An Effective General Framework for Localized Content Optimization

Yoshiyuki Inagaki
Yahoo! Labs
Sunnyvale, CA 94089
inagakiy@yahoo-inc.com

Jiang Bian
Microsoft Research
Mountain View, CA 94043
jibian@microsoft.com

Yi Chang
Yahoo! Labs
Sunnyvale, CA 94089
yichang@yahoo-inc.com

ABSTRACT

Local search services have been gaining interests from Web users who seek the information near certain geographical locations. Particularly, those users usually want to find interesting information about what is happening nearby. In this poster, we introduce the localized content optimization problem to provide Web users with authoritative, attractive and fresh information that are really interesting to people around the certain location. To address this problem, we propose a general learning framework and develop a variety of features. Our evaluations based on the data set from a commercial localized Web service demonstrate that our framework is highly effective at providing contents that are more relevant to users' localized information need.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

Keywords

Localized content optimization, Online local service

1. INTRODUCTION

Localized Web services, such as Yelp (yelp.com) and Yahoo! Local (beta.local.yahoo.com), have become a popular and effective paradigm for users who seek information within the certain geographical regions. Most of such services also deliver various types of information about what's happening near the users' specified geographical location directly via their portals. For example, as shown in Figure 1, the portal of Yahoo! Local presents the user with multi-facets of information, such as news, events, and deals, and for each facet, the portal provides users with corresponding Web content near the users' specified location in a timely fashion.

To provide users with relevant localized contents, it is necessary to perform effective localized content optimization. Localized content optimization in fact aims at finding contents that are of precise neighborhood identification, high quality and authority, strong attractiveness to large popularity and relative freshness. Although human editors can be employed to prune low-quality information and ensure the geographical and temporal constraints, such human effort is quite expensive and cannot guarantee that the selected content items are the most interesting ones to users especially when they must be chosen from a large pool of candidates. To address these challenges, in this poster, we propose a general learning framework for localized content optimization. In particular, we first develop a variety of features based on the pair of the content and the user-specified location; then, we employ the learning-to-rank based approach to build a ranking model that automatically ranks content items according to their relevance to user's interests given the specific location. Our evaluations based on the data collected from a commercial localized Web service demonstrate that our framework is highly effective at optimizing localized contents and improving users' satisfaction with the local experience. In the next section, we will introduce our learning framework in details.

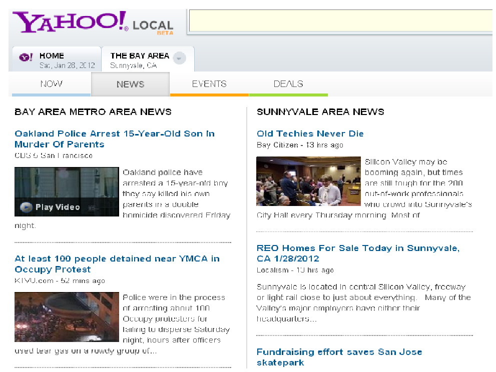

Figure 1: A snapshot of Yahoo! Local page specified by the city of *Sunnyvale, CA*.

2. LOCALIZED CONTENT OPTIMIZATION

In localized Web services, there are a large amount of content items posted by many authorized and ordinary Web users everyday. These items can be of various types, such as news, events, deals, photos, videos, etc. To achieve localized content optimization for each type of contents, we take advantage of the learning-to-rank framework. In particular, given a user's specified location, our goal is to learn a ranking function for pairs of $\langle$*location*, *content item*$\rangle$ and to order the set of candidate content items according to their relevance to the user's interests around this location. The relevance in the localized content optimization consists of multiple aspects, such as precise neighborhood identification of the location, high quality and authority of the content item, high attractiveness and freshness of the content to a large population.

To reach effective ranking, it is crucial to extract useful features. Analogous to feature extraction in general Web search which usually derives *query features*, *document features*, and *query-document correlation features*, we represent each pair of $\langle$*location*, *content item*$\rangle$ as a combination of *location features*, *content item features*, and *location-content correlation features*:

- **Location features** describe characteristics of the user-specified location.
- **Content item features** indicate the content's authority, attrac-

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.

Table 1: **Sample *location*, *content item*, and location-content correlation features and their correlations with four aspects of the relevance.**

	Neighborhood identification	Content quality	Content Attractiveness	Content Freshness
Location	The size of the location's metro area Popularity of the location's metro area ...			
Content item		Content publisher's authority score	Average click-through rate over the content item's semantic categories	Content item's age
		Authority scores of the content item	Click-through rate of the content item's top related semantic category	Novelty score of the content
		...	...	...
Location-content correlation	Distance between the user's location and content publisher's location	Difference between the publisher's authority score and the average authority over items that were clicked from the area of this location	Similarity between the content item and clicked items from the area of this location	Difference between the content item's age and the average ages of items that were clicked from the area of this location
	If user's location is in the same metro area as content publisher's location	Difference between the authority of the item and the average authority over items that were previously clicked from the area of this location	Correlation between the content item's semantic categories and those of top clicked items from the area of this location	Difference between the novelty score of the content item and the average novelty score of items that were clicked from the area of this location.
	...	...	...	...

tiveness, and freshness, and describe characteristics of content publisher.

- **Location-content correlation features** can identify the neighborhood relationship between the user's location and content publisher's location. They can also imply the content item's authority, attractiveness, and freshness by measuring the correlation between the content item and those previously rated by users from this location.

Table 1 presents sample features of each kind as well as their correlation with four aspects of the relevance. Once the features are extracted, we can apply any learning-to-rank algorithm to obtain the ranking function. In this poster, we explore two popular algorithms: RankSVM [3] and Gradient Boosted Decision Tree (GBDT) [1].

3. EXPERIMENTS

• **Dataset:** We gathered user visiting logs from a commercial localized Web service, i.e. Yahoo! Local, during one week in October, 2011. While our framework is applicable to any type of content item, we focused on *news* items in our experiments. To learn the ranking function and conduct offline evaluations based on user click logs, we sampled 742 user visiting sessions in total, each of which consists of one user-specified location and top-20 news ranked by Yahoo! Local's default ranking. As a result, we collected 12735 $\langle$*location*, *content item*$\rangle$ pairs, and 4213 of them triggered users' clicks while others indicate only users' views.

• **Evaluation Metrics:** In our experiments, we use Precision@K ($P(K)$) and Mean Average Precision (MAP) as the evaluation metrics for offline evaluations. For a given session, $P(K)$ reports the average fraction of items ranked in the top-K results that were clicked by the user. MAP is defined as the mean of average precisions of all sessions in the test set. These two metrics are used to measure user's overall satisfaction with the ranking results. For online editorial test, we employ human editors to manually judge the relevance for new ranking results. After that, we use Normalized Discounted Cumulative Gain (NDCG) [2] to evaluate the ranking performance on real localized services.

• **Offline Results:**

To learn the ranking function, we generate the training and testing data as follows: we randomly select 600 sessions from total 742 ones as training set; we use five-fold cross validation to perform training of the ranking function on these 600 sessions; then, we conduct testing on the remaining 142 sessions.

Table 2 reports Precision@1 and MAP values for the hold-out testing set. Besides RankSVM and GBDT, we also evaluate the performance of a simple rule-based method, which orders news based on their ages. The table shows that both RankSVM and GBDT can predict users' localized interests with much higher accuracy than the simple rule-based one.

• **Online Editorial Test:**

We also applied our new ranking function to the real localized Web service and tested its performance. In particular, we implemented our new ranking functions on 6 cities in the U.S., and take 10 snapshots of news rankings under each city during three days with each of the baseline ranking and our new ranking functions. We employed local experts to manually judge the top-5 ranked items for each of ranking methods. The relevance judgment levels are defined as shown in Table 3.

Table 2: **Prec@1 and MAP for various ranking algorithms**

	Prec@1	Prec@3	Prec@5	MAP
Age	0.141	0.118	0.047	0.120
RankSVM	0.503	0.477	0.3294	0.429
GBDT	**0.622**	**0.513**	**0.381**	**0.483**

Table 3: **Relevance judgment levels**

Level	High	Medium	Low
Score	2	1	0

Figure 2 illustrates the NDCG values of various ranking algorithms based on the human editors' relevance judgments. From the figure, we can find that both RankSVM and GBDT can achieve much higher performance in terms of NDCG for online editorial tests.

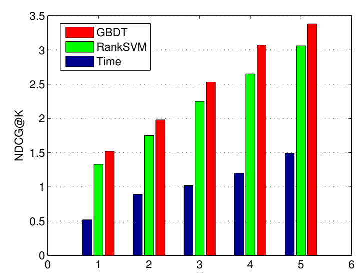

Figure 2: **NDCG for various ranking algorithms in online test.**

4. CONCLUSION AND FUTURE WORK

In this poster, we introduced the problem of localized content optimization and proposed a general learning framework with a variety of features for achieving this task. Our experiments over the dataset from a commercial localized Web service, demonstrated that our framework is highly effective. In future, we will investigate how to build personalized localized content optimization, and explore more features to indicate the localized relevance.

5. REFERENCES

[1] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[2] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20:422–446, 2002.

[3] T. Joachims. Optimizing search engines using clickthrough data. *Proc. of ACM SIGKDD Conference*, 2002.