Business Lead Generation for Online Real Estate Services: A Case Study

Md. Abdur Rahman, Xinghui Zhao, Maria Gabriella Mosquera, Qigang Gao and Vlado Keselj
Faculty Of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
{mrahman, xzhao, mosquera, qggao, vlado}@cs.dal.ca

Hai Wang
Sobey School of Business
Saint Mary's University
Halifax, Nova Scotia, Canada
hwang@smu.ca

***Abstract*—Business leads generation is a crucial and challenging task for online business to attract customers and improve their services. This paper presents a case study of an online real estate service company which provides potential home buyers useful neighborhood information, and accordingly offers them business leads to real estate companies. The company's current business lead generation is based on a "brute-force" method which requires tedious manual efforts. We developed an automatic business lead generation system which includes data integration as well as data mining tasks of classification and association rule mining. We demonstrated through experiments that this system can empower online real estate service companies to quickly and effectively generate targeted business leads.**

Keywords- Business lead generation; Data modeling; Data integration; Data mining; Business Intelligence.

I. INTRODUCTION

Over the past three decades, there has been a significant change in which information was collected, disseminated, and used. Form the customers' point of view, the information overload that they experience on a daily basis can hinder their decisions when selecting products and services to purchase. From the business organizations' point of view, the vast array of products and services being offered has made data mining a critical competitive advantage. Data mining techniques allow business organizations to improve their services and attract more customers. This paper presents a case study of business lead generation for an on-line real estate service company through data integration and data mining. A business lead, also referred to as a lead for simplicity, is defined as a potential sales contact, i.e., a potential customer who expresses an interest in the company's products or services [2].

In the real estate research literature, a recent study shows that almost 80% of home buyers start a home search online [8]. There are many online real estate service companies that provide potential home buyers useful neighborhood information, and accordingly provide them business leads to real estate companies. We conducted a study on a Canadian based online real estate service company. Currently, the lead products being offered by the company are based on a "brute-force" method which requires tedious manual efforts. There is no automatic lead generation system available on the market that could empower the company to quickly and effectively generate leads.

This paper proposes a business lead generation system which is capable of identifying effective landing pages on the websites of online real estate service companies. A landing page, also referred to as a lead capture page, is a web page to be displayed in response to clicking on an online advertisement which is posted on the Internet, either on the website of an online real estate service company or on a third-party website. For an online real estate service company, these landing pages may generate leads as potential home buyers may leave contact information on them. Data mining techniques such as classification and association rule mining are employed for lead generation.

The proposed automatic lead generation system can help the online real estate service company to increase revenue by selling more leads to real estate companies. It will also equip the real estate service company with a platform to reduce the cost of generating leads. Fig. 1 shows the business model of a typical online real estate service company:

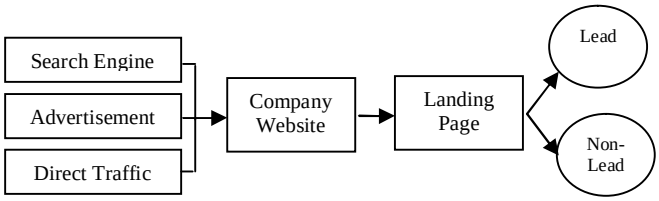

Figure 1. Business model for a typical online real estate service company.

Most of the users visit the online real estate service company's site through search engines or third-party websites where the company's advertisements are posted. While visiting the company's website, the users are exposed to a landing page. If a user visits the landing page and leaves his/her contact information then, the user is considered as a lead. On the other hand, if a user leaves the site at any point during their visit, the user is considered as non-lead.

To attract users, landing pages are designed carefully by considering the text size, style, color and contain various business offers and rewards. Generally, more than one landing pages are used to attract various user groups. The biggest challenge is how to provide the right landing page to the right user so that the user can become a lead.

 ISBN: 978-1-61208-185-4

The rest of the paper is organized as follows. Section 2 shows related work. Section 3 shows the system architecture of the lead generation engine; Section 4 describes the data model design and integration based on activity flow of users who use online services for home buying/selling purpose; Section 5 shows the data mining tasks, results, as well as interpretation of the results. Finally, Section 6 concludes the paper.

II. RELATED WORK

Currently, there are no lead generation tools available on the market. One of the main difficulties is due to large volume of data, which are distributed into several sources. This large volume of distributed data needs to be integrated into a single source before conducting any analysis. The data also needs to be transformed into a suitable format for data mining tasks to discover leads and non-leads.

In the research literature, automatic lead generation has been studied in a few cases for typical B2B and B2C types of e-commerce. [7] studied automatic sales lead generation based on online customers' survey, and [5][11] studied lead generation based on online customers' purchasing patterns. An online real estate service company differs from typical B2B or B2C e-commerce companies because it acts like a broker between the real estate companies and potential home buyers. To our best knowledge, there is no previous research on business lead generation for online real estate service companies.

III. SYSTEM ARCHITECTURE

For lead generation two major data mining tasks are involved in this system: Association rules mining and Classification. Association rule generation is a method for discovering interesting relationships among various item sets in a dataset and classification is the task of discovering a model first and applying it to predicate a new data object into one of the several classes of a predetermined target.

The system architecture is illustrated in Fig. 2. At first, the proposed system will gather all available data from the application. The primary data sources are the server log files, which include web server access logs and application server logs. Some additional data are also essential for both data preparation and pattern discovery, including the site files and meta-data, operational databases, application templates, and domain knowledge.

After gathering data from all the required sources, the web data will go through a pre-processing phase to clean the data by removing irrelevant or redundant entries from the data sets. Following the pre-processing phase, the data will be formatted appropriately according to the application business model. After formatting, the log data will be converted into a form suitable for data-mining task. The transformation will be accomplished according to the transaction model for that particular task (i.e., classification or association rule mining). Finally, after discovering the hidden statistics patterns from the data, the new discovered knowledge will be evaluated and filtered to only keep those which are statistically sound and novel to the business.

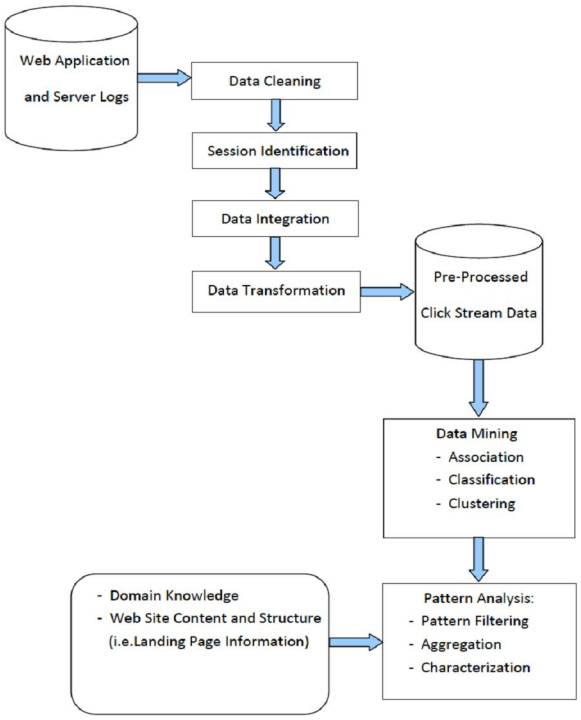

Figure 2. System Architecture.

IV. DATA MODEL DESIGN AND DATA INTEGRATION

Data model design is a critical step for data warehousing and data mining. It mainly involves four steps: identifying relevant datasets, feature selection, data integration and data transformation.

The current data repository setup, for our study of the online real estate service company, contains 3 datasets, which hold information for the period of January to June 2011. The "Dataset 1" contains the data pertaining to what users' actions were while browsing the website. This dataset also contains information about landing page visits where users' preferences about home buying/selling were recorded. The "Dataset 2" and "Dataset 3" contain search type information such as whether a user searched for schools, banks, and other retailers while searching for a home or neighborhood, as well as which ads were served. The "Dataset 2" also contains more detailed information about each search (e.g., search strings, referrer pages, user agent or browser used). A summary of these datasets is shown in Table 1.

TABLE I. SUMMARY OF DATASETS

Name	Size	#Cases	#Fields
Dataset 1	668.3 MB	2,197,320	10

 ISBN: 978-1-61208-185-4

Dataset 2	81.8 MB	1,015,332	7
Dataset 3	109.1MB	349,850	18

"Dataset 1" is dynamically generated in such a way that a new data instance is created for every different action performed by an user and the properties of a particular instance depends on the action of the user (i.e., submitting form, searching etc). In other words, the dataset contains many data instances for one user if the user performs more than one action. Therefore, this dataset is a semi-structured table in which one field may contain different types of information, and for one user there might be various numbers of rows associated with them. For data mining purposes, we normalized this table, combined all information related to one user into one single row, and then selected a fixed number of representative fields, which contain useful information for data mining.

Based on our analysis on user activity flows, we have divided the users into two groups and created a separate data model for each group. The two groups are "General User Data Model" and "Landing Page User Data Model". Landing page users are important and treated specially since they are more likely to become leads and they have more information recorded in the dataset.

A. General User Data Model

The general user data model covers all users including both the leads users and the non-leads users. The purpose of this data model is to form a single integrated dataset, which contains all activities of all users, i.e., including who did not reach any landing pages, as well as the activities of the users prior to reaching any landing pages. Table 2 shows the features for the general user data model.

Furthermore, to facilitate data mining tasks such as classification, we artificially created a new field "landing page related" to indicate whether the user eventually reached landing page and whether the user eventually became a lead. This new field is for the purpose of differentiating landing-page-users from non-landing-page-users, and differentiating leads from non-leads. Specifically, we use an integer value for the "landing page related" field. We assign the value "-1" for users who did not reach the landing page, "0" for non-leads users who reached the landing page, and "1" for leads.

TABLE II. GENERAL USER DATA MODEL.

Category	#	Features	Description
Before reaching website	1	referrer to website	referrer page to website
	2	ad served	ad served on referrer page
	3	ad name	ad name
Activity on website	4	page view on website	Website's page(s) being visited
	5	search address	address being searched
User info	6	IP location	IP location of user
	7	landing page visit	whether the user visit landing page or not
Additional	8	error	whether there is any error while performing any action.
	9	time stamp	activity time stamp

B. Landing Page User Data Model

For all users who reached landing page (leads or non-leads), we designed a separate data model, in order to accommodate the extra landing page related activities, as shown in Table 3.

TABLE III. LANDING PAGE USER DATA MODEL

Category	#	Features	Description
Before reaching website	1	referrer to website	referrer page to website
	2	ad served	ad served on referrer page
	3	ad name	ad name
On website, before landing page	4	page view on website	website page(s) being visited
	5	search address	address being searched (Prov.)
	6	referrer to LP	referrer page to landing page
Landing page info (LP)	7	LP selection method	method of selecting the LP
	8	LP version	version of LP
User info	9	Total 8 features in this category.	User's preferences about buying/ selling home.
Leads or not	17	submit	whether user submitted form to be contacted
Additional	18	error	whether there is an error
	19	time stamp	activity time stamp

We selected 19 features for the landing page user data model, which belong to 6 categories: activities before reaching website, activities on website (before landing page), landing page information, user information, lead or not, and additional information.

As shown in Table 3, most of the features are landing page related features, which do not exist for users who did not reach a landing page. Having separate data models can facilitate data mining tasks such as classification and association rule mining.

C. Data Integration Design and Implementation

We integrated the data from different sources into a single dataset for data mining. For each of the data models, we need to generate the integrated dataset. The data integration consists of two steps:

1. Integrate features from different tables into one table.
2. Integrate multiple rows which are related to one user into one row.

 ISBN: 978-1-61208-185-4

For Step 1, because the three original datasets (i.e., Dataset 1, Dataset 2 and Dataset 3) share a common id, we used that common id to join all tables. In Step 2, users' sessions were identified and all the activities during a single session were integrated into a single instance. The detailed algorithm is as follows and the corresponding flow chart is depicted in Fig. 3.

Algorithm:
Step 1: All Landing Page visits are identified.
Step 2: For each Landing Page visit, attributes are checked to identify whether this user has activities on the website's search page or not. If the user has search activities on the website's search page, all search related attributes are extracted for this user, if there is no search activity performed by the user, the program move to next step.
Step 3: User session is identified from "Dataset 1".
Step 4: All required attributes, for data model, are extracted from each row within the session.
Step 5: All extracted attributes, from a session, are stored in a new table within a single row where each single row hold data for only one session.
At the end, if all sessions are identified and data extraction is completed then the integrated file is generated for data mining purposes.

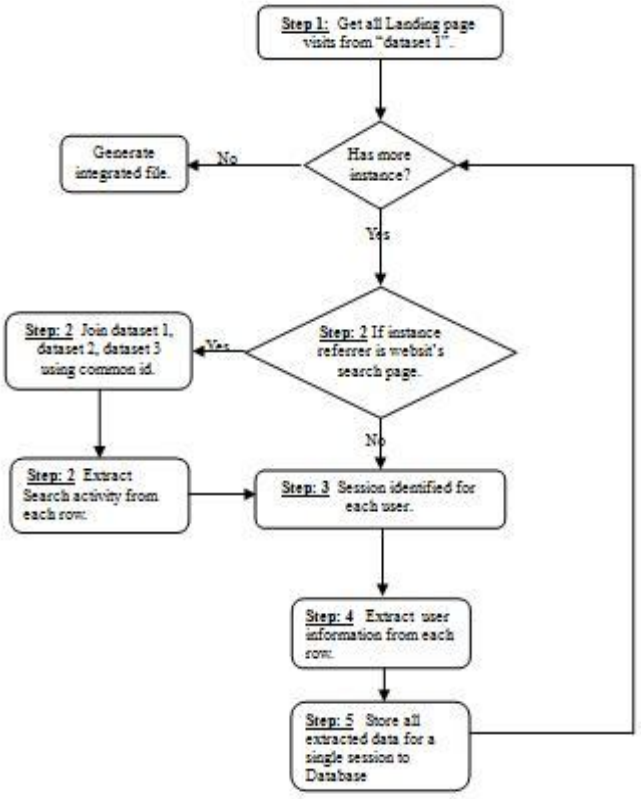

Figure 3. Flow Chart of Data Integration

V. DATA MINING AND EXPERIMENTS

We carried out data mining tasks of classification and association rule mining on the integrated datasets to identify effective landing pages and generate rules for leads.

A. Classification

The goal of classification is to find out whether a user profile belongs to lead user group, or non-lead user group. The output of our classification rule mining algorithm is a decision tree, in which the user models are represented as sets of rules for lead users and non-lead users. These rules can then be used to analyze the behavior patterns of the two groups of users. In this research, the classification algorithm C4.5 is applied to generate decision trees [9][6].

Fig. 4 shows an example of the decision tree generated by the C4.5 algorithm. We use the landing page version to show how decision tree can help to select the right landing page for right user. For example, landing page "lp11" has 100% lead turn out for the users who visit the site using advertisement named "X" and contain home buying price range for 200-400k. On the other hand, 90% of users who visit landing page "lp11", using advertisement for "Z" with the same price range, become non-leads. Clearly, for this category of users, "lp11" is not working well in terms of generating lead.

Using the decision tree, the company can analyze which landing page is working well for a specific category of users instead of choosing landing page randomly. In addition, the landing page can also be preselected based on the performance rate in converting users to leads.

```
pricerange = 200-400k
|       pageview = lp5 : Non Lead (0.96)
|       pageview = lp6 : Lead   (0.80)
|       pageview = lp7 : Lead   (0.50)
|       pageview = lp11
|       |       ad_name = Y : Non Lead (1.00)
|       |       ad_name = Z : Non Lead   (0.90)
|       |       ad_name = X : Lead   (1.00)
|       |       ad_name = M : Non Lead   (1.00)
|       pageview = lp10
|       |       intent = buy-sell: Non Lead  (0.75)
|       |       intent = buy : Lead   (0.70)
|       |       intent = sell : Non Lead  (1.00)
|       pageview = lp0 : Lead   (0.61)
pricerange = 400-600k
|       iplocation = ON : Non Lead  (1.00)
|       iplocation = NS : Lead   (1.00)
pricerange = 600-800k : Lead   (1.00)
pricerange = 800-1000k : Non Lead  (0.66)
pricerange = 1000k
|       error = no_error : Non Lead  (1.00)
|       error = yes : Lead   (1.00)
```

Figure 4. Decision tree from classification

B. Association rule mining

In data mining, association rule generation [10] is used

for discovering interesting relationships among various item sets of a dataset. We used association rule mining to find useful relationships among the user activities, landing pages and advertisements information of the company.

Association rules can be expressed in the form X => Y where X and Y are two disjoint subsets of all available items. We used the Apriori algorithm [1] to generate association rules. The algorithm can generate large quantities of patterns from the data, however most of which are of no interest. To remove the uninteresting rules, interestingness measurements techniques are applied to the result sets. Two commonly used measures are support rate (*supp*) and confidence rate (*conf*), which are defined as the following:

• Support rate of X => Y: The percentage of transactions in dataset containing both X and Y. Support(X,Y) = P(X∪Y).

• Confidence rate of X => Y: The percentage of transactions containing X and also containing Y. Confidence(X,Y) = P(Y|X).

However, the combination of support and confidence measurements are insufficient at filtering out the uninteresting association rules [4]. As a result, another measurement technique, "Lift" [3], is applied. Lift value is equivalent to the ratio of the confidence of the rule and the expected confidence of the rule. The formula for measuring the lift can be expressed as : Lift (X, Y) = P(Y) / P(X)*P(Y).

The value of Lift can be from 0 to infinity and can be interpret in the following way:

- If the value is greater than 1 then the rule (X=>Y) occurs more often than expected, which means that the occurrence of the rule body(Y) has a positive impact on the occurrence of the rule head(X).
- If the value is smaller than 1 then the rule (X=>Y) occurs less often together than expected, which means that the occurrence of the rule body(Y) has a negative impact on the occurrence of the rule head(X).
- If the value is near 1 then the rule (X=>Y) occurs almost as often together as expected, which means that the occurrence of the rule body(Y) has almost no effect on the occurrence of the rule head(X).

In our experiment, we set the support threshold to be 0.10 and Lift threshold to be greater than 1 to filter all positively correlated rules. To compare the effectiveness of the rules in terms of interestingness measures, both Lift and Confidence values are calculated for each rule. Some examples of the association rules are shown in Fig. 5.

The first rule shows that the users from "ON" (ip location, ON means Ontario, which is a province in Canada), who visited the site through advertisement "X" are more attracted to landing page version 10 and website page "page3". Therefore, the landing page "lp10" proved to be a high lead conversion page for users with iplocation "ON" and who visit the site through advertisement "X".

1. iplocation="ON", ad_name= "X" ==> site_pageview="page3", landing_page= "lp10" [conf:(0.73) , lift:(3.88**)]**
2. ad_name="Y" ==> landing_page="lp10", site_pageview = "page3" [conf:(0.69), lift:(3.67)]
3. pricerange="200-400k", intent="buy" ==> landing_page ="lp10", site_pageview ="page3" [conf:(0.67) , lift:(3.56)]
4. ad_name= "Y" ==> pricerange="200k" [conf:(0.5), lift:(1.6)]
5. landing_page ="lp11" ==> pricerange= "200k" [conf:(0.6), lift:(1.92)]
6. lp_referrer= "search engine x", choice= "NA" ==> landing_page= "lp0", address= "Other", iplocation= "AB" [conf:(0.75) , lift:(3.33)]
7. landing_page = "lp0" ==> lp_referrer="search engine x", site_pageview= "page1" [conf:(0.5), lift:(3.33)]
8. landing_page = "lp6", reward_choice= "yes" ==> ad_name= "Z" [conf:(1) , lift:(4)]
9. ad_name="X" ==> landing_page="lp10", iplocation="ON" [conf:(0.56) , lift:(3)]

Figure 5. Association rules generated from landing page data set

Rule number 7 shows that users whose referrer is "search engine x" and visit website's "page 3" are more attracted to landing page "lp0".

This knowledge can be used by the online real estate service company to increase the sales lead from different interest of users.

VI. CONCLUSION AND FUTURE WORK

This paper presented a framework for business lead generation in the area of online real estate services via a business case study. The framework includes data model design, data integration from multiple sources, data mining, and lead pattern evaluation. The results of classification rules can tell which groups of users are more likely to become leads, so that the corresponding pages/ads on the website may be emphasized for attracting those users. The association rules provide useful information for web designers, in that the togetherness of certain association patterns are important and the correlations among attributes in terms of desired target analysis. The knowledge base of lead generation should be updated periodically in order to best support business practices. In future research, the web page content data may be combined with web click stream data together form improving the lead generation results. This system may also be extended to build a real time lead generation system, which can select the most effective landing page by matching the user's request and the existing knowledge base. In addition, more rigorous experiments can be conducted to compare different data mining approaches

 ISBN: 978-1-61208-185-4

for lead generation, including the validity testing and performance evaluation of the data mining methods.

REFERENCES

[1] R. Agrawal, T. Imielienski, and A. Swami. “Mining association rules between sets of items in large databases”. Proceeding of the SIGMOD Conference, pp. 207–216. ACM Press, New York, NY, USA 1993.

[2] S. J. Bigelow. "Lead", August 2007. <http://searchitchannel.techtarget.com/definition/lead> [accessed October 1, 2011].

[3] D. S. Coppock. "Data Modelling and Mining: Why Lift?", June 2002. <http://www.information-management.com/news/5329-1.html> [accessed October 4, 2011].

[4] J. Han and M. Kamber, “Data mining Concepts and Technique”, Morgan Kaufmann Publishers, San Francisco, CA, 2006.

[5] D. V. D. Poel and W. Buckinx. “Predicting online-purchasing behaviour”, European Journal of Operational Research, Volume 166, Issue 2, 16 October 2005, pp. 557-575.

[6] J. R. Quinlan. “C4.5: Programs for Machine Learning. Morgan”, Kaufmann Publishers, 1993.

[7] G. Ramakrishnan, S. Joshi, S. Negi, R. Krishnapuram, and S. Balakrishnan. "Automatic Sales Lead Generation from Web Data”, Proceedings of the ICDE Conference, pp. 101-101, 2006.

[8] E. Weintraub. "Writing Purchase Offers in a Buyer's Market", <http://homebuying.about.com/od/offersnegotiations/tp/BuyersMKTOffers.htm> [accessed October 2, 2011].

[9] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand and D. Steinberg, "Top 10 algorithms in data mining", Knowledge and Information Systems, Volume: 14, Issue: 1, 1 January 2008. pp. 1-37.

[10] Y. Woon, W. Ng and E. Lim, Association Rule Mining, Infomration System, 77-82, (2009).

[11] X. Zhang, W. Gong, and Y. Kawamura. “Customer behavior pattern discovering with web mining”, Lecture Notes in Computer Science, vol. 3007, pp. 844–853. Springer-Verlag Berlin, Heidelberg, 2004.

 ISBN: 978-1-61208-185-4