Data collection and analysis of usages from connected objects: some lessons

Sara Meftah[1], Antoine Cornuéjols[1], Juliette Dibie[1], and Mariette Sicard[2]

(1) UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay,
75005, Paris, France
(2) Research Cooking & Food, Groupe SEB, 21261 Selongey - France
{sara.meftah,antoine.cornuejols,juliette.dibie}@agroparistech.fr,
msicard@groupeseb.com
http://www.springer.com/lncs

Abstract. The emergence of widely available connected devices is perceived as the promise of new added-value services. Companies can now gather, often in real time, huge amounts of data about their customers' habits. Seemingly, all they have to do is to mine these raw data in order to discover the profiles of their users and their needs.
Stemming from an industrial experience, this paper, however, shows that things are not that simple. It appears that, even in an exploratory data mining phase, the usual data cleaning and preprocessing steps are a long shot from being adequate. The rapid deployment of connected devices indeed introduces its own series of problems. The paper shares the pitfalls encountered in a project aiming at enhancing the cooking habits and presents some hard learnt lessons of general import.

Keywords: Data Mining, Internet of Things, Data Preprocessing.

1 Introduction

1.1 The promise of gathering data from connected objects

Recent years have witnessed the arrival of a new concept, that of smart connected products and devices, which, all together, will make the *Internet of Things*.

One major driving force is to get a direct access to product usage data. By analyzing massive amounts of data about usages, companies aim at forming new kinds of relationships with customers. The accumulation and analysis of product usage data should enable them to gain fresh insights into how to create new values for the customers, therefore ensuring closer ties and increased loyalty.

This is in this perspective that the industrial project that serves as a case study in this paper has been launched. The project is about cooking devices and habits. It is part of a large undertaking to improve public health by measuring alimentation behaviors. Its goal is to gently try to nudge consumers towards healthier behaviors if needed, and at least to offer and suggest them a more varied diet (see [3] for a related purpose).

This new public health approach, user-centered and in real-time, is enabled by the possibility to equip kitchens with connected devices that both send data about the cooking procedures followed and offer new interfaces to the users. These interfaces provide them with descriptions of recipes and suggestions and gather information from the customers.

In the following, we first present in Section 2 the exchanges of data that are possible with the new devices, while Section 2.1 reports the kind of questions that are expected to be solved by analyzing the data. Section 3 then turns to the data mining processes that were attempted and shows the difficulties encountered. Section 4 shows that these difficulties are largely intrinsic to the deployment of smart connected devices, independently of the field of cooking study. It is thus beneficial to draw general lessons from this experience.

2 Case study: analyzing data about cooking behaviors

2.1 The questions

When equipping users with smart connected devices, e.g. e-health watches, smartphones, or cooking devices, and then gathering data about the users' habits, companies share general questions such as:

1. Does a *categorization of the users* emerge from the collected data, either
   - directly from their profile available during the buying procedure or before the first interaction with the device
   - or, indirectly, from their measured interactions with the device(s)
2. Is there, and what is, a *typology of the usages*? For instance,
   - from the recipes that are consulted on the application's website
   - from the recipes that are effectively followed
3. Is it possible to identify interesting *relationships between classes of users and classes of behaviors* or classes of recipes?

2.2 The available data

When a connected device is rented or sold to a customer, the information gathered by the provider is of two qualitatively different types.

1. The first is akin to a *factual description of the client*. That could be the company or client's name, the revenue, number of employees, number of dependents, age, and so on. Sometimes this description is readily available because the contract cannot be signed without it, sometimes, it comes from the voluntary filling of information from the user.
2. The second type includes all the information that can be collected during the *interactions of the user with the device*, or during the device's operations. Such data can take the form of logs listing temporal actions or procedures, or they can also trace the user's connections to a website purported to provide operating or maintenance information. In that case, the logs of interactions can be completed with the content of the web pages that have been consulted.

As an illustration, in the context of the industrial project on the analysis of cooking habits, data was collected on approximately 100,000 users, covering a period of more than 24 months, from January 2014 to March 2016. Data were obtained when the user voluntarily filled a form. The file describing the *users' profiles* thus provides information about 'age category', 'city', 'gender', 'number of children', 'number of adults' in the family, and 'type of device'.

The behaviors of the users, and their habits, were measured thanks to an application available for smartphones. On one hand, this application allows the users to access a list of recipes approved by nutritionists, to select some of them, to look for information, and to rate them. On the other hand, the application can also be used to remotely control the cooking device in order to automatically perform complex cooking operations. In this way, sophisticated recipes can be realized, but also, data about the user's usage can be gathered. Therefore, each cooking session by a user produces a mixture of operation logs and sequences of textual contents describing the web pages and recipes that had been looked onto during the session and the appreciations possibly provided by the user.

In addition, there is a file recording *relations between users and recipes*, in the form of bookmarks (plus date and time) that each user can put on recipes he/she would like to remember for future use. For instance, this file recorded approximately 9,000 bookmarks put by 2,000 users ($\approx 2\%$ of all users) on 400 recipes ($\approx 11\%$ of the recipes). Conversely, there is also a file about the bookmarks that were removed after a single session.

The users can also evaluate the recipes by grading them. And there are files collecting data about the usages of the cooking devices. One such file contains information about automatic launches of cooking operations. Another file keeps details about the web page navigation by the users on the supplier's website. More than 4,850,000 events were thus recorded from more than 4,500 users at the time of this study. Each event in the file is associated with specific information about the webpage accessed, the time and date, the duration of the consultation, and so on.

2.3 The methods: Exploratory data analysis

In order to better understand how the cooking devices and the associated services are used and how the users' habits are related to the users' profiles, a wealth of machine learning techniques were used, including:

- *Univariate analysis*, e.g. using histograms of distributions for each variable.
- *Visualizations* in 2D or 3D to help discover correlations between variables.
- *Clustering* in order to detect categories in users and in recipes.
- Discovery of *Frequent Item Sets* and *association rules* both within the users' profiles or the habits descriptions
- Modeling of the dynamics of the user's habits using *Markov chains*.
- *Supervised classification* in order to understand what determined that some recipes would be tagged as "favorites" or "not favorites".

All methods required the use of preprocessing steps before they could be carried out. Some operations were standard, others involved specific knowledge in the form of domain ontologies. Briefly, the following methods were used:

- *Data cleaning.* E.g., lots of variations were encountered in the address fields.
- Processing of *missing information.* As is so unfortunately too often the case, there existed default values for some data fields, like number of children, or number of adults in the household. We devised rules that estimate the probability that these default values were in fact missing values given the answers to other data fields.
- *Data enrichment.* The descriptions of the recipes were too fined grain for useful subsequent data analysis. For instance, there were 483 different ingredients mentioned, such as 'salt', 'oil', 'ham', 'parmesan'. In order to perform clustering or association rules discovery, we decided to derive more abstract description of the recipes, using an ontology[1]. 'Tomato' could thus be replaced by 'vegetable'. (see also [4, 5]). In addition, ontologies allowed us also to *add* information, such as the type of diet implied by the use of some ingredient, or the type of course during the meal: entry, main course or dessert.
- *Value imputation.* One important input is whether the user likes or dislikes a process, here a recipe. In our setting, users could provide a grade to the consulted recipe through the interface of the application, however, they were a minority to use this fixture. We thus decided to infer the like/dislike values from the behavior of the customers. One source was the observation of the bookmarks put by the users on recipes. We decided for instance that a user that puts a bookmark on a recipe and consults it at least another time is likely to 'like' this recipe. Conversely, a user that removes a bookmark after a single consultation is viewed as 'disliking' this recipe. Likewise, we decided that a user who launches an automatic cooking session using a recipe probably 'likes' this recipe. Finally, a user sharing a recipe on social networks was also deemed to 'like' this recipe. In this way, we were able to substantially increase the proportion of recipes qualified by like/dislike appreciations.

3 Analyzing the data

The analysis was organized along three main objectives: *first*, categorizing the users' profiles, *second*, detecting patterns of cooking habits and cooking preferences, and, *third*, see whether there exist some relationships between users's profiles and cooking habits and preferences.

3.1 Analysis of the data describing the users

A preliminary study involved the examination of the distribution of the values for each field: 'age category', 'gender', 'type of device', 'number of adults' and 'number of children' in the household.

[1] We used the TAAABLE ontology [6], which is the most encompassing one for analyzing nutrition and food in general.

In the process of computing statistics for each attribute, it was found that a significant proportion of them were not filled out by the users, or, worse, were filled with a default value. In fact, even the type of the cooking device was not provided by about 10% of the users. This is remarquable since it could be expected that this information should be collected automatically. As was mentioned in Section 2.3, we had to resort to heuristic rules to detect and, if possible correct, the values resulting from filling by the default value.

Altogether, these defects in the information collected about the users were detrimental for more refined analyses, such as clustering.

3.2 Analysis of the data describing the cooking behaviors

In this case study, the idea was to identify typical behaviors measured through the logs of interaction of the users with the devices. The goal was to examine whether there were characteristic temporal patterns in the use of the recipes, for instance during the week or during the year, and, generally, to measure in which way the application's services were used, whether there were steps that were bypassed, others that should be more informative, and so on.

To answer these questions, it is crucial to be able to determine exactly what were the webpages that were consulted, in which order and what was the duration of the consultation for each page. It has been underlined in the literature (for instance, [7]), that there can be impediments from the way Web servers are organized and operate. For instance, because of proxy and local caching, it can be difficult to detect that a user is going back to a page already viewed. With connected devices, however, the operations can be thought anew, and such hindrances should be limited or eliminated.

In our application, it can be determined that a web page has been accessed thanks to the 'page load event'. The duration of consultation of a page can be determined through the records of the 'page load' and 'page unload' events since each is associated with a timestamp.

The application environment draws a distinction between 'content pages', which essentially describe recipe steps, and 'navigational' or 'mobile application pages' which allow the user to navigate between the application services. There are 76 such different types of navigational pages. When trying to analyze the behaviors of the users, it quickly became apparent that problems were looming.

First of all, looking at the content pages, it was readily obvious that the distribution of the durations of consultation exhibited abnormal results. Altogether, 262,488 consultations of pages describing recipe steps were recorded, with a mean duration of consultation of 98s and a standard deviation of 3,326s, clearly out of normal range. Actually, it was discovered that the maximal duration time was 119 hours! An histogram of the duration times showed that there was a significant proportion of outliers, and that this proportion was higher for the last steps of the recipes that for the first ones. What then was the reason behind this odd and unhelpful set of observations?

After some analysis and discussion with the development team, it was found that the consultation of a page did not have to be formally closed by the user,

before he/she was moving to another page, or discontinuing altogether the current cooking session. The current page could be put in the background or the application could be left in its current opened state for days, before another session was started. The confort of use of the application was the foremost concern for the development team, while the demands of any future data analysis was not really considered.

Another source of frustration emerged from the study of the consultations of the 'navigation' pages. There were 1,841,044 pages consulted in the course of 27 months, from January 2014 to March 2016. We quickly discovered that some pages were redundant, having closely related meanings for the users, or providing exactly the same information. We found also that there had been changes in the application navigation system over time, with various modifications in the architecture of the system. For instance, 56 out of the 76 types of navigation pages were added starting from November 2015, without notification to the data analysis team. Again, the changes in the application were driven by concerns about the usage of the device, with no regard for the data analysis needs.

Finally, it was difficult to analyze the search behavior of the users, since if a search conducted among the 'navigation' pages was concluded by a click on a recipe, the search was erased from the user's history, and he/she had to repeat it entirely if needed, causing havoc in the statistics about the viewed pages. When trying to figure out if there were typical search paths, this resulted in obtaining Markov chains that were difficult to interpret.

3.3 Relationships between types of users and cooking behaviors

One way to search for relations between the users' characteristics and types of recipes is to perform *clustering*: on the users' description, on the one hand, and on the recipes on the other hand. Another way is to look for *association rules* between user's profiles and types of recipes.

In both analyses, we got interesting and interpretable results. Clustering allowed us to identify marked relationships between clusters of users (2 clusters) and clusters of recipes (3 clusters). Likewise, we uncovered association rules that made sense, like, for instance: women with more than one child look for easy recipes, or men at least 53 years old prefer recipes with less cholesterol.

However, it must be said that these findings are tentative since they rest on fragile and fault prone measurements.

4 Lessons for the design of IoT for collecting data on usages

Several general lessons can be drawn from the difficulties encountered in analyzing the data about cooking habits and preferences. They are interesting because we believe they potentially apply to many industrial projects that aim at delivering connected devices to their customers in order both to bring them new services and to gather valuable data about their usages. We list them in the following.

1. All too often, *data collected about the users* and their characteristics *are incomplete*, when they are not *erroneous*. The fundamental reason lies in the fact that users will not spend time to fill information if they do not perceive in which way this is critical to the service they get. For instance, users of connected e-health devices, like connected watches, know that they better provide the service with their precise age and weight, because otherwise the assessment of their performance is senseless. In the case of the connected cooking devices, such a link between the information asked to the user and the service provided was far less apparent.
   The remedy is therefore to make obvious to the user that it is in his/her own interest to provide "useful" information in order to get a true benefice from the service. This is what has been done since this study.
2. The *information available from the logs of the device operations and/or the user's interactions was altogether almost useless*. Indeed, the team of designers and developers of the devices and of the user's interface was naturally obsessed with their ease of use and with the technical aspects of the device and interface. If the team was aware of the data collecting role of the devices and interfaces, this role was not a foremost concern.

   One remedy is to mix together in a single team the designers, the technical staff and the data scientists, or, at the very least to ensure a strong communication channel between them. Another remedy is to teach the basics and demands of data science to every one implied in the project.
3. When the data collected over several months of operation was analyzed, it quickly became apparent that there had been *changes in the types of data collected, or, even worse, in the semantics of some measurements*. This was due to the rapid deployment of the devices to the end-users while the design and realization of the device, and above all of the user's interface was not stabilized. Again, the primary concern of the designers and technicians tend to be the proper working of the device and interface. And because the field of connected devices is in such a frenzied state, designers and technicians are in a agile state of mind with rapid prototyping of new softwares and their deployment to the users. The data collecting goal is second.

   One remedy is again to provide education about data science to every one involved. Then, when changes affecting the data available are in order, programs for translating data from one period to another should be produced, or, at least, meta data should accompany any data that is collected.

5 Conclusion

In the last two or three years, the Internet of Things has been widely heralded as a revolution in the making, that could shadow even the already "old" Internet revolution. Nonetheless, among high expectations about what the new area could bring to our lives, some warnings were voiced. For instance, Vinton Cerf, chief inventor of the Internet, observed that a lack of standards could hinder the

development and operations of the Internet of Things [1]. Ease of interoperability between different systems, and ease of communication in networks composed of millions of objects, plus resilience in front of possible attacks and privacy preservation were also recognized as essential assets that must be secured if IoT is to become a reality.

But, while all these technical hurdles started to be appraised, other concerns, of as much fundamental importance, have been largely overlooked. Indeed, when connected objects are delivered to a client and installed, the provider and/or the client expect that one prominent service will be the collection, sometimes in real-time, of day to day data. One goal can be to better predict breakdowns, therefore ensuring a smoother service. Quite often, though, it is deemed even more important to gather information about the usage of the connected devices in order to bring improved experience to the users, and possibly new services.

It has been said that the largest challenge for businesses will be determining how to use the tremendous volume of new data that Internet of the Things will generate (see for instance [2]). Ironically, as this paper shows, the data collecting role itself is often overlooked by the people who design and deploy the connected devices. One reason is that they are already overtaxed with trying to solve all the technical problems mentioned above. Another reason is the frantic pace with which the technology evolves, a pace which conducts the technical teams to adopt an agile strategy, with many adaptations along the way. Unfortunately, as our experience demonstrates, these changes and the lack of a clear perception of the demands of data analysis, can ruin the very purpose of the whole operation.

This paper, we hope, will thus help promote a new awareness of the challenges set when the production of exploitable data from connected objects is aimed at.

References

1. Fisher, L.M.: Cerf Cites Challenges Facing the Internet of Things. Communications of the ACM (ACM News). November 6 (2015)
2. Violino, B.: The 'Internet of things' will mean really, really big data. InfoWorld, July 29 (2013)
3. Teng,C. and Lin, Y., and Adamic, L..: Recipe recommendation using ingredient networks. In Proceedings of the 4th Annual ACM Web Science Conference, pages 298–307. ACM, (2012).
4. Jonsson, E.: Semantic word classification and temporal dependency detection on cooking recipes. (2015).
5. Amano, S., Aizawa, K., and Ogawa, M.: Food category representatives : extracting categories from meal names in food recordings and recipe data. In Multimedia Big Data (BigMM), 2015 IEEE International Conference on, pages 48–55. IEEE, (2015).
6. Cordier, A., Dufour-Lussier, V., Lieber, J., Nauer, E., Badra, F., Cojan, J., Gaillard, E., Infante-Blanco, L., Molli, P., Napoli, A., et al.: Taaable : a case-based system for personalized cooking. In Successful Case-based Reasoning Applications-2, pages 121–162. Springer, (2014).
7. Mobasher, B. and Cooley, R. and Srivastava, J.: Automatic personalization based on web usage mining. Communications of the ACM, Vol.43, No.8,142–151, (2000).