because many businesses may not belong to any business chain. For example, our data analysis shows that about 45% businesses do not belong to any business chain. Nevertheless, businesses in the same business chain usually not only belong to the same category but also tend to share similar reputation, popularity, and other properties. We thus believe that the two types of clusters would complement each other.

With business chains, we can smooth the signal of an individual business by aggregating the signal values of other businesses belonging to the same chain in similar ways as we do for category-based smoothing.

3.2. Smoothing Ranking Signals

One standard smoothing approach is the Bayesian prior smoothing, which assumes that the data follows some probability distributions, and uses the Bayesian method to take the reference values as prior to adjust the observed data, e.g., [17, 16]. Bayesian smoothing usualy introduces parameters that need manual tuning to control the prior. However, in our case, we have multiple signals for smoothing, and we also need to optimally combine all the smoothed signals into a single ranking function, making it hard to tune the smoothing parameters.

To overcome this problem, we use a machine learning algorithm to do smoothing and to optimize the combination of multiple signals at the same time in a supervised way. Specifically, we provide not only the customer rating scores, numbers of reviews, and clickthrough rate values, but also the category aggregation values and the business-chain aggregation values (by averaging the corresponding signal values in the category / business chain) as additional features to the learning algorithm. We also build a baseline ranking model based on "global smoothing", which essentially uses the aggregation values of all businesses instead of the aggregation values of a specific category/business-chain in the machine learning process. Our major goal is to use machine learning as a black-box to evaluate the effectiveness of the proposed cluster-based smoothing.

4. EXPERIMENTAL SETTING

In our experiments, the query is sampled from the search log, while the candidate businesses are all those businesses that were shown to the user for that query. We choose to follow the previous work on mobile local search [1, 2, 3] and use clicks to approximate the relevance judgments. Therefore our task is, given a query, to predict if a candidate business would be clicked, and then rank the candidate businesses based on the click prediction. To learn and evaluate a click prediction model, we split the query log into four parts. The first 9 months of data is kept out as the "history" data, and is used purely for estimating the clickthrough rate of a business. We sample 60475, 18491, and 23152 queries from the next 3, 1, and 1 months of data respectively for training, validating, and testing the click prediction models.

Since we need to leverage multiple signals for click prediction, we seek help from machine learning. We adopt MART [18], a learning tool based on Multiple Additive Regression Trees. MART is based on the stochastic gradient boosting approach described in [19, 20] which performs gradient descent optimization in the functional space.

We construct a training instance for each query-business pair, which consists of a set of features (e.g., distance, rating, etc.) and a click label which indicates if the user clicks the business (1 for click and 0 otherwise). The training and validation data are fed into MART to build a binary classification model, which we use to estimate the probability of clicks in the test data.

Our basic feature set contains 4 representative features that are selected based on previous research studies [21, 10, 2, 3]. These features are: (1) the distance between the query and the business locations, (2) the clickthrough rate (CTR) of the business in the history data as defined by the number of clicks divided by the number of impressions and defined as 0 if it did not occur in the history data, (3) the customer rating score of the business in a range of $[0, 10]$, and (4) the number of customer reviews of the business.

We evaluate the retrieval performance in terms of MAP (Mean Average Precision) and the precision at 1, 3, and 5 results (i.e., P@1, P@3, and P@5).

5. EXPERIMENTAL RESULTS

We first compare methods for smoothing rating scores in Table 1. The baseline run involves all the 4 basic features, plus a global smoothing of those sparse signals without using any cluster structures. Throughout this section, "ChainSmth", "CategorySmth", and "HybridSmth" represent for business-chain based smoothing, category-based smoothing, and both business-chain and category-based smoothing, respectively. We use two different aggregation strategies to calculate the smoothing reference, i.e., averaging by businesses and averaging by reviews, which are labeled as "-1" and "-2" respec-

Methods	MAP	P@1	P@3	P@5
Baseline	.409	.249	.183	.151
ChainSmth-1	.416^{b}	.258	.183	.152
ChainSmth-2	.415^{b}	.258	.184	.151
CategorySmth-1	.413^{b}	.253	.184	.151
CategorySmth-2	.412^{b}	.253	.184	.151
HybridSmth-1	.419^{bc}	.262	.186	.152
HybridSmth-2	.418^{bc}	.261	.186	.152
HybridSmth-1&2	.419^{bc}	.261	.186	.152

Table 1. Comparison of methods for smoothing rating scores. The best run is highlighted. b and c indicate the significance over Baseline and other non-hybrid smoothing methods respectively is at the 0.01 level using the Wilcoxon non-directional test.

Methods	MAP	P@1	P@3	P@5
Baseline	.409	.249	.183	.151
ChainSmth	.417^{b}	.259	.184	.151
CategorySmth	.413^{b}	.253	.184	.152
HybridSmth	.419^{bc}	.261	.186	.152

Table 2. Comparison of methods for smoothing the number of reviews. The best run is highlighted. b and c have the same meaning as in Table 1.

.

tively.

Table 1 shows that the proposed cluster-based smoothing of rating scores outperforms the baseline significantly, suggesting the effectiveness of cluster structures for smoothing sparse ranking signals. We can also see that combining ChainSmth and CategorySmth can further improve the performance, which confirms that the two cluster structures are complementary to each other. However, it is interesting to see that averaging by businesses turns out to be more effective than averaging by reviews, though the latter looks more likely to generate the true rating score. One possible explanation is that, users can directly observe the rating score of each business (including the default 0 for missing value) in the mobile local search result list, so averaging by businesses may make more sense to the users, though may not as accurate as averaging by reviews. Unfortunately, it fails to lead any further improvement by applying both averaging by businesses and averaging by reviews. So we will discard the averaging by reviews strategy in the following experiments.

We next compare methods for smoothing the number of reviews in Table 2 and the clichthrough rate in Table 3. There are similar observations as in Table 1: cluster-based smoothing works effectively and two cluster structures are complementary to each other. With the cluster-based smoothing of these two signals, we can also achieve significantly better ranking precision as compared with the baseline run.

In addition, we also examine the sensitivity of the ranking

Methods	MAP	P@1	P@3	P@5
Baseline	.409	.249	.183	.151
ChainSmth	.420^{b}	.264	.186	.152
CategorySmth	.420^{b}	.261	.187	.154
HybridSmth	.426^{bc}	.269	.188	.154

Table 3. Comparison of methods for smoothing clickthrough rate. The best run is highlighted. b and c have the same meaning as in Table 1.

Methods	MAP	P@1	P@3	P@5
Baseline	.409	.249	.183	.151
Rating + Reviews	.422	.264	.187	.153
Rating + Clickrate	.429	.273	.191	.155
Reviews + Clickrate	.428	.271	.191	.155
All	.430	.275	.192	.155

Table 4. Sensitivity Analysis. It shows that combining the proposed cluster-based smoothing methods (i.e., "All") can improve the Baseline over 5% in terms of MAP.

precision to each features in Table 4, where "Rating", "Reviews", and "Clickrate" indicate that we include features used in the highlighted runs in Table 1, 2, and 3, respectively. On the one hand, we can see that leveraging multiple features can clearly improve the performance. On the other hand, the results also suggest that the effect of these signals may have a overlap; as a result, although these features perform well as individual signals, their performance may not simply add together. Nevertheless, the proposed cluster-based smoothing methods boost the MAP by more than 5%.

Finally, throughout all these tables, we notice that the ranking precision increases more for the top positions. For example, P@1 increases by more than 10%, while P@5 only increases less than 3%. This suggests that the cluster-based smoothing appears to be particularly useful for top-ranked positions: this observation is encouraging, because the preccision of top-ranked positions are more important in mobile local search results, due to the relatively small screen of mobile devices.

6. CONCLUSIONS

The problem of data sparseness is amplified in mobile local search and becomes one of the major bottlenecks for effective ranking. To address this problem, we described and evaluated cluster-based smoothing techniques that leverage business domain knowledge, such as business categories or chain store information, to smoothen out a business' sparse ranking signals. Even though the proposed techniques can improve ranking when applied on real search logs, we believe that this work has only slightly touched the data sparseness problem in mobile local search.

7. REFERENCES

[1] Klaus Berberich, Arnd Christian König, Dimitrios Lymberopoulos, and Peixiang Zhao, "Improving local search ranking through external logs," in *SIGIR '11*, 2011, pp. 785–794.

[2] Dimitrios Lymberopoulos, Peixiang Zhao, Arnd Christian König, Klaus Berberich, and Jie Liu, "Location-aware click prediction in mobile local search," in *CIKM '11*, 2011.

[3] Yuanhua Lv, Dimitrios Lymberopoulos, and Qiang Wu, "An exploration of ranking heuristics in mobile local search," in *SIGIR '12*, 2012, pp. 295–304.

[4] Maryam Kamvar and Shumeet Baluja, "A large scale study of wireless search behavior: Google mobile search," in *CHI '06*, 2006, pp. 701–709.

[5] Maryam Kamvar and Shumeet Baluja, "Deciphering trends in mobile search," *Computer*, vol. 40, pp. 58–62, August 2007.

[6] Karen Church, Barry Smyth, Paul Cotter, and Keith Bradley, "Mobile information access: A study of emerging search behavior on the mobile internet," *ACM Trans. Web*, vol. 1, May 2007.

[7] Karen Church and Nuria Oliver, "Understanding mobile web and mobile search use in today's dynamic mobile landscape," in *MobileHCI '11*, 2011, pp. 67–76.

[8] Jaime Teevan, Amy Karlson, Shahriyar Amini, A. J. Bernheim Brush, and John Krumm, "Understanding the importance of location, time, and people in mobile local search behavior," in *MobileHCI '11*, 2011, pp. 77–80.

[9] Alia Amin, Sian Townsend, Jacco Ossenbruggen, and Lynda Hardman, "Fancy a drink in canary wharf?: A user study on location-based mobile search," in *INTERACT '09*, 2009, pp. 736–749.

[10] Nicholas D. Lane, Dimitrios Lymberopoulos, Feng Zhao, and Andrew T. Campbell, "Hapori: context-based local search for mobile phones using community behavioral modeling and similarity," in *Ubicomp '10*, 2010, pp. 109–118.

[11] Roderick J A Little and Donald B Rubin, *Statistical analysis with missing data*, John Wiley & Sons, Inc., New York, NY, USA, 1986.

[12] Rui Xu and Donald C. Wunsch II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[13] Xiaoyong Liu and W. Bruce Croft, "Cluster-based retrieval using language models," in *SIGIR '04*, 2004, pp. 186–193.

[14] Fernando Diaz, "Regularizing ad hoc retrieval scores," in *CIKM '05*, 2005, pp. 672–679.

[15] Qiaozhu Mei, Duo Zhang, and ChengXiang Zhai, "A general optimization framework for smoothing language models on graph structures," in *SIGIR '08*, 2008, pp. 611–618.

[16] Xuerui Wang, Wei Li, Ying Cui, Ruofei Zhang, and Jianchang Mao, "Click-through rate estimation for rare events in online advertising," in *Online Multimedia Advertising: Techniques and Technologies*, pp. 1–12. IGI Global, 2011.

[17] ChengXiang Zhai and John D. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *SIGIR '01*, 2001, pp. 334–342.

[18] Qiang Wu, Christopher J.C. Burges, Krysta Svore, and Jianfeng Gao, "Ranking, boosting, and model adaptation," Tech. Rep. MSR-TR-2008-109, Microsoft Research, 2008.

[19] Jerome H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[20] Jerome H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, pp. 367–378, February 2002.

[21] Jon Froehlich, Mike Y. Chen, Ian E. Smith, and Fred Potter, "Voting with your feet: An investigative study of the relationship between place visit behavior and preference," in *Ubicomp'06*, 2006, pp. 333–350.