**2016 Microsoft Research Asia**

**IFP for Korea Collaboration (KoCo)**

## **Project Name**

|  |
| --- |
| **Prediction of Customer Revisit Intention using Indoor Movements in Stores** |

## **Principal Investigator(s)**

*Name, title, department, university or institute, email, phone number, personal web URL*

|  |
| --- |
| Name: **Jae-Gil Lee**  Title: Associate Professor  Department: Graduate School of Knowledge Service Engineering  University: KAIST  Email: jaegil@kaist.ac.kr  Phone Number: +82-42-350-1617  Personal Web URL: <http://dm.kaist.ac.kr/jaegil> |

## **Abstract (150 ~ 300 words)**

*Give a clear summary of the overall project and the problem you propose to solve, your approach to investigating the problem and the tangible outputs. State what your proposed project will contribute to the area you work in and why it is relevant and worthy of MSRA’s support.*

|  |
| --- |
| **Customer revisit intention** is defined as the willingness of a customer to revisit a place (e.g., store). One of the major marketing principles is that customer retention is always a higher priority than customer acquisition. Thus, it is very important for merchants to identify who can be converted to **loyal customers** (repeat buyers). The purpose of this project is to **investigate the factors influencing customer revisit intention to offline stores by analyzing indoor movements**. We will obtain a license for the WIFI router data sets that hold the information of indoor movements in two or more KOLON Sport stores. We will carefully design the classification features that can be derived from such data sets and build a predictive model with machine learning or deep learning techniques. The outcome of the project is the methodology of constructing a predictive model, including a sophisticated design of discriminative features. Since the data sets in hand are very novel and this kind of problems has not been addressed before, we believe that the project will significantly contribute to the in-depth analysis of customer behaviors in offline stores. Our project is relevant and worthy of MSRA’s support, especially from the Social Computing group, because we aim at discovering human behaviors at a massive scale to support data-intensive business for retailers. |

## **Keywords (up to five)**

|  |
| --- |
| Revisit Intention; Repeat Buyer; Indoor Movement; WIFI Log; Machine Learning |

## **Project Description (No more than five, single-space pages, in 10 pt. font or larger)**

* *Description of the problem to be investigated and its technical importance*
* *Objectives to be achieved in contributing to its solution*
* *Proposed methodology to achieve the objectives, including diagrams or algorithms if this is helpful in describing your methodology clearly and concisely*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Introduction**:  The goal of this project is to develop a novel methodology of inferring the revisit intention of a customer in offline stores by using his/her movement data, which is illustrated in Figure 1. The revisit intention is defined to be the willingness of a customer to revisit a place in the near future. The project is based on the expectation that the customers satisfied and those unsatisfied may show different movement patterns in the stores. If we can predict whether a new customer will become a loyal customer early on (e.g., at his/her first visit), merchants can take advantage of this information to further increase the possibility of his/her becoming a loyal customer—e.g., by sending him/her discount coupons.    Figure 1 Illustration of indoor movements.  **Data sets**:  In order to work on this interesting problem, real-world data sets are required for testing a methodology. Currently, there are data sets from **two KOLON offline stores** shown in Figure 2. As shown by Figure 3, several **WIFI hotspots** are installed in each store. If the WIFI of a customer’s smartphone turns on, the smartphone is connected to one of these hotspots. Based on the hotspot to which the customer is connected, it is possible to know his/her current location in the store.    Figure 2 Two offline KOLON stores data are being collected.    Figure 3 Installation of WIFI hotspots for data collection.  For the Munjeong-dong branch, three hotspots (1F-inner, 1F-left, and 1F-right) are installed on the first floor, three hotspots (2F-inner, 2F-left, and 2F-right) on the second floor, and one hotspot (3F) on the third floor, as shown in Figure 4. A **region** is determined by the coverage of a WIFI hotspot. The region is the finest granularity of the location.    Figure 4 Floor plan of the Munjeong-dong branch.  Accumulating the connection information from all WIFI hotspots, a log record having the attributes in Table 1 are kept **per connection** in the server. Thus, in order to reconstruct a trajectory of each visit, we need to combine multiple log records if *device\_id*’s are the same and *ts*’s are consecutive. *device\_id* represents the MAC address of a smartphone. We do not know the identity of a customer, but we can at least figure out whether two customers are the same or not. *revisit\_count* and *revisit\_period* are automatically calculated by the server. A staff was required to register the MAC address of his/her smartphone. Then, his/her connection information is ignored by the server.  Table 1 Attributes of WIFI log records.   |  |  |  | | --- | --- | --- | | **Attribute** | **Description** | **Example** | | *device\_id* | MAC address of a mobile phone | aafc10f061882da0e1a043c95bc149e1 | | *dwell\_time* | Dwell time at a region | 35 | | *area* | Region in a store (see Figure 4) | 1F-left | | *revisit\_count* | Accumulate number of visits to a store | 52 | | *revisit\_period* | Days passed since the last visit | 1 | | *deny* | Staff or customer | true | | *ts* | Timestamp (Unix time) | 1449457815168 |   Table 2 shows the statistics of the two real-world data sets. It is worthwhile to note that the first store has a higher number of revisits than the second store. We are currently negotiating with the data owner to get the data sets from more stores.  Table 2 Statistics of the two real-world data sets.   |  |  |  | | --- | --- | --- | | **Characteristics** | **Munjeong-dong Branch** | **Hongik University Branch** | | Collection Period | 2015-08-28~2016-08-31 | 2015-08-28~2016-08-31 | | Total number of WIFI log records | 4,126,399 | 4,063,446 | | Total number of valid visits | 23,576 | 16,319 | | Total number of unique customers | 18,856 | 15,464 |   **Overall procedure**:  The features are extracted from each visit (i.e., combined trajectory), and the features from a single visit form a feature vector. As in Figure 5, the classification label—**revisit**—is true if the subsequent visit is within three months, and it is false otherwise.    Figure 5 Labeling of a store visit in the training set.  Figure 6 shows the overall procedure of our methodology, which is a common flow of predictive analytics. First, the features discussed in the next section are extracted, and each training sample is labeled according to Figure 5. Second, the training samples are provided to a classifier (supervised or deep learning) to build a predictive model. When a new customer comes in, the same set of features are extracted and fed to the predictive model, in order to estimate his/her revisit intention.    Figure Overall procedure of the methodology.  **Feature candidates**:  The most important task is to determine what to extract from indoor movements. We will empirically determine the important features and are currently considering the following features.   1. Stay-time features   In Figure 7, the **stay point** is defined as a region where the customer stayed longer than a given threshold, and its **stay time** indicates how long the customer stayed there. We expect that the total stay time as well as the distribution of stay times at different regions are related to the customer revisit intention.   * *Stay time*: e.g., 26 mins * *Number of stay points*: e.g., 6 * *Variance of the stay times at each stay point*: e.g., 21, 1, 1, 1, 1, 1 mins → * *Transit patterns between stay points* (aka temporally annotated sequence):   e.g., 3 mins  1f-right 2f-left (support = 0.2, confidence = 0.3)    Stay point: a region where the customer stayed longer than a threshold  Figure 7 Concept of the stay-time features.   1. Group-movement features   In Figure 8, a **group** is defined as a set of customers who entered a store at the same time as well as left the store at the same time. We also expect that the existence of accompanying persons influences the customer revisit intention.   * *Number of accompanying persons*: e.g., 2 persons * *Interaction with accompanying persons Proportion of the time* *being together*: e.g., 15 mins / 26 mins = 0.58     Figure 8 Concept of the group-movement features.   1. Personal or external features (optional)   If possible, we may like to use the personal information of a user (e.g., age, gender, address, and income level) and the social media articles that the user has uploaded before.  **Preliminary results**:  We conducted a quick test using part of the stay-time features (without transit patterns) on the Munjeong-dong branch data set. See the “X\_Basic” column in Figure 9. The accuracy is slightly above 0.6, which is a little better than a random classifier. Thus, there is a lot of room to improve using more sophisticated features. We aim at achieving accuracy of over 80%.    Figure 9 Preliminary results (accuracy) using only basic features. |

## **Major Expected Outputs and Plan for Disseminating the Results**

* *Social impact - [Benefit to a large number of users or plans to solve real-world environmental, social, research, government, and/or other challenges]*
* *Benefit to talent - [Training of high-quality research and/or engineering talent]*
* *Publications - [Books, chapters, journal papers, conferences, white papers and technical reports, as well as a plan for dissemination]. Note - we appreciate your acknowledgement of MSRA’s sponsorship in your publication.*
* *Data sharing – [What data is to be collected? Is data being made available to other researchers?]*

|  |
| --- |
| **Social impact**:  By targeting the potential **loyal** customers who are likely to revisit, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI). Similar studies have been conducted for online stores and online text reviews. In contrast, there has been no previous work for offline stores based on indoor movements. Since **offline purchases still account for a major proportion (over 90%) of total retail sales**, we believe that our research for offline stores will make a bigger social impact than previous studies.  **Publications**:   * We will try to publish our result at top-level conferences (e.g., ACM KDD 2017, UbiComp 2017, and VLDB 2017) or top-level journals (e.g., IEEE TKDE and ACM TIST). * We are willing to share the source code of our software through GitHub after the papers are accepted by the conferences or journals.   **Data sharing**:  Since the data sets we will use are proprietary, unfortunately, it is unlikely that we can share the data sets. However, we may share a part of the data sets with the grant from the data owner. |

## **Collaboration with Microsoft Research / Microsoft Research Asia**

*List possible collaborating researcher(s) and/or research group(s) and how the collaboration will be facilitated (for example, in the form of meetings, workshops, short stay at Microsoft Research lab, or through students as interns)*

|  |
| --- |
| **Dr. Xing Xie**  The collaboration will be done in the form of a short stay at Microsoft Research lab (more likely) or through students as interns (less likely). |

## **Adoption of Microsoft Technologies**

*Describe the planned use of Microsoft Research or Microsoft technologies, including operating system, programing language, server platform and/or MSR research technologies. For some topics, Windows Azure is the recommended platform.*

|  |
| --- |
| Most of the tasks for data preprocessing, learning, and prediction will run on Microsoft Azure instances. We already received a grant from Microsoft Azure for Research in 2016.   * <https://www.microsoft.com/en-us/research/academic-program/microsoft-azure-for-research/>   In addition, we may use Azure Machine Learning to run various machine learning algorithms.   * <https://azure.microsoft.com/en-us/services/machine-learning/>   The operating system and programming language are not that dependent on Microsoft technologies.   * Operating system: Microsoft Windows or Linux * Programming language: Python or R |

## **Windows Azure Resource Estimation (optional)**

*Describe your best estimation of Azure resources required by your project.*

|  |
| --- |
| Data storage: up to 5 TB (size of the raw log files and feature files)  Computation time: 24 hours 30 days 6 months 5,000 hours  We may need the Azure instances with GPU capability to run the GPU version of TensorFlow. |

## **Estimated Budget (US$)**

*Proposed funding request with detailed breakdown by deliverables. Purchase of hardware and software are not encouraged; instead, you may reserve a part of project funding for Windows Azure resources. The estimated expected budget is 10K-20K USD.*

|  |
| --- |
| The estimated budget to deliver the **prediction methodology** as well as **software** is as follows.   * Research assistant salary: $10,000/year 2 students $20,000 * MSRA visit and short stay: $5,000 * Purchase of additional data sets (optional): $5,000   ━━━━━━━━━━━━━━━━━━━━━━━━━  Grand total: $30,000/year |

## **Project Schedule**

*Please list the planned timeline of project milestones.*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The tentative schedule is as follows.   |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  | 2016 | | | | 2017 | | | | | | | | | **Task** | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | Data set preprocessing | → | → |  |  |  |  |  |  |  |  |  |  | | Algorithm design | → | → |  |  |  |  |  |  |  |  |  |  | | Algorithm development |  | → | → |  |  |  |  |  |  |  |  |  | | Feature engineering |  | → | → |  |  |  |  |  |  |  |  |  | | Accuracy tuning |  |  |  | → |  |  |  |  |  |  |  |  | | Paper writing |  |  |  |  | → | → | → |  |  |  |  |  | | Code clean-up & release |  |  |  |  |  |  |  | → | → |  |  |  | | Subsequent paper writing |  |  |  |  |  |  |  | → | → | → | → | → |   As the first step of the project, we would like to publish our result at top-level conferences such as ACM KDD 2017 or UbiComp 2017. After the paper is accepted, we are willing to make our algorithm and source code public. Then, while we will try to write a subsequent paper, eventually to make a big impact to the business, we will contact a few more retailers to materialize our methodology. |

## **Leverage other Resources**

*Additional sources of funding and/or how the project will leverage other projects or resources in the field.*

|  |
| --- |
| There is no additional source of funding up to this point. However, we are negotiating with the data owner to get additional data sets for free.  If the project succeeds and shows its potential, we expect that several retailers such as Walmart and Target are willing to adopt our technology, and the researchers in the field will be provided with abundant real-world data sets in the near future. |

## **Previous and Related Work**

*List previous projects (limited to top 3) related to the proposed topic, include project URL - if available, and papers (limit to top 3) you have published previously in the same area.*

|  |
| --- |
| **Related work**:  IJACAI 2015 held a competition for repeat buyer prediction. The input data set was online user behaviors including transactions and click behaviors. The competition home page is available at <http://ijcai-15.org/index.php/repeat-buyers-prediction-competition>. In addition, the papers by the winners are available at <http://socinf2015.isistan.unicen.edu.ar/winners>.  Since quite many customers are inclined to write reviews about restaurants on social media, e.g., Yelp, Yan et al. analyzed the quantitative scores of 10,136 restaurant reviews collected from an online life community in China and found that food quality, price and value, service quality, and atmosphere affect restaurant customers' revisit intention.  [Yan et al., 2015] Xiangbin Yan, Jing Wang, and Michael Chau, Customer Revisit Intention to Restaurants: Evidence from Online Reviews, *Information Systems Frontiers*, 17(3): 645~657, 2015.  However, as far as we know, none of the studies addressed the customer revisit intention in **offline** stores based on the customers’ indoor movements.  **Our previous work**:  As a related topic, we proposed a predictive analytics technique for user interruptibility on smartphones. **Interruptibility** is defined as the degree of how opportune it is to interrupt a person. This problem is useful especially in determining the proper moment for sending messages or notifications. Similar to the project, this previous work used vast amounts of activity log data and predicted human behaviors.  [Choy et al., 2016] Minsoo Choy, Daehoon Kim, Jae-Gil Lee, Heeyoung Kim, and Hiroshi Motoda, Looking Back on the Current Day: Interruptibility Prediction Using Daily Behavioral Features, In *Proc. UbiComp 2016*, pp. 1004~1015, Sept. 2016. |