Incorporating Seasonal Time Series Analysis with Search Behavior Information in Sales Forecasting

Yuchen Tian,Yiqun Liu,Danqing Xu,Ting Yao,Min Zhang,Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, China, 100084
taylortianyuchen@gmail.com

ABSTRACT

We consider the problem of predicting monthly auto sales in mainland China. First, we design an algorithm using click-through and query reformulation information to cluster related queries and count their frequencies on monthly-basis. By introducing Exponentially Weighted Moving Averages (EWMA) model, we measure the seasonal impact on the sales trend. Two features are combined using linear regression. The experiment shows that our model is effective with high accuracy and outperforms conventional forecasting models. [1]

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

Search Log Mining,Time Series Analysis,Sales Forecasting

1. INTRODUCTION

As more and more people turn to search engine for advice before purchases, search query volume has become an important indicator to predict the sales trend. In a recent work, Choi and Varian[1] use Google Trends data to predict economic indicators like retail sales. However, their approach didn't adequately leverage the rich user behavior information in search log and couldn't model complex seasonal trend. Our contributions in this paper are:*(i)*.An algorithm using both click-through and query reformulation information to identify auto-sales related queries.*(ii)*.An effective prediction model which incorporates seasonal time series analysis with search behavior information.

2. CLUSTERING RELATED QUERIES

As queries related to auto-sales are very popular, simple clustering using only click-through information performs badly, the results of which contains a lot of noise. Here we introduce the user behavior model described in [2] using not only click-through but also query reformulation information. First, we divide the search log into user sessions. A session is a sequence of queries submitted by a user within a given time limit consecutively. Here we set the time limit to 10 minutes, as in [2]. Then, for each query q in the search log we extract the following information:***(1)***.K most clicked documents set $D(q)$:$D(q)$ is a set that contains the top K most clicked documents given q. Empirically, K is set to 15 which can cover most of the clicks.***(2)***.Reformulation query set $RQ(q)$:$RQ(q)$=$\{q_1|\ \frac{\#(q,q_1)}{\#(q)} > \tau,\ \tau = 0.001\}$, where $\#(q)$ is the number of sessions q appears, $\#(q,q_1)$ denotes the number of sessions in which q_1 appears after q. We filter some most popular queries from $RQ(q)$(like *Baidu*) which tends to be formulations of most other queries.***(3)***.Co-occur query set $CQ(q)$:$CQ(q)$=$\{q_2|q_2$ co-occurs with q in the same session$\}$

We put together the above information in a directed graph $G(q)$ as in [2]. $G(q)$ can be seen as a combination of click-through bipartite and query flow graph. Unlike the graph in [2] which includes only reformulation queries, here we add q to $G(q)$ as a query node just like its reformulations. By performing 3 steps random walk on $G(q)$, for each query node q', we compute a *Document Visiting Probability Distribution Vector*, denoted by $\vec{d}_{q'}$. Then, the *relatedness* of two queries is judged using the cosine similarity between their documents visiting probability vectors. Following this, we propose an iterative algorithm to cluster related queries in Algorithm 1.

Algorithm 1: clusterRelatedQueries

```
Input: query set seeds ,RQs,CQs,Ds,iteration times
       t,and similarity threshold θ
Output: relatedQueries
relatedQueries = Φ
Add seeds to relatedQueries
for i = 1 to t do
    newSeeds = Φ
    for seed in seeds do
        Construct G(seed) using RQs,CQs,and Ds
        for query node q in G(seed) do
            if cosineSim(d_seed, d_q) > θ and q not in
            relatedQueries then
                Add q to relatedQueries and newSeeds
            end
        end
    end
    seeds = newSeeds
end
return relatedQueries
```

In the experiment, 5 months of anonymized search query log from a widely-used commercial Chinese search engine

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1]This work was supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A205) of China

is used, from April to August in 2009, and divided into sessions. We select 20 popular queries related to auto-sales and famous auto brand names as seed set. After 5 iterations, we harvest 573 unique auto-sales related queries and only 15 of them are unrelated with θ set to 0.1.

3. DATASETS AND FEATURES

We collect monthly auto sales data reported by China Association of Automobile Manufacturers(CAAM) which could be viewed on their official statistics site[2], from January 2005 to December 2010. Taking data integrity and continuity into consideration, we also use data from the Autohome[3], which is the biggest automobile related site in China.

Our forecast model consists of two features: search query feature and seasonal feature. As to the query feature, we count the monthly occurrences of the related queries collected by Algorithm 1 using query log from April 2009 to December 2010. We rank all the related queries according to their occurrences correlation with the sales data from April 2009 to February 2010. Then, we define $Query_t$, the fraction of the top 100 auto-sales related queries in all search query records of month t as the query feature.

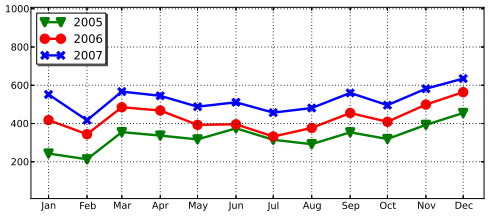

Figure 1: Auto sales trend of 2005,2006,2007

From Figure 1, we can see a strong influence of seasonal impacts on the sales trend. Following this observation, we introduce a variant of EWMA model proposed in [3], which takes seasonal impact and linear trend into account. The seasonal feature, denoted by $Seasonals_t$, is the prediction sales of month t made by the EWMA model. It could be described using Equation (1):

$$Seasonals_t = (\tilde{S_{t-1}}(A) + R_{t-1}(C)) * F_{t-L}(B) \quad (1)$$

Here, L stands for the periodicity of sales trend, and in our case is 12 months. $\tilde{S}_t$ is the seasonal adjusted sales data. F_t is a sales ratio factor represents the seasonal effects estimated from last year. R_t, the linear trend factor, controls $\tilde{S}_t$ to increase or decease. $\tilde{S}_t$,F_t,R_t are smoothed by parameter A,B,C, respectively.

To make predictions, we first use the method described in [3] to setup the initial values of $\tilde{S}_t$,F_t,R_t using 36 months sales data from January 2005 to December 2007. To decide the values of (A, B, C), different from the method described in [3] of using fixed (A, B, C), we adopt a new strategy that makes the model more adjustable: After each prediction, we enumerate value tuples of (A, B, C) from the set $\{(0.1, 0.1, 0.1),(0.2, 0.1, 0.1),\cdots,(1.0, 1.0, 1.0)\}$ to choose the one whose predictions have the highest correlation with the latest 24 months of real sales since January 2008 and use it in the next prediction.

Combining both features in a linear model, our model is established as follows:

$$log(Auto_t) = a * log(Query_t) + b * log(Seasonals_t) + c \quad (2)$$

[2]http://www.auto-stats.org.cn/

[3]http://club.autohome.com.cn/bbs/thread-a-100002-8753794-1.html

4. EXPERIMENTS AND ANALYSIS

We compare performance of 4 models: (1).query feature only.(2).seasonal feature without EWMA model: we use Seasonal Autoregressive(AR) model in [1], a tradition seasonal impacts measuring model, as a substitute for EWMA model which combines $log(Sales_{t-1})$ and $log(Sales_{t-12})$ in a linear model.(3).seasonal feature with EWMA model.(4).the proposed model (seasonal feature with EWMA combined with query feature). We use the data from April 2009 to February 2010 as training set and perform linear regression on those models. The sales data from March 2010 to December in 2010 is used as test set. Before experiment, all data are normalized by dividing their maximum values. The results are showed in Table 1. From Table 1, we can see that Model 3 has a much higher correlation than Model 2, improves the correlation by 59% from 0.537 to 0.855. This clearly states the importance of measuring seasonal effects in prediction of the sales trend, like auto sales. While seasonal feature using EWMA alone gains good performance in prediction, after combining with query feature in Model 4, the correlation boosts to 0.895. This can be explained by the fact that the proposed model integrates the feature which represents the *present*, the Query feature, instead of just using data from past, thus has the highest correlation of all models.

Table 1: Results of Different Models on Test Set

Model	Features	Correlation
1	Query feature only	0.773
2	Seasonal feature without EWMA	0.537
3	Seasonal feature with EWMA	0.855
4	Proposed model	**0.895**

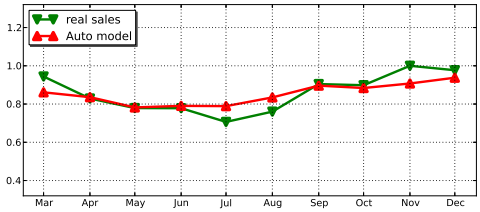

Figure 2: Curves of $Auto_t$ and sales data

5. CONCLUSIONS

In this paper, we propose an algorithm to identify and cluster related queries using rich user behavior information in search log, and a forecast model combining present query feature and past seasonal feature. Experiments comparing prediction performance of several forecast models confirm the effectiveness of our model.

6. REFERENCES

[1] H. Choi and H. Varian. Predicting the present with google trends. Technical report, Google Inc., 2009.

[2] E. Sadikov, J. Madhavan, L. Wang, and A. Halevy. Clustering query refinements by user intent. In *Proceedings of the 19th World Wide Web conference*, 2010.

[3] Peter R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, Vol.6(No.3):pp. 324–342, April 1960.