Intelligent Combination of Food Composition Databases and Food Product Databases for Use in Health Applications

Alexander Muenzberg[1,2(✉)], Janina Sauer[1,2], Andreas Hein[2], and Norbert Roesch[1]

[1] University of Applied Science Kaiserslautern, Amerikastr. 1, 66482 Zweibrücken, Germany
{alexander.muenzberg,janina.sauer,norbert.roesch}@hs-kl.de
[2] Carl von Ossietzky University Oldenburg, Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany
{alexander.muenzberg,janina.sauer,andreas.hein}@uni-oldenburg.de

Abstract. The necessity of using food data in mobile health applications is often linked with difficulties. In Europe no standardized and quality-controlled food product databases are accessible. Data from third party sources are often incomplete and have to be checked carefully before use for errors and inconsistencies. The purpose of this approach is to improve data quality and to increase information density by developing a dedicated food data warehouse. By using the extract, transform and load processes known from data warehouse technologies, multiple data sources will be combined, inserted and evaluated. The data is cleaned up by using data profiling techniques. Data mining methods are used to merge the datasets from food composition databases and food product databases to increase information density. The aim is to analyze, if and how Big Data technologies can increase performance of data processing significantly.

Keywords: Food data · Data analysis · Data mining · Big Data

1 Introduction

The use of mobile health applications (health apps) is constantly increasing in the app stores of mobile platforms and many of them are focused on nutrition. However, no standardized sources of food product data are available in Europe with complete information of available food products on the market. Most app developers are dependent on their own data or on data collections of third parties. Such datasets are often incomplete or have not been sufficiently checked to guarantee a data quality which is sufficient for medical use. Missing or incomplete datasets can have negative effects on the data quality and therefore on the quality of the app. This circumstance reduces the user's trust in the app to a high degree. Provider of quality-controlled and verified food datasets are limited to a few and usually offer their data at a very high fee [1, 2].

Within the project "Digital Services in Nutritional Counselling" (DiDiER) [3], funded by the German Federal Ministry of Education and Research (BMBF), a food data

G. M. P. O'Hare et al. (Eds.): MobiHealth 2019, LNICST 320, pp. 305–319, 2020.
https://doi.org/10.1007/978-3-030-49289-2_24

warehouse system was developed, which combines the food data of different, mostly free or public data sources and stores them in a uniform data format and data structure. Using data profiling tools, errors and duplicates were detected and eliminated [1]. By combining food composition databases with food product databases and creating ontologies between their data elements, inconsistencies among the data are detected and the datasets are completed. In order to obtain hidden information of useful value from the data and ontologies, data mining methods are used.

2 State of the Art

2.1 Food Data Warehouse

The Food Data Warehouse (FDWH) contains a collection of information about natural foods, packaged foods and branded products. The data was obtained from the sources of various platform operators and food manufacturers. The data was extracted from the external data sources, transformed into a uniform data structure as well as data format and then loaded into a central relational database by the Extract, Transform and Load process (ETL process) known from data warehousing. The FDWH currently contains data with information about approximately 40,000 foods from the following extern data sources [1].

- Food Composition Databases
  - German Federal Food Key (Bundeslebensmittelschlüssel [4]
  - Swiss nutrition database (Schweizer Nährwertedatenbank) [5].
- Food Product Databases
  - WikiFood.eu [6]
  - das-ist-drin.de [7]
  - OpenFoodFacts.org [8]
  - FoodRepo.org [9]
  - Danone and its subsidiary Nutritcia [10, 11].

Data profiling methods were used to detect and correct defective characters, incorrect data formats and duplicates of data records. Data profiling includes the following tasks in detail.

- Checking patterns and data types
- Outlier detection
- Characterization of missing and preset values
- Data rule analysis (e.g. recognition of values corresponding to certain regular expressions)
- Analysis of column properties (validity check of all values in a table column)
- Analysis of value dependencies across columns
- Recognition of functional dependencies or foreign key dependencies in databases.

Using special metrics, which enable data quality to be measured, quantifiable quality values in terms of completeness and consistency were determined in order to obtain an initial assessment of the data quality improvement using the above-mentioned data profiling methods [1].

2.2 Food Information for Health Applications

For the analysis of food intolerances and allergies, correct and complete lists of ingredients are required e.g. for the detection of related elicitors. Although it is required in the EU that information on 14 major allergens must be labelled by manufacturers on food packages, also other ingredients and additives have the potential to provoke allergic reactions [1]. In the field of Frailty Syndrome, the focus is on energy intake. Information about nutrition facts, in particular of energy, fat, protein and carbohydrates is needed. Information about Nutrition Facts is important for a wide range of health fields, such as healthy nutrition and fitness as well as the prevention of diseases such as obesity, cancer, diabetes, heart attack or stroke [12].

2.3 Differences Between Food Composition and Food Product Databases

Food Composition Databases (FCDB) provide the name of foods in conjunction with their nutritional information. Unlike Food Product Databases (FPDB), FCDBs do not provide branded products but natural food without branding. The following nutritional information is provided using FCDB data [5].

- Energy value in kilo joules (kJ) and kilo calories (kcal)
  - calculated from the sum of the kJ values of
    - carbohydrates 17 kJ/g
    - protein 17 kJ/g
    - fat 37 kJ/g
    - alcohol 29 kJ/g
    - food fibers 8 kJ/g.
- values of macronutrients
  - carbohydrates, food fibers, fat, cholesterol, protein, alcohol, water.
- values of vitamins
  - vitamins a, b1, b2, b6, b12, c, d, e, etc.
- values of the minerals among other things
  - sodium, salt, potassium, chloride, calcium, magnesium, iron, iodine, zinc, etc.

In contrast to the FCDB, the FPDB data also provides the information in the following.

– brand
– origin
– Global Trade Item Number (GTIN)
– information about packaging and special features
– content quantity (for packaged products)
– ingredient lists
– allergens requiring labelling.

2.4 Food Information Service

The Food Information Service (FIS) is a webservice that provides FDWH data via an Application Programming Interface (API) [13]. The FIS receives a request from the application with username and salted password hash [14] (for secure authorization) as well as a search string. Within the FIS, the identifiers (IDs) of suitable foods with reference to the search string are selected and returned to the requesting application. A reference of the food to the search string exists if the search string completely or partially corresponds to the food name, the manufacturer name or the category as well as the soundex values (code generated by the sound of a name) [15] of the search string and the food name correspond. Based on the selected IDs, the application can now request more detailed information about the associated food. Both the requests and the answers provided by the FIS are in the human- and machine-readable JavaScript Object Notation Format (JSON format) [1, 16].

2.5 Ontologies Between Food Data

In the context of information theory, ontologies are special data models that formally define data objects of a subject area [17]. When evaluating the data of the FDWH such data models are generated in order to extract specific knowledge about individual data elements, with the help of which it can be recognized whether it concerns a consistent value. Furthermore, in many cases missing values in the databases can be derived with the help of such models.

In the following, the ingredient lists, which are available in text form, are preprocessed using Text Data Mining (Text Mining) [18] so that further ontologies can be formed between the FCDB and FPDB datasets and their attributes in order to extract missing information in data as far as possible from other similar or identical datasets.

3 Methods

3.1 Preprocessing of Ingredient Lists by Text Mining

Before information about the ingredients can be generated by the food data records, the ingredient lists, which are available in text form, must be processed using text mining methods. The supplied ingredient lists of the external data sources were also stored in the FDWH as comma separated text strings, for presentation in health apps. In the further course, the individual ingredients are extracted from these strings.

By tokenization [18], the ingredient lists are split into individual ingredient character strings at the points where a predefined character (in this case a comma) occurs. The ingredient strings are stored separately related to the food ID to which the ingredient belongs. All strings are stored in lower case. These strings, or parts of them, that correspond to numeric values, special characters, and specific words from a stopword list are then removed. Stopwords are filler words, adjectives and articles (e.g. with, a, the, big, etc.) as well as words that describe a processing form of an ingredient (e.g. cooked; for the entry “cooked egg”). Ingredient lists often contain additional explanations on quantity relationships using numerical or percentage values. After removal of stop words, double ingredients may appear in the tokenized table and will be deleted afterwards (Fig. 1).

food_id	ingredient_list
1015797	salt, vegetables 20% (onions 8,1%, carrots 6,4%, leek 2,4%, celery, roasted onions, parsnips, tomatoes), glucose syrup, corn, sunflower oil, concentrated vegetable juice (celery, carrots, onions), garlic, turmeric, lovage leaves, parsley, nutmeg, pepper, lovage root, ginger
…	…

⇨

food_id	ingredient
1015797	salt
1015797	vegetables
1015797	onions
1015797	carrots
1015797	leek
1015797	celery
1015797	onions
1015797	parsnips
…	…

⇨

food_id	ingredient
1015797	salt
1015797	vegetables
1015797	onions
1015797	carrots
1015797	leek
1015797	celery
1015797	parsnips
…	…

Fig. 1. Tokenization of the content list including stopword removal and removal of duplicates

Stemming generate the word stem of a certain word [18]. In order to be able to compare ingredients and FCDB Food names, the word stems of the ingredients are generated by using the snowball stemmer method [19]. Differently named ingredient words of the same meaning are linked language-spreading to a representative english term. For example, the North German word “Apfelsine” means “Orange” in Southern Germany. These two words are combined with the corresponding English translation “orange”. Thereby the library Thesarus [20] helps, which contains synonyms of many words. In addition, the Google Translate Library [21] is used to translate and a specially created library with synonymous food names. By linking the words of the same meaning and the stemmed words a new linking table is created (Fig. 2). Using this linking table, all individual ingredients can now be converted into a standardized main term. This will later help to identify the same ingredients in different ingredient lists.

main_term	synonym
orange	orange
orange	apfelsine
orange	orang
orange	apfelsin
...	...

Fig. 2. Linking of ingredients with synonyms and word stems

3.2 Method for Similarity Analysis Between Datasets

In order to achieve a higher information density of the food data, a scheme was developed that combines the information from the FCDB and the FPDB. To explain it, the following scenario is considered.

Scenario. A user of a health app, which provides information on food ingredients and nutritional values, uses the app's search function to search for the drink named "Cola" and enters this name in a search mask of the app. In the background, the app receives various results from the FDWH via the FIS, including the food named "Coladrink" from the FCDB and the two (fictitious) branded products named "Cola X" and "YZcoke". By selecting one of the displayed foods, the user can display its information on ingredients

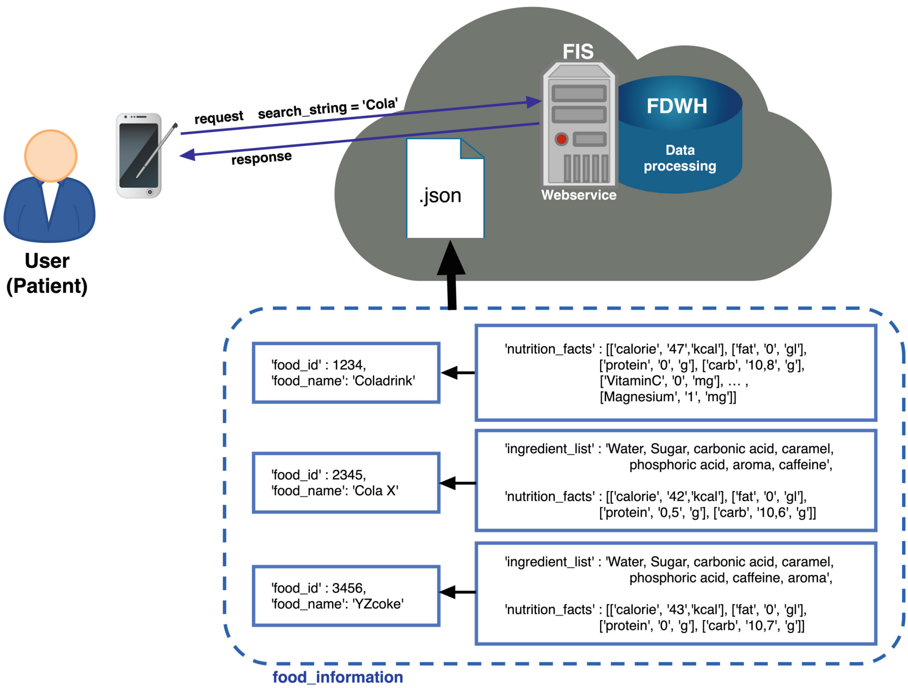

Fig. 3. FIS response (from food information from the FDWH) in JSON format, which matches to the search term "Cola"

and nutritional values. However, the food Coladrink from the FCDB does not provide any ingredients and the two foods from the FPDB, Cola X and YZcoke do not provide all Nutrition Facts that Coladrink provides (Fig. 3). It would be useful if the different information of similar foods would be combined and contained in each food dataset. For example, the information of the ingredient lists would then also be contained in the food data set of Coladrink and the information about vitamin C and magnesium (among others) would also be contained in the food data sets of Cola X and YZCoke (see Fig. 6). However, for this must be developed a method that recognizes similar foods on the basis of the given information.

In order to develop a scheme that links the information of such similar foods with each other, the attribute ranges presented in the following list are analyzed to identify similarly composed foods.

– Food names:
  - Parts of the food names are identical
– Ingredient lists:
  - products with the same or to a large extent the same ingredients
– Nutrition values:
  - The nutrition values are in the same range.

For the foods in the above scenario, the analysis of the food names shows that in the textstrings "Coladrink" and "Cola X" the same substring "Cola" is contained. The ingredient lists of Cola X and YZcoke are only different in the order of the last two ingredients (see Fig. 3). Furthermore, the distance between the nutritional values of energy, fat, proteins and carbohydrates among the foods is small, so that all three foods are similar in this attribute range (see Fig. 3).

With the help of the data mining method of decision tree generation [22], decision trees were developed on the basis of 1000 food test data (Fig. 4 and Fig. 5), with the help of which it is determined to what extent the individual ingredients and nutritional values may differ, so that the foods can be regarded as similar in these attribute ranges.

The two ingredient lists of a food pair, food X and food Y, are transmitted to the decision tree in Fig. 4. First, it will be checked if both ingredient lists contain more than two ingredients (the length of the lists must be greater than two). The first three ingredients of both lists must be identical. If the lists are greater than three elements and do not differ in the number of elements by more than three elements, all other elements (from the fourth to the last ingredient) are compared with each other. Finally, the food pair is classified as similar in the attribute range of the ingredient lists, if the number of elements of the ingredient in the larger of the two ingredient lists is less than 15 and the ingredients are contained in both ingredients lists except for a maximum of one element (the order does not matter). If the number of elements in the smaller of the two ingredient lists is greater than 15 elements, the ingredients must be contained in both ingredients lists up to a maximum of two elements.

The decision tree in Fig. 5 receives as input the nutritional values carbohydrates, fat, proteins (each in gram per 100 g) and energy (in kcal per 100 g) of food X and food Y. These values of the both foods are compared with each other. The foods are similar in the attribute range of the nutritional values if none of the compared nutritional value pairs differ by more than the value 10. The nutritional values for carbohydrates, fat, proteins and energy are the most widely contained nutrition facts in the food data sets.

For the similarity analysis of the food names, the individual words of the food names of two foods are tokenized and filtered with the help of stopword lists (as already described in Sect. 3.1). The words of both food names are then compared with each

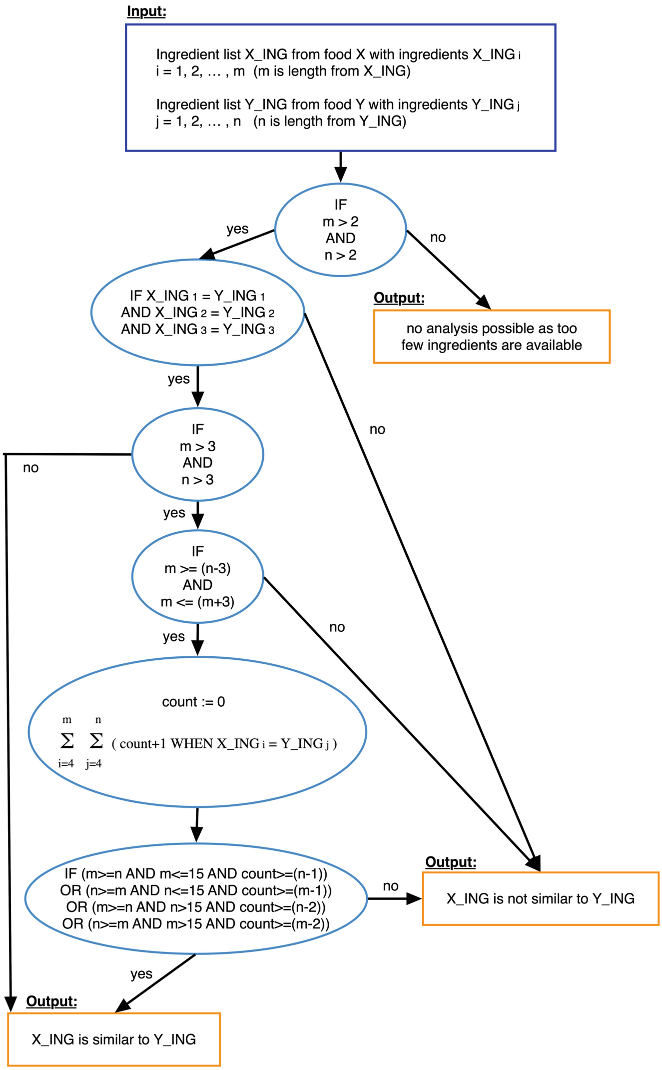

Fig. 4. Decision tree of the similarity analysis of ingredients in two food datasets

other. If at least two words in the two food names are identical, the food is classified as similar in the name attribute range.

If two foods are similar in at least two of the attribute ranges, their data sets are considered to be similar overall. This fact is true by the pairs Coladrink and X Cola as well as with X Cola and YZcoke (see Fig. 3). The similarity of X Cola to the other two foods results in a transitive dependency, so that the foods Coladrink and YZcoke are considered similar to each other. Finally, the following scheme is summarized with the help of which all food pairs of the FDWH can be analyzed for their similarity.

1. Similarity analysis of food names between all foods of the FPDB among themselves
2. Similarity analysis of the ingredients between all food products of the FPDB using the decision tree in Fig. 4
3. Similarity analysis of the nutritional values between all foods of the FPDB using the decision tree in Fig. 5
4. Similarity analysis of the food names between all foods of the FCDB among each other
5. Similarity analysis of the nutritional values between all foods of the FCDBs among themselves whose similarity of the food names is already confirmed by the analysis in 4. using the decision tree in Fig. 5
6. Similarity analysis of food names between FPDB and FCDB foods
7. Similarity analysis of nutritional values between all foods of FPDB and FCDB whose similarity of food names is confirmed by the analysis in 6. Already confirmed using the decision tree in Fig. 5
8. Determination of foods that were classified as similar in two of the similarity analyses in different attribute ranges
9. Determination of transitive dependencies between similar food pairs.

If no similarity is confirmed after the analysis according to 4. or 6., an analysis according to 5. or 7. is no longer necessary because at least two similarities must occur in different attribute ranges and the foods of the FCDB have no lists of ingredients and can therefore only be analyzed in two attribute ranges.

The IDs of the foods whose data sets were classified as similar overall are stored in a special database table linked with each other. The FIS is modified so that it obtains missing information of a food data set from a linked food data set and delivers it to the requesting app (Fig. 6).

In the above scenario, this means that when the user selects the entry Coladrink, the ingredients of the other two linked foods are also displayed. If he selects one of the other two foods, additional nutritional values of the FCDB data set Coladrink will be displayed which are not contained in the selected data set.

If a specific nutritional value of the linked data sets Y and Z (e.g. carbohydrate value of Y and carbohydrate value of Z) is to be displayed for a data set X, the mean value of the supplied information (mean value of carbohydrate value of Y and Z) is calculated. Ingredients are only obtained from linked data sets if the ingredient list of the selected food dataset is missing, so that no existing entry is falsified. For FCDB foods where ingredient information is missing, all ingredients from linked data sets are included. Now it can happen that this information does not correspond exactly to the food the user is looking for in the health app, but the user still has the important information which ingredients might be present in the food (this circumstance must be specially marked by the FIS or the displaying app). Now, for example, a nutritionist who looks at this information, in the case of a food allergy, can take a closer look at such a food data set if an occurrence of the allergen is apparent from the ingredient list (even if the occurrence was only vaguely determined) and the user of the app who suffers from the allergy can avoid eating the food for its own safety.

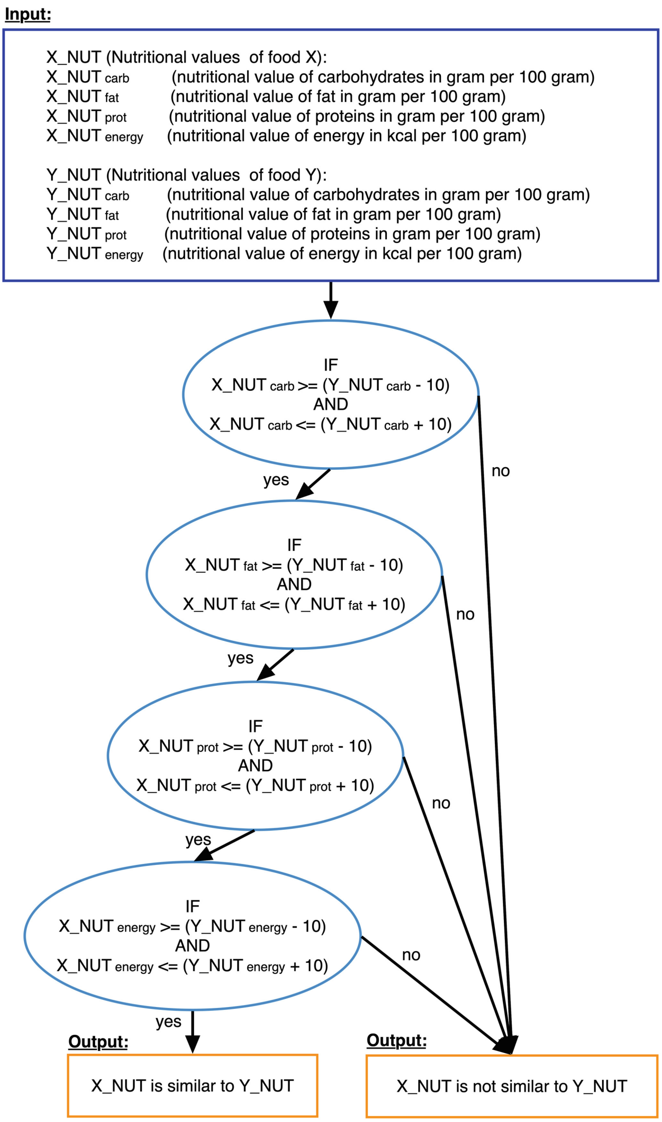

Fig. 5. Decision tree of the similarity analysis of nutrition values in two food datasets

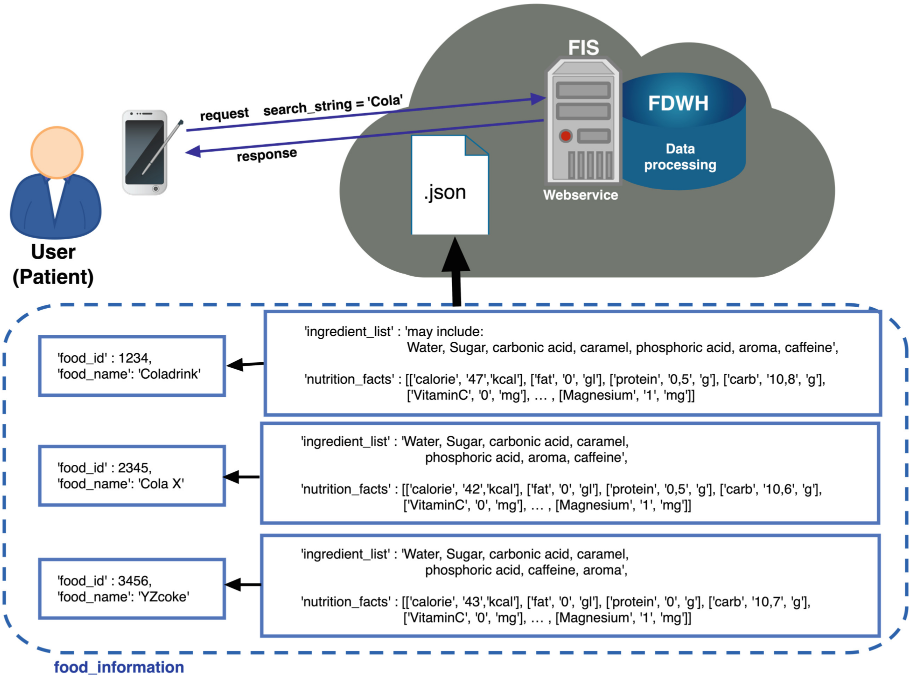

Fig. 6. FIS response which matches the search term "cola", with foods whose information has been completed with the help of similarity analysis

3.3 Using Big Data Technologies to Increase Performance

By comparing and mapping all data sets with each other using decision trees, conventional data analysis frameworks and tools quickly reach their performance limits. For example, the similarity analysis of the food names of 1000 food data sets already took more than 15 min. With an increased number of food data sets, the time taken by nested iterations during data processing increases exponentially. By using the Big Data Framework Spark of the Apache Foundation [23] it is possible to process the data sets very quickly with the help of special techniques (map, reduce, filter, text mining methods, etc.) and so-called lambda functions (anonymous functions without names that directly provide the return value) [24]. With Spark the data can be processed on different computing clusters directly in the main memory with the help of resilient distributed datasets (RDD) [23]. For example, RDDs can be represented and edited as data frames, which are known to developers of the programming language Python [25], which is often used in the data science field. Spark can be integrated into Python using the library PySpark. By operating on the data frames with the help of the PySpark library, the execution time is minimized many times over. Figure 7 shows a code snippet, as example for the similarity analysis of food names, in the programming language Python, using the PySpark library and using lambda functions. The food names were selected from a comma separated values file (csv file) and stored as RDDs. Further, the names were tokenized (splitted in single terms by space character) and stopwords were eliminated in each tokenized term.

The terms were compared to each other to determine similar name parts. The results, which name pairs are similar to each other, are written into a new csv file.

```
rdd = sc.textFile(
        ‚data.csv‘).map(lambda line :line.split(‚,‘))
data_frame = rdd.toDF(['x_food_id', 'y_food_id',
                        'x_food_name', 'y_food_name'])
tokenizer = Tokenizer(inputCol=‚x_food_name‘,
                      outputCol=‚x_vector‘)
data_frame = tokenizer.transform(data_frame)
tokenizer = Tokenizer(inputCol="y_food_name",
                      outputCol="y_vector")
data_frame = tokenizer.transform(data_frame)
remover = StopWordsRemover()
remover.loadDefaultStopWords('german')
remover.setInputCol("x_vector")
remover.setOutputCol("x_vector_without_stopw")
data_frame = remover.transform(data_frame)
remover.setInputCol("y_vector")
remover.setOutputCol("y_vector_without_stopw")
data_frame = remover.transform(data_frame)
differencer = udf(lambda x, y: list(set(x) - set(y)), Ar-
rayType(StringType()))
data_frame = data_frame.withColumn(
      'difference', differencer('x_vector_without_stopw',
      'y_vector_without_stopw'))
comparer = udf(lambda x, y, z: not len((set(x) - set(z)))
               == 0 and not len((set(y) - set(z))) == 0)
data_frame = data_frame.withColumn(
          'comparing', comparer('x_vector_without_stopw',
          'y_vector_without_stopw', 'difference'))
result_data_frame = data_frame.filter(
                     data_frame["comparing"] == True)
result_data_frame.toPandas().to_csv('new_data.csv')
```

Fig. 7. Code snippet as example for the similarity analysis of food names

4 Evaluation

To evaluate the similarity analysis, 1000 food test data, of which 200 dataset pairs had to be classified as similar, were evaluated by computer-based similarity analysis. The result shows that the analysis of the ingredients of 1000 foods has classified 249 dataset pairs as similar in this attribute range. The analysis of the nutritional values identified 207 similar dataset pairs and those of the Food Names 502 pairs. A total of 196 food data set pairs were classified as similar, including transitive dependencies. Only four data sets were not classified. This results in a recognition rate of 98% for the overall similarity analysis (Fig. 8). An incorrect similarity classification of dissimilar data sets did not take place.

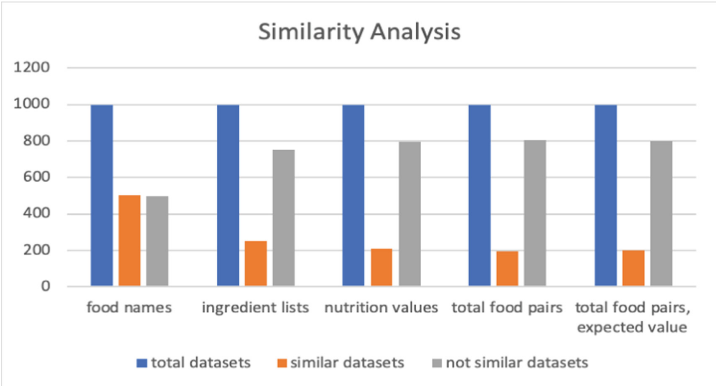

Fig. 8. Results of similarity analysis using test data

By using the Big Data Framework Spark and the additional Lambda functions used, the execution time of the similarity analyses was reduced by a factor of 90[1] (Fig. 9).

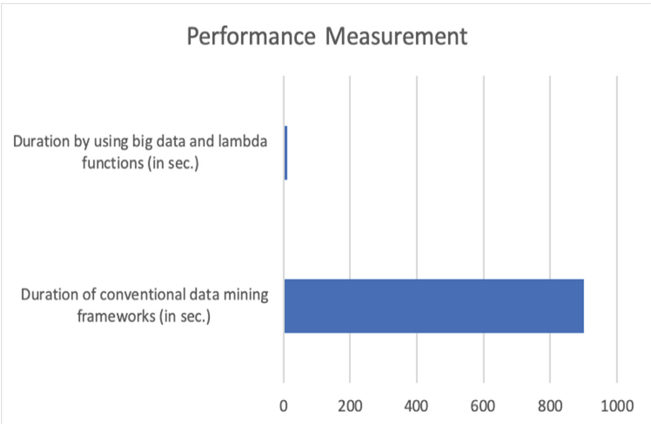

Fig. 9. Performance measurement before and after the use of big data technologies for the similarity analysis of food names

1 On a single cluster system with 4 × 2.5 gigahertz (GHz) processor and 16 gigabyte (GB) main memory.

5 Conclusion and Outlook

In order to deliver food data sets of sufficiently good data quality to e-health apps, the data sets of several data sources from both FCDBs and FPDBs were standardized and stored in a standardized data format in the FDWH. The FIS is used to deliver the data in the FDWH to the respective apps. Using data profiling methods, erroneous and duplicate data sets were detected and corrected or eliminated [1]. Data mining methods were used to link information from similar data sets in order to achieve a significant increase in information density. The evaluation with test data showed that the methods used achieved a high degree of recognition of the similarity analyses. By linking data sets through the similarity analysis, it can happen that ingredients and nutritional data are not determined 100% accurately. Nevertheless, these data can serve the experts in medical settings (nutritionists, physicians) as clues for a diagnosis [26]. The processing of food data records has shown that the time taken to process data increases exponentially with conventional data mining methods, as the amount of data records increases and thus the performance of the data processing system decreases, even if the 40,000 data records already stored are not yet classified as "Big Data". The number of food data records is constantly increasing as the amount of data in the data sources increases and new data sources are added. For these reasons, the use of Big Data technologies is unavoidable. This use has already resulted in a considerable increase in data processing performance.

In the further course of the DiDiER project, the quality improvement of the data sets is to be further promoted with the help of machine learning and the detection of inconsistencies. The aim is to achieve the greatest possible coverage of high-quality food information. In a study of the project, the methods used will be evaluated progressively.

References

1. Muenzberg, A., Sauer, J., Hein, A., Roesch, N.: The use of ETL and data profiling to integrate data and improve quality in food databases. In: 2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WI Mob), pp. 231–238 (2018). https://doi.org/10.1109/wimob.2018.8589081
2. Arens-Volland, A., Roesch, N., Feidert, F., Herbst, R., Moesges, R.: Change frequency of ingredient descriptions and free-of labels of food items concern food allergy sufferers. Allergy: Eur. J. Allergy Clin. Immunol. **65**, 394 (2010). https://doi.org/10.1111/j.1398-9995.2010.02393.x
3. Elfert, P., Eichelberg, M., Tröger, J., Britz, J., Alexandersson, J., et al.: DiDiER digitized services in dietry counselling for people with increased health risk related to malnutrition and food allergies. In: IEEE Symposium on Computers and Communications (ISCC), pp. 100–104. IEEE (2017)
4. Max Rubner Institue (MRI) Karlsruhe: Bundeslebensmittelschlüssel. https://www.mri.bund.de/de/service/datenbanken/bundeslebensmittelschluessel/. Accessed 10 May 2019
5. Schweizer Nährwertedatenbank. https://www.naehrwertdaten.ch/. Accessed 10 May 2019
6. WikiFood: Knowing what's inside, the Wiki for foodstuffs. http://www.wikifood.eu/wikifood/en/struts/welcome.do. Accessed 10 Mar 2019
7. snoopmedia GmbH, das ist drin: gemeinsam besser leben. http://das-ist-drin.de/. Accessed 12 May 2019
8. OpenFoodFacts – World. https://world.openfoodfacts.org. Accessed 12 May 2019
9. The Open Food Repo. https://www.foodrepo.org/. Accessed 12 May 2019

10. Danone GmbH. http://www.danone.de/. Accessed 12 May 2019
11. Nutricia GmbH. http://www.nutricia.de/. Accessed 12 May 2019
12. Biesalski, H.K., Pirlich, M., Bischoff, S., Weimann, A.: Ernährungsmedizin: Nach dem Curriculum Ernährungsmedizin der Bundesärztekammer, 5th edn. Thieme, Stuttgart (2017)
13. Upwork. https://www.upwork.com/hiring/development/intro-to-apis-what-is-an-api/. Accessed 12 May 2019
14. Eckert, C.: IT-Sicherheit: Konzepte – Verfahren – Protokolle, 7th edn. Oldenbourg Verlag, München (2012)
15. Ourcodeworld. https://ourcodeworld.com/articles/read/249/implementation-of-the-soundex-algorithm-function-in-different-programming-languages. Accessed 12 May 2019
16. Friesen, J.: Java XML and JSON, 1st edn. Apress, New York (2016)
17. Hochschule Augsburg: Glossar. https://glossar.hs-augsburg.de/Ontologie. Accessed 12 May 2019
18. Hotho, A., Nuernberger, A., Paaß, G.: A brief survey of text mining. LDV Forum - GLDV J. Comput. Linguist. Lang. Technol. **20**, 19–62 (2005)
19. Korenius, T., Laurikkala, J., Järvelin, K., Juhola, M.: Stemming and lemmatization in the clustering of finnish text documents. In: Proceedings of the Thirteenth ACM International Conference on Information and knowledge management (CIKM 2004), pp. 625–633. ACM, New York (2004). http://dx.doi.org/10.1145/1031171.1031285
20. Thesarus. https://www.thesaurus.com/. last accessed 2019/05/12
21. Google Cloud. https://cloud.google.com/translate/docs/. Accessed 12 May 2019
22. Cleve, J., Lämmel, U.: Data Mining, 1st edn. De Gruyter Verlag, Berlin (2014)
23. Apache Spark. https://spark.apache.org/. Accessed 12 May 2019
24. Python-Kurs. https://www.python-kurs.eu/lambda.php. Accessed 12 May 2019
25. Python. https://www.python.org/. Accessed 12 May 2019
26. Roesch, N.: Der Einsatz von Informations - und Kommunikationstechnologie bei Nahrungsmittelallergie. Shaker Verlag, Aachen (2010)