DREXEL UNIVERSITY

MIS 698 Final Project

Technical Approach Document

Group 8

Malhar Deshpande
Gavin McCullion

8/10/2015

Malhar Deshpande
Gavin McCullion

Technical Approach Proposal
MIS 698, Summer 2015

1. INTRODUCTION

We will be pursuing project B, the Expedia Search Ranking problem. Rankings and user recommendations are used in a variety of online services, including search engines and entertainment services such as Netflix. The large dataset will necessitate using Hadoop for working with the data, as well as working with machine learning algorithms to determine a methodology for ranking hotel searches.

1.1 PROBLEM STATEMENT

Online travel agencies (OTAs) find themselves in a very competitive marketplace, with millions of potential customers searching for the best deals on a variety of websites. Websites need to deliver the best search results in order to win sales. The difficulty comes in matching users to available hotel rooms, without having the user find a better deal on a different website. A ranking system tailored to specific consumers with price competitiveness helps ensure that an OTA will get the sale.[1]

The daunting amount of user data and available algorithms to create rankings means that even for the best in the business, there is likely room for improvement and refinement, especially as new data becomes available and consumer preferences shift. We seek to explore search data for Expedia.com, and explore a number of descriptive statistics and machine learning algorithms to understand which attributes contribute to consumer purchasing and propose an algorithm of our own for search ranking.

1.2 BACKGROUND

There are more than 10,000 travel agencies in the United States, and the industry has been growing at a rate of 3.8 % since 2010. Online travel agencies have revolutionized the way travelers book their flights and accommodations, allowing consumers to research competitor pricing and book their own reservations without the assistance of a travel agent. Expedia owns the highest market share in the industry, at 9.5%, with competitor Preceline.com coming in second at 5.4% of the market. Profit margins are thin for the industry, at only 2.6% of revenue, so companies need to maximize the amount of customers purchasing through their service in order to keep profits high.[2]

[1] https://www.kaggle.com/c/expedia-personalized-sort

[2] IBISWorld iExper Industry Summary: Travel Agencies in the US. http://clients1.ibisworld.com.ezproxy2.library.drexel.edu/reports/us/iexpert/default.aspx?entid=1481

2. PROPOSED TECHNICAL APPROACH

2.1 MAP REDUCE

2.1.1 What will your mapper accomplish? (What intermediate key/value and what output key/value pairs?)

We will be examining the relationship between hotel star ratings (prop_starrating) and location desirability scores (prop_location_score1 and prop_location_score2). Our mapper will use the star rating as our keys, and the location scores as the values.

2.1.2 What will your reducer accomplish? (What intermediate key/value and what output key/value pairs?)

Our reducer will take the average of the location scores, aggregated against each star rating.

2.1.3 Are there any intermediate steps between the mapper and the reducer that you will implement?

During the shuffle and sort operation, star ratings will be aggregated on each node, and the value list will become a list of location scores.

2.2 DATA

The Expedia dataset has over sixteen million data points, compiled from user behavior. Users can search for hotels by visiting Expedia.com or through search engines such as google. Once users submit a search, a ranked list of hotels is returned. Sometimes Expedia knows the customer's purchase history, and sometimes they know competitor's pricing on the same hotels.

The dataset contains five categories of variables:

- Search Criteria
- Hotel Characteristics (static)
- Hotel Characteristics (dynamic)
- Visitor Information
- Competitor Information

The size of the data set can present a challenge, and may be too large for certain algorithms to run on. In addition, there are a number of columns present with varying amounts of missing data.

2.3 METHODOLOGY

2.3.1 How do you plan to accomplish the ranking?

We will be running our analysis of the ranking using R. First, we will narrow down the number of features for the dataset using a logistic regression to determine important features. Once we have our features selected, we will run a number of different models to determine which gives us the best normalized discounted cumulative gain. Candidates for model testing include:

- Logistic regression
- Neural Networks
- Support Vector Machines
- Random Forest
- LambdaMART

2.3.2 How will you calculate the relevant metrics?

We will be performing our calculations with R.

2.3.3 What statistics are you considering?

We will be looking at the NDCG, as well as ROC curves to compare models.

2.3.4 What type of visualization are you considering?

We will overlay the ROC curves of our various models to compare their performance.