Food Data Automatic Classification and Clustering

Author: Ade Romadhony

Email: ade.romadhony@gmail.com ade_romadhony@students.itb.ac.id

Slack Id: @aromadhony

OpenFoodFacts User Id: @aromadhony

Author Info

Author's short profile:

I am a 5th year PhD student in Institut Teknologi Bandung and also a lecturer in Telkom University Bandung. My dissertation topic is "Template Extraction from Open Information Extraction Result". My research interests include: text mining, machine learning, and natural language processing. In Telkom University I taught several courses: algorithm and programming, information retrieval, machine learning, and natural language processing. During my third year of PhD study, I collaborated with other PhD students from University of Indonesia and TU-Wien. We built a Strike-Sensor project and my contribution was building an Indonesian Semantic Role Labeling (SRL) system. Our work has been awarded Bernd-Rode Award (BRA)*.

*http://asea-uninet.org/scholarships-grants/bra-laureates-2017/

Project experiences:

- Online hybrid recommender system
- Indonesian Hadith retrieval system
- Truecasing for Indonesian text
- Initial study on Indonesian Semantic Role Labeling (SRL) system
- Online mobile collaborative annotation tools (KataKita)
- Open Information Extraction (Open IE) Based event extraction

Summary

In this GSoC, I am interested to participate in Data Science project, since it is relevant with my interests. According to the information on Data Science ideas list, there are two tasks: product classification and error detection. To understand the data structure and the task better, I have examined the dataset and wishlist listed on github AI issues. I also tried to solve an issue and discover several task suggestions.

I pick the automatic **category classification** task as a starting point, since I think category and product name fields are two important information on performing other tasks. Below are my suggestions:

1. **First of all, we should organize the classification structure.** From my understanding, the category structure only consist of 1 level and parent and more specific categories were collapsed into the same level. For example, in main category fields, there are *tuna-steaks*, *fish and egg and meats*, and also *meals* category. It makes the automatic classification and clustering tasks more complicated. Based on my initial observation, automatic category classification can be solved as multi-label classification task. Another finding is no mapping information for category names in different languages.

   As for label classification, we can use information from ingredients, allergens, traces, additives, and energy fields.

2. For the **clustering** task, we must **determine which category** that need to be splitted further and **which fields/attributes that have high information gain**. One of basic assumption on splitting the existing category is based item numbers in a category. If the item numbers of a category exceed certain threshold, then we need to split it into several sub-categories.
3. Error detection -> could be performed using outlier detection technique.

Proposed Solution

1. Reorganize category structure, based on a food classification standard. If there is no suitable standard, a minimal work should be done on redefining and reorganizing the main category.

2. Category and label classification using information from the following fields: product name, ingredients, allergens, traces, additives, and energy fields. The attributes are represented by bag-of-words model. Proposed method: SVClassifier (for main category field) and multi-label SVClassifier (for category field).
3. Brand classification -> I tried to examine the code/barcode prefix, but still have no clue on employing this information. Utilizing heuristic rules derived from barcode standards will be helpful..
4. Clustering, automatically construct subcategories.
    a. Identifying the category that need to be classified further. -> assumption: category with large members
    b. Identifying the attributes for clustering -> proposed method: feature selection using hierarchical clustering [1]
    c. Giving label to clusters -> proposed method: topic detection by clustering keywords [2]
5. Error detection. Performing outlier detection, based on ingredients and other fields: allergens, traces, additives. Proposed method: kNN distance based outlier scoring [3]

References:

[1] Butterworth, R., Piatetsky-Shapiro, G., & Simovici, D. A. (2005, November). On feature selection through clustering. In *Data Mining, Fifth IEEE International Conference on* (pp. 4-pp). IEEE.

[2] Wartena, C., & Brussee, R. (2008, September). Topic detection by clustering keywords. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on* (pp. 54-58). IEEE.

[3] Ramaswamy, S., Rastogi, R., & Shim, K. (2000, May). Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record* (Vol. 29, No. 2, pp. 427-438). ACM.

Workplan

No	Activities	Week												
		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
1	Analysis on category structure/hierarchy													
2	Reorganizing category structure													
3	Automatic classification: category													
4	Automatic classification: label													
5	Automatic classification: brand													
6	Evaluation of automatic classification													
7	Clustering: creating subcategories													
8	Evaluation of clustering													
9	Error detection													
10	Evaluation of error detection													
11	Overall evaluation, final product													

	and final report submission													

Week:

I: May 14

II: May 21

III: May 28

IV: June 4

V: June 11

VI: June 18

VII: June 25 (Eidil Fitr holidays)

VIII: July 2

IX: July 9

X: July 16

XI: July 23

XII: July 30

XIII: August 6

I made the work-plan based on the assumption that I will allocate 6 hours per weekday for the GSoC project. I also have considered the proposed methods that I already mentioned before. Although the proposed methods are not the most sophisticated and the most recent methods, I think the methods are suitable to to the problems, and feasible to be implemented in such tight schedule (including the learning curve on the existing system).