A Taxonomy of Local Search: Semi-Supervised Query Classification Driven by Information Needs

Jiang Bian
Yahoo! Labs
Sunnyvale, CA 94089
jbian@yahoo-inc.com

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

ABSTRACT

Local search service (e.g. Yelp, Yahoo! Local) has emerged as a popular and effective paradigm for a wide range of information needs for local businesses; it now provides a viable and even more effective alternative to general purpose web search for queries on local businesses. However, due to the diversity of information needs behind local search, it is necessary to use different information retrieval strategies for different query types in local search. In this paper, we explore a taxonomy of local search driven by users' information needs, which categorizes local search queries into three types: *business category*, *chain business*, and *non-chain business*. To decide which search strategy to use for each category in this taxonomy without placing the burden on the web users, it is indispensable to build an automatic local query classifier. However, since local search queries yield few online features and it is expensive to obtain editorial labels, it is insufficient to use only a supervised learning approach. In this paper, we address these problems by developing a semi-supervised approach for mining information needs from a vast amount of unlabeled data from local query logs to boost local query classification. Results of a large scale evaluation over queries from a commercial local search site illustrate that the proposed semi-supervised method allow us to accurately classify a substantially larger proportion of local queries than the supervised learning approach.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Measurements

Keywords

Local query taxonomy, Local search, Semi-supervised learning

1. INTRODUCTION

Local search site is a viable method for seeking information about geographically constrained local businesses, products, and services online. Beyond general purpose web search engines, local search sites, such as Yelp [1] and Yahoo! Local [2], provides a more effective alternative for web users' queries on local businesses. As general purpose web search is driven by users' information needs [4, 8], the central tenet of local search is that web users are driven by *location-targeted* information needs. However, those information needs behind local search queries are usually diverse. For example, some local search queries, such as '*walmart*' and '*cheesecake factory*', intend to retrieve the information of the certain business near the user-specified geographical location, while other local search queries, such as '*Italian restaurant*' and '*gas stations*', infer users' intents to find businesses of the certain category near the user-specified location. Note that, in local search sites, usually the location information is specified by users as one of search settings and not included in the query.

Due to the diversity of information needs behind local search queries, it is inadequate to use a single information retrieval approach to serve all local search queries. For example, to retrieve relevant results for the query '*walmart*', the feature of textual matching between the query and the name of the business might be more essential than others; while, for the query '*Italian restaurant*', the search strategy might count more on the correlation between the query and the topical taxonomy of the business. Therefore, in order to better serve local search, it is necessary to employ different strategies to deal with the various information needs of web users. And, the indispensable prerequisite is to introduce a taxonomy for local search which can categorize local search queries according to diverse information needs.

In this paper, we propose and deeply analyze a taxonomy of local search, which classifies local queries into hierarchical categories according to the hierarchical information needs:

- Business category query
- Business name query
  - Chain business query
  - Non-chain business query

where a business category query, such as '*Italian restaurant*', represents the user's intent to find business of a certain category; a business name query represents the user's intent to retrieve a specific business, which is further categorized into chain business query (e.g. '*walmart*') v.s. non-chain business query (e.g. '*uc berkeley*').

Better understanding of the information needs of queries in terms of the proposed taxonomy can significantly benefit local search, since correct query categorization results in higher relevance with reduced computation for local search service via selecting different search strategies for different query categories. With the goal of automatically and precisely classifying the local search query stream, using solely supervised learning techniques is too limited to be of much practical use since local search queries yield very few online features and it is too expensive to obtain enough editorial la-

[1] http://yelp.com
[2] http://local.yahoo.com

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

bels to train an accurate classifier. In this paper, to address these problems, we develop a semi-supervised approach, which mines large scale local search engine query logs as the source of unlabeled data to aid in automatic classification. In particular, based on analysis over search logs using a small set of labeled queries, we propose a click-based as well as a location-based label propagation method to automatically generate query category labels for unlabeled queries from search logs. Results of a large scale evaluation demonstrate that our semi-supervised method can substantially boost the accuracy of local query classification over the supervised learning approach. Moreover, our method is quite efficient since it only requires very little labeling effort and uses very cheap word unigram features. The specific contributions of this paper include: (1) A general taxonomy for local search driven by local information needs (Section 2). (2) Deep-dive analysis on local search logs. (3) A click-based and a location-based label propagation technique for semi-supervised query classification (Section 3).

2. A TAXONOMY OF LOCAL SEARCH

The information needs behind local search queries are quite different in nature from most of those in general purpose web search, since users of local search sites target finding geographically constrained local businesses rather than traditional web-pages. After observing large scale local search query logs from a commercial local search site, we generalize local search queries into a hierarchical taxonomy according to users' intents:

- **Business category queries:** The purpose of such queries is to search for local businesses of a certain category that the user has in mind. The user will select the specific business after comparing the retrieved ones of the same category. Some examples are '*Italian restaurant*', '*gas station*', '*car dealer*', etc.
- **Business name queries:** The purpose of such queries is to reach the specific business that the user intends to find around a geographically constrained location. According to the scale of such specific business, these queries can be further classified into two sub-categories:
  - **Chain business queries:** The specific business that user intend to reach is of large scale and has many chain stores, such as '*walmart*', '*cheesecake factory*', '*citi*', etc.
  - **Non-chain business queries:** The specific business that user intend to reach is not a chain, such as '*mass general hospital*', '*stanford university*', '*san bruno library*', etc.

3. SEMI-SUPERVISED CLASSIFICATION

Automatic accurate local query classification can benefit local search since it allows us to employ various search strategies to serve different types of information needs. Supervised learning methods [6, 8, 5, 1] have demonstrated their success in building accurate query classifier, however, they usually require high quality features and much human labeling effort, both of which are expensive to obtain. To address these problems, some of previous studies [3, 2] have investigated mining the vast amount of unlabeled data in web query logs to improve automatic web query classification. In this section, we will first conduct a deep analysis on the local search query logs and then develop automatic classifiers by leveraging click and location related information mined from local search logs. It is naturally a semi-supervised learning approach based on label propagation. And, the proposed method can use only cheap features, such as unigram word features, during learning.

3.1 Analysis on Local Search Logs

Local search query logs record the activities of users and reflect the users' actual information needs when conducting local search. They generally have the following information: text queries that users submitted, the URLs displayed to users after they submit the queries, the activities (click or not) of users on URLs, the timestamp of users' activities, and especially the locations that users search for the businesses nearby. Local search logs are separated by sessions, each of which represent one user's single search activity and thus includes one query, one location and all the URLs displayed to the user. Note that the location information can be identified either explicitly by users or implicitly by IP address. The schema and examples of local search logs is shown in Figure 1.

Session	Query	Location	URL	Activity	Time
S_1	Q_1	L_1	D_1^1	view/click	xxxx
S_1	Q_1	L_1	D_2^1	view/click	xxxx
...	...	...	...	...	...
S_2	Q_2	L_2	D_1^2	view/click	xxxx
...	...	...	...	...	...

Figure 1: The schema and example entries of local search logs

Our idea of using local search logs to improve local query classification is to treat these logs as past history of users' information needs, extract more critical information from these logs for better describing diverse information needs, and leverage such a critical information to propagate existing query taxonomy labels to larger set of unlabeled local queries.

In particular, to classify between business category and business name queries, we consider the click information critical: the number of clicks per search session of business name queries is much more likely to be smaller than that of business category queries, because each business name query is intended to reach one specific business that the user has in mind, but users with business category queries tend to compare (click) several search results before selecting her/his favorite business. Moreover, to distinguish between chain and non-chain business queries, the diversity of users' locations becomes vital: chain business queries are submitted by users from more locations since chain businesses have larger geographic scale, but non-chain ones are likely to be bound to fewer locations.

To examine these two intuitions, we analyze a large amount of local search logs from a commercial local search site. In particular, we collect all the search logs from Jan, 2010 to Nov, 2010. Then, we randomly sampled 5074 local queries which occur frequently in the logs and asked human experts to give the editorial taxonomy labels for all the 5074 queries. As a result, there are 1874 queries labeled as business category, 1998 labeled as chain business, and 1202 labeled as non-chain business. Then, we compute the average numbers of clicks per session for all three query categories using different timespans, as shown in Table 1. From the table, we can see that the average clicks per session of business category queries is explicitly larger than that of business name queries, while chain business and non-chain business queries have very close average clicks per session, and that each query category yields very small variance. All these observations indicate that average clicks per session can be used as an effective signal to distinguish between business category queries and business name ones.

Table 1: Average clicks per session for all query categories

Timespan	Category	Aver clicks/session (variance)
1 month (Nov, 2010)	Business category	1.484 (0.554)
	Chain business	1.098 (0.0090)
	Non-chain business	1.105 (0.027)
3 month (Sep-Nov, 2010)	Business category	1.481 (0.581)
	Chain business	1.101 (0.0088)
	Non-chain business	1.103 (0.030)
12 month (Jan-Nov, 2010)	Business category	1.482 (0.574)
	Chain business	1.099 (0.0087)
	Non-chain business	1.102 (0.026)

Additionally, we count the average number of different locations where one query occurs per month for all business name queries, as shown in Table 2. In our work, we use each distinct city to denote one specific location. The table shows that chain and non-chain business queries have very distinct average numbers of locations per month. Although their respective variances are relatively large,

Table 2: Average number of different locations per month for all query categories

Category	Aver num of locations per month (variance)
Chain business	248.16 (2.88×10^3)
Non-chain business	65.36 (4.65×10^2)

we will illustrate in our experiments that this location-based information can be still used as a vital signal to distinguish between chain and non-chain business queries.

In the following, we will introduce how to take advantage of click-based signals to propagate labels of business category and business name to unlabeled queries from search logs, as well as how to leverage location-based signals to automatically assign labels of chain and non-chain business to queries from search logs.

3.2 Click-based Label Propagation

As Table 1 shows, business category queries tend to attract more clicks per session than business name queries; we propose a **click-based label propagation** method to automatically assign labels of business category or business name to unlabeled queries from search logs. In particular,

> *Queries whose average number of result clicks per session is more than a threshold will be labeled as business category; otherwise, they will be labeled as business name.*

The threshold can be decided based on analyzing those queries labeled by human judgments. Specifically, we use this click-based label propagation to automatically generate new class labels for those human labeled queries. We use human labels as truth and select the threshold which can optimize the recall of both query categories based on the new generated labels.

In our experiments, we evaluate the recall of two query categories on those 5074 queries when using the click-based label propagation with different thresholds. The results are shown in Table 3, where the average clicks per session is computed based on three months (Sep-Nov,2010) of local search logs. From the table, we

Table 3: Recall of two query categories (CAT: business category, BIZ: business name) when using the click-based label propagation method with different threshold values.

Threshold	1.1	1.15	**1.2**	1.25	1.3	1.4
CAT recall	0.959	0.929	0.905	0.849	0.797	0.636
BIZ recall	0.646	0.840	0.916	0.956	0.969	0.980
Weighted recall	0.802	0.884	**0.911**	0.902	0.883	0.808

can find that the clicked-based label propagation method can reach nearly optimal performance when we set the threshold as 1.2. We will use this threshold in the following experiments. Note that, with finer selection, we may find more precise thresholds than 1.2, but the following evaluations show that we can achieve good enough accuracy using 1.2 as the threshold.

3.3 Location-based Label Propagation

As shown in Table 2, chain business queries tend to occur at different locations compared to non-chain business queries, thus, we propose a **location-based label propagation** method to automatically assign labels of chain or non-chain business to unlabeled queries from search logs. In particular,

> *Queries which on average occur at more than a certain number (threshold) of locations will be labeled as chain business; otherwise, they will be labeled as non-chain business.*

Similar to Section 3.2, this threshold can be decided by analyzing our set of 5074 human labeled queries. Specifically, we apply this location-based label propagation to generate new class labels for those human labeled queries, and we use human labels as truth and select the threshold which can optimize the recall of both query categories based on new generated labels.

Algorithm 1 Semi-Supervised Local Query Classification

Inputs: Q_E: queries with existing editorial judgments
Q_L: more queries from search logs without editorial judgments
L: a large scale local search log

Outputs: M_{lev_1}: the classifier at 1st level to classify between business category and business name queries.
M_{lev_2}: the classifier at 2nd level to classify between chain and non-chain business queries.

Algorithms:

Step1: Decide the thresholds of click-based and location-based label propagation by using the labeled queries Q_E and search log Q_L.

Step2: Assign the label *business category* or *business name* to each query in Q_L based on click-based label propagation method.

Step3: Train M_{lev_1} based on human labeled Q_E and automatically labeled Q_L.

Step4: For those queries labeled as *business name* in *Step 2*, assign the label *chain* or *non-chain* based on location-based label propagation method.

Step5: Train M_{lev_2} based on human labeled Q_E and automatically labeled Q_L.

We evaluate the recall of two query categories when using the location-based label propagation with different thresholds to generate new labels. As shown in Table 4, the location-based label

Table 4: Recalls of two query categories (BIZ-CH: chain business, BIZ-NC: non-chain business) when using the location-based label propagation with different threshold values.

Threshold	27	28	**29**	30	31	32
BIZ-CH recall	0.720	0.712	0.703	0.694	0.688	0.681
BIZ-NC recall	0.691	0.702	0.718	0.723	0.729	0.734
Weighted recall	0.706	0.707	**0.711**	0.709	0.709	0.708

propagation method can reach nearly optimal performance when setting the threshold to 29. We will use this threshold value in the following evaluations. Also note that we may find more precise thresholds via finer selection, but the following evaluations show that we can achieve good enough accuracy using this threshold.

3.4 Semi-Supervised Classification

Using the above two label propagation methods, we can collect large amounts of automatically labeled local search queries for training more reliable classifiers to categorize queries into the taxonomy introduced in Section 2. In particular, we develop a hierarchical classification system which consists of two classifiers: the first one takes charge of classifying queries between business category and business name, while the second one is used for classification between chain and non-chain business if the result of first classifier is business name. Both of these two classifiers follow the standard classification framework: we derive only cheap unigram word features to represent each query; the label of each query is assigned either by existing human judgments or by automatic label propagation methods. The whole semi-supervised classification framework is generalized in Algorithm 1. For learning method, we will explore three families: SVM, Naive Bayes, and Decision Trees, all using implementations from the Weka [7] framework.

4. EXPERIMENTS

4.1 Datasets

All the queries in our datasets are collected from a commercial local search site. We first sample 5074 queries and ask human experts to give the editorial class labels. Then, to enrich the training data, we collect all local queries in three months (Sep-Nov, 2010), about 405K queries in total, on which we apply our proposed label propagation methods with the thresholds decided by analyzing those 5074 labeled queries in search logs. About 37% of

Table 5: Accuracy of the 1st level semi-supervised classifier (implemented by Naive Bayes(NB), SVM, and Decision Trees(DT)), compared with supervised method trained with small training set.

Classifier	Business category			Business name		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	0.667	0.583	0.622	0.773	0.830	0.800
NB	0.969	0.825	0.891	0.786	0.992	0.877
SVM	0.949	0.940	0.945	0.984	0.986	**0.985**
DT	0.936	0.958	**0.947**	0.988	0.943	0.972

those queries are automatically labeled as *business category*, about 40% as *chain business*, and 23% as *non-chain business*. All the log queries and 5074 human-labeled ones are used as the training set in the experiments. We additionally sample another 3023 local search queries and ask human experts to assign editorial class labels, which are used as the testing set in the evaluations. In total, this testing set contains 508 *business category* queries, 1448 *chain business*, and 1027 *non-chain business* queries.

4.2 Evaluation Metrics

Since our local search query classification system consists of two level classifiers, we can evaluate the performance of each classifier separately. We use standard evaluation metrics in classification, *Precision* and *Recall*. *Precision* of one query class is the ratio of the number of queries correctly assigned to this class divided by the total number of queries assigned to this class. *Recall* of one class is the ratio of the number of queries correctly assigned to this class as compared with the total number of queries truly in this class. We also compute the *F-measure*, which combines two metrics into a single number, the geometric mean of precision and recall, computed as $\frac{2\times \text{precision}\times \text{recall}}{\text{precision}+\text{recall}}$.

4.3 Classification Results and Analysis

In our experiments, we train the two-level classifiers. For the first one, we use 5074 human-labeled queries with 405K automatically labeled queries from three month of search logs as our training set and use another 3023 human-labeled queries as our testing set; for the second classifier, we use all business name queries from those 5074 queries with all logs queries that are automatically labeled as business name as our training set and use all business name queries from 3023 human-labeled queries for testing.

Table 5 reports the accuracy of our semi-supervised first level classifier with different implementations of classification algorithms (Naive Bayes, SVM, and Decision Trees), compared with the supervised baseline which utilizes only 5074 human-labeled queries for training and 3023 queries for testing. The supervised baseline is implemented with Decision Trees which performs best among three algorithms. From the table, we can find that semi-supervised methods can achieve much better accuracy than the supervised baseline, which indicates the effectiveness of our click-based label propagation method for classifying between business category and business name queries. We can also observe that semi-supervised methods implemented by SVM and Decision Trees can result in similar performance. We will use SVM in further analysis since it costs less time for training than Decision Trees.

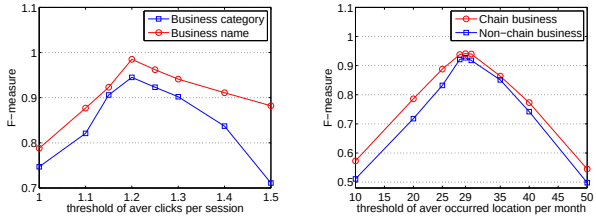
Table 6 demonstrates the accuracy of our second level classifier with different implementations of classification algorithms, compared with the supervised baseline implemented with Decision Trees. From the table, we can find that the semi-supervised methods can achieve much better performance than the supervised baseline, indicating the effectiveness of our location-based label propagation method for classifying between chain and non-chain business queries. Similarly, since SVM reaches similar accuracy compared to Decision Trees but costs much less time, we use SVM in further analysis.

In our semi-supervised learning methods, the threshold of average number of clicks per session for click-based label propagation and that of average number of unique locations associated for location-based label propagation are essential for the classification accuracy. To examine whether optimizing the thresholds,

Table 6: Accuracy of the 2nd level semi-supervised classifier (implemented by Naive Bayes(NB), SVM, and Decision Trees(DT)), compared with supervised method trained with small training set.

Classifier	Chain business			Non-chain business		
	Precision	Recall	F1	Precision	Recall	F1
Baseline	0.702	0.518	0.596	0.636	0.744	0.686
NB	0.923	0.836	0.877	0.802	0.942	0.866
SVM	0.974	0.909	0.941	0.889	0.968	**0.927**
DT	0.969	0.917	**0.942**	0.882	0.958	0.918

which are decided based on analyzing the corresponding characteristics of relatively small set of human-labeled queries in the search logs, can result in better performance, we compare the classification accuracy against varying thresholds during label propagation. Figure 2 demonstrates the classification accuracy in terms of F-measure when we use different values for the two thresholds for label propagation during training, respectively. From these figures, we can find that if we select the values of these two thresholds (i.e. 1.2 for first-level classifier and 29 for second-level classifier) based on analyzing the corresponding characteristics of human labeled queries in search logs, respectively, we can reach the better accuracy of both two classifiers than using different threshold values.



(a) Varying threshold of average clicks per session (b) Varying threshold of average occurred locations per month

Figure 2: Classifiers' accuracy in terms of F-measure against varying values of two thresholds

5. CONCLUSION

In this paper we presented, to our knowledge, the first attempt to define a taxonomy of local search driven by users' diverse information needs. To automatically decide the classes of information need for local search queries, we propose a semi-supervised approach for mining the large amount of local search logs to boost local query classification. Our evaluations using queries from a commercial local search site demonstrate that our proposed methods can substantially outperform the state-of-the-art supervised learning methods. In the future, we will study quantitatively on how such local search taxonomy and corresponding accurate classification can benefit the relevance of local search.

6. REFERENCES

[1] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Proc. of SIGIR*, 2004.

[2] S. Beitzel, E. Jensen, D. Lewis, A. Chowdhury, A. Kolcz, O. Frieder, and D. Grossman. Automatic web query classification using labeled and unlabeled training data. In *Proc. of ICDM*, 2005.

[3] S. Beitzel and D. Lewis. Improving automatic query classification via semi-supervised learning. In *Proc. of ICDM*, 2005.

[4] A. Broder. A taxonomy of web search. In *SIGIR Forum*, 2002.

[5] A. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *Proc. of SIGIR*, 2007.

[6] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *Proc. of CIKM*, 2003.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. In *SIGKDD Explorations*, 2009.

[8] I. Kang and G. Kim. Query type classification for web document retrieval. In *Proc. of SIGIR*, 2003.