About the Accommodation Search Ads (ASA) Data Sets

Guanting Tang Yupin Yang Jian Pei

This document describes the accommodation search ads (ASA) data sets used in [3]. In total, there are 3 data sets, namely the query data set, the search advertisement data set and the advertiser data set.

The data sets are stored in *.txt* files. Attributes are separated by *tab* delimiters. Attribute names are presented at the first row of each file. All data was collected in April 2011.

1 Query Data Set

The overall statistics of the query data set are shown in Table 1.

Table 1: Statistics of the Query Data Set

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Real, Integer, Text
Number of Instances:	1,542
Number of Attributes:	4

1.1 Data Set Information

The queries and their characteristics, including global monthly searches and approximate cost per click, were collected from the accommodation category in the Travel & Tourism section of Google AdWords [2]. We included 1,542 queries that are related to the accommodation industry here.

1.2 Attribute Information

There are 4 attributes in total, which are

Query: the detailed query. Format: *text*.

Length: the number of words in the query. Format: *integer*.

CPC: Cost per Click, the cost of a query that a customer might have to pay if bidding on it. It is calculated by the average cost of ads on all positions of that query [1]. Format: *real number*.

Global-Monthly-Search: the average number of searches for each query on Google over the past 12 months in all locations, languages, devices, and query match types [1]. Format: *integer*.

2 Search Advertisement Data Set

The overall statistics of the search advertisement data set are shown in Table 2.

Table 2: Statistics of the Search Advertisement Data Set

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Real, Categorical, Integer, Text
Number of Instances:	11,818
Number of Attributes:	7

2.1 Data Set Information

The search advertisement data set was crawled by submitting the queries from the query data set to the Google search engine. Each query was submitted multiple times. The replicated advertisements were deleted from the data set. In total, we included 11,818 advertisements that are related to accommodation. We regard two advertisements identical if they satisfy three conditions: (1) their entire content, including the title and description, is the same; (2) they are triggered by the same query; and (3) they come from the same advertiser. We also did some data cleaning on the advertisements included in this data set.

2.2 Attribute Information

There are 7 attributes in total, which are.

Query: the detailed query. Format: *text*.

Advertiser: the advertiser, which is presented by the domain part of the URLs appearing in the advertisements. Format: *text*.

Title: the title of an advertisement. Format: *text*.

Description: the description of an advertisement. Format: *text*.

Ad Length: the number of words in an advertisement, including the title and description. Format: *integer*.

Ad Position: the position of an advertisement. Format: *categorical*. The value "R" refers to the right-hand side position; and the value "U" refers to the underneath the search bar position.

Average Rank on Right-hand Side: the average rank that an advertisement appeared on the right-hand side. Format: *real number*.

3 Advertiser Data Set

The overall statistics of the advertiser data set are shown in Table 3.

Table 3: Statistics of the Advertiser Data Set

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Binary, Real, Integer, Text
Number of Instances:	963
Number of Attributes:	14

3.1 Data Set Information

Advertisers here are represented by the domain part of the URLs appearing in the advertisements. For example, we treated `www.expedia.ca` and `www.expedia.com` the same advertiser *expedia* instead of two. The advertisers that are not in the accommodation business (e.g., `ask.com`) were excluded

from the data set. Most of the advertiser characteristics were obtained manually based on the advertisers' website information (especially on the "About Us" pages).

3.2 Attribute Information

There are 14 attributes in total, which are

Advertiser: the advertiser, which is presented by the domain part of the URLs appearing in the advertisements. Format: *text*.

Service Provider: an indicator on if the advertiser is a service provider or not. Format: *binary*, "1" represents yes; and "0" represents no.

Travel Agency: and indicator on if the advertiser is a travel agency or not. Format: *binary*, "1" represents yes; and "0" represents no.

Price Search Website: an indicator on if the advertiser is a price search website or not. Format: *binary*, "1" represents yes; and "0" represents no.

Information Provider: an indicator if the advertiser is an information provider or not. Format: *binary*, "1" represents yes; and "0" represents no.

Hostels: an indicator on if the advertiser is a hostel service provider or not. Format: *binary*, "1" represents yes; and "0" represents no.

Hotels: an indicator on if the advertiser is a hotel service provider or not. Format: *binary*, "1" represents yes; and "0" represents no.

Other Accommodations: an Indicator on if the advertiser provides accommodation service other than hostel and hotel or not. Format: *binary*, "1" represents yes; and "0" represents no.

Multi. Accommodation: an indicator on if the advertiser provides multiple accommodation service types or not. Format: *binary*, "1" represents yes; and "0" represents no.

Global Rank: the global rank of the advertiser's website traffic from the traffic estimator website Alexa. Format: *integer*.

Advertiser Rank: the advertiser's relative rank according to the global rank. Format: *integer*.

Other Products Offering: an indicator on if the advertiser provide products other than accommodation service products. Format: *binary*, "1" represents yes; and "0" represents no.

Chain: an indicator on if the advertiser is a chain hotel service provider or not. Format: *binary*, "1" represents yes; and "0" represents no.

Hotel Star Rating: the hotel star rating of the advertiser. Format: *real number*.

4 Citation Request

We are pleased to make these data sets public and for non-commercial use under the MIT License protocol. If any of the data sets described in this document are used in a publication, we appreciate a proper citation to the following paper.

> G. Tang, Y. Yang, and J. Pei. Price information patterns in web search advertising: An empirical case study on accommodation industry. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM'13)*, Dallas, TX, USA, December 7-10, 2013.

References

[1] G. AdWords. Understanding keyword tool columns. `http://goo.gl/rchmu`, 2012.

[2] Google. https://adwords.google.com, 2012.

[3] G. Tang, Y. Yang, and J. Pei. Price information patterns in web search advertising: An empirical case study on accommodation industry. In *13th IEEE International Conference on Data Mining*, ICDM'13, Dallas, Texas, USA, December 7-10, 2013.