Nutrition Data Analysis and Visualization for Deterministic Food Categorization

Martin Tiefengrabner, Simon Ginzinger

University of Applied Sciences, Salzburg
Department of MultiMediaTechnology

simon.ginzinger@fh-salzburg.ac.at

Abstract

In the treatment of non-communicable diseases such as diabetes or cardiovascular diseases logs of the patients' food intake are important information. The assessment of the nutritional values of meals is a complicated task because of the wide variety in different foods – due to the variation in ingredients and serving sizes. Serving sizes can be measured or estimated. However, using scales or measuring cups in everyday life can be tedious for the patients. Visual estimation is preferred to simplify the determination of food quantity. Therefore food groups with similar energy densities per volume have to be defined. We present a method for building food groups based on their volumetric carbohydrate and kilocalorie densities. The underlying food properties were taken from the USDA National Nutrient Database for Standard Reference (USDA-NNDSR). Carbohydrate and kilocalorie densities where computed for every food. For the categorization of foods OPTICS, a density-based clustering algorithm was applied on the dataset. The algorithm computes reachability-distances, which represent the different densities in the structure of the data and are the basic information for detecting clusters. The resulting reachability plot can be analyzed in an interactive visualization tool to investigate the structure of the food data. Clusters in the data represent food categories.

1 Introduction

Non-communicable diseases are a major concern of the World Health Organization (WHO). The most important risk factors are high blood pressure, high concentrations of cholesterol in the blood, inadequate intake of fruit and vegetables, overweight or obesity, physical inactivity and tobacco use (cf. WHO 2004). Four out of theses six risk factors are directly related to people's nutrition.

Self-awareness about nutritional values of food intake is the first step to a healthier diet. However, measuring nutritional values of food intake can be tedious and also error-prone. Categorizations of food into groups with similar nutritional values have proven effective in achieving a healthier diet (cf. Lowe et al. 1999; Dansinger et al. 2004). Nevertheless, food assessment is mostly done based on weight (if not possible from standardized portion sizes). This is in contrary to the human perception, which first of all gives information about the volume of the different foods on one's plate. Therefore a categorization based on energy per volume is expected to significantly increase the accuracy of individual estimations.

In this work data from the USDA National Nutrient Database for Standard Reference (cf. US Department of Agriculture 2014) is analyzed. Carbohydrates and kilocalories per volume are calculated for all foods where volume information is available. The resulting 1718 foods were clustered with respect to this data. The clustering was refined by comparing the textual description of different foods resulting in a semantically reasonable ordering of foods with highly similar energy values. As no single optimal categorization for every potential application exists, a tool for analysis and visualization of the consolidated data was designed. This tool is freely available to the scientific community.

2 Methods

The data in the USDA-NNDSR is collected by the U.S. Department of Agriculture, Agricultural Research Service (cf. US Department of Agriculture 2014). In its latest release (27) it contains data for 8618 different foods. The analysis presented here relies on nutritional information per volume in contrast to nutritional information given per weight. Therefore, all foods measured in cups were extracted from the database. The manual of the USDA-NNDSR gives no information about which definition concerning the metric volume of cups where used. Therefore, the definition in the Electronic Code of Federal Regulations (e-CFR) was used which defines one cup as 240 ml (cf. Electronic Code of Federal Regulations 2014). In a next step the metric volume of all extracted foods where calculated and the information about kilocalories and carbohydrates was translated from weight-based to volume-based values.

$$E_{ml} = \frac{E_{100g}}{100} \bullet \frac{weight_{cup}}{240}$$

Equation 1 Translation of weight-based to volume-base values

After the conversations 1718 foods remained for further processing. Carbohydrates where translated to kilocalories reflecting industry practices (cf. US Department of Agriculture 2014). All the mentioned translations and computations where done in MySQL. This resulted in a dataset containing textual descriptions, kilocalories and carbohydrates per volume of the 1718 different foods.

Data was exported to "R" for further processing. R is an environment and programing language for statistical computing based on "S" (cf. R Core Team 2014). The Euclidean distance between every pair of foods (d_{eucl}) was calculated using the kilocalories per volume as the x-coordinate and the carbohydrates per volume as the y-coordinate. To be able to refine any grouping using the semantic information in the textual descriptions, the short descriptions in the USDA-NNDSR were split in chunks; see Figure 1 for an example.

MARGARINE-LIKE,VEG OIL SPRD,60% FAT,STICK/TUB/BOTTLE,W/ SALT

⇩

MARGA-RINE-LIKE	VEG OIL SPRD	60% FAT	STICK/TUB/ BOTTLE	W/ SALT

Figure 1 Example for splitting the short description

The similarity of two foods in terms of descriptions was then measured by summing-up the number of identical chunks included in both food descriptions and multiplied by two and divided by the total sum of chunks in both descriptions. The similarity measure is translated into a distance measure (d_{text}) by inversion.

To include this semantic information into the distance measure the following formula is used:

$$d(a,b) = d_{eucl}(a,b) * d_{text}(a,b) * w_{text} + d_{eucl}(a,b) * w_{eucl}$$

Equation 2 Translation of weight-based to volume-base values

This results in a dataset containing all pairwise distances in terms of descriptions, carbohydrates and kilocalories. The dataset was generated eleven times using the following weighting combinations:

Table 1: Distance weightings used

w_{text}	0	0.1	0.2	0.3	0.4	0.5	0.4	0.3	0.2	0.1	0
w_{eucl}	1	0.9	0.8	0.7	0.6	0.5	0.6	0.7	0.8	0.9	1

The distance matrices are used as input for the clustering algorithm OPTICS. OPTICS is an acronym for "Ordering points to identify the clustering structure" and describes a density-based clustering algorithm (cf. Ankerst et al. 1999). In this work the implementation of OPTICS in the "Environment for Developing KDD-Applications" (ELKI) was used (cf. Achtert, Kriegel & Zimek 2014). Usually the raw dataset is used as an input for OPTICS and the pairwise distances are computed as part of the algorithm. In our case the precomputed distance were used as an input. As a result of OPTICS the reachability distances for all entries in the given dataset are computed which can later be visualized as a so-called *Reachability Plot*.

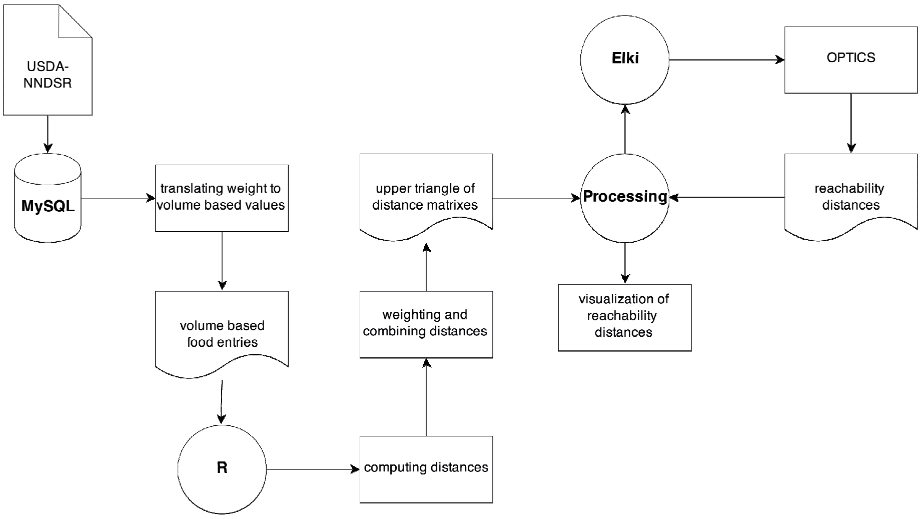
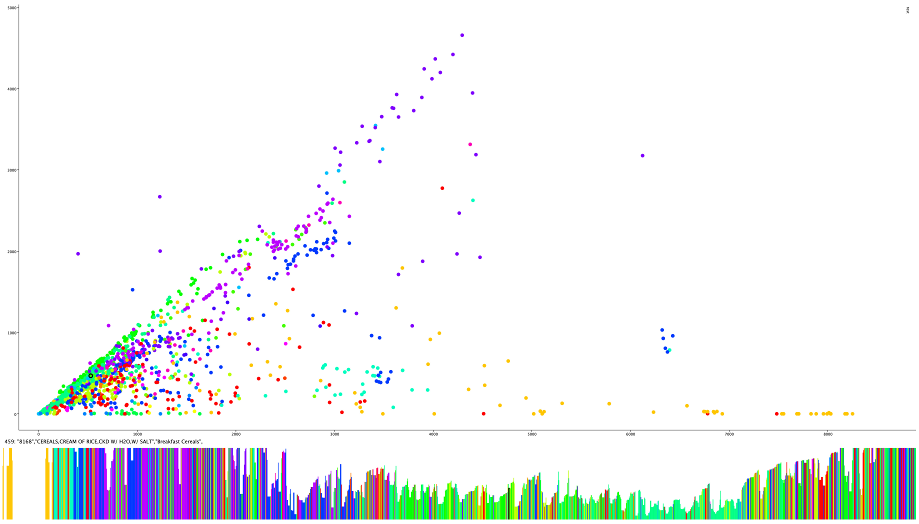

USDA-NNDSR
MySQL
translating weight to volume based values
volume based food entries
R
computing distances
weighting and combining distances
upper triangle of distance matrixes
Elki
OPTICS
Processing
reachability distances
visualization of reachability distances

Figure 2 Flowchart of the data processing

To create an interactive visualization that helps in finding clusters and groupings in the results of OPTICS a Processing application was written. Processing is a Java-based programming language and development environment for scientific and artistic purposes developed at the MIT.[1] The application was combined with the ELKI java package such that precomputed distances matrices can be selected in the Processing application and used as input for OPTICS. The results of OPTICS, the reachability distances, are then sent back to Processing and visualized. Therefore the application consists of two main parts, a scatterplot of the carbohydrates and kilocalories of the foods and the reachability plot (see Figure 3). To automatically find clusters/groupings in the foods, a self-developed algorithm was implemented which finds clusters based on the different steepness in the reachability plot. The threshold of the steepness of the clusters to be found can be set in the graphical user interface of the application. By changing the threshold, clusters of different densities can be found. For further investigations the reachability plot gives the possibility to pan and zoom and select entries to get further information of the particular foods. Besides we tried to give further information of the food groups in the USDA-NNDSR by coloring the plots.



459: "8168","CEREALS,CREAM OF RICE,CKD W/ H2O,W/ SALT","Breakfast Cereals",

Figure 3 The Visualization Tool: In the upper part a scatter plot of carbohydrate per volume vs. kilocalories per volume is shown. In the lower part the reachability plot of the OPTICS clustering algorithm is shown.

[1] http://www.processing.org/

3 Results

The main result of this work is a Processing package, which provides the user with the ability to browse the nutritional data and identify groups of foods with similar energy to volume ratios (shown as valleys in the reachability plot in Figure 3). The user may switch through the different distance datasets by using the up- and down-keys. Coloring corresponds to food groups defined in the USDA-NNDSR.

```
findClusters(reachabilityDistances, steepnessThreshold)
   clusterOpened = FALSE
   currentClusterNumber = -1
   lastPosIdx = -1
   reachedStepness = FALSE

   for (i in 1 to (reachabilityDistances.length - 2)):
       curDist = reachabilityDistances[i]
       nextDist = reachabilityDistances[i+1]
       nextNextDist = reachabilityDistances[i+2]
       if(clusterOpened):
          setCluster(i,currentClusterNumber)
       else:
          distDelta1 = nextDist - curDist
          distDelta2 = nextNextDist - nextDist
          deltaDeltaDist = delta1 - delta2
          if(distDelta1 > 0 AND ABS(deltaDeltaDist) > steepnessThreshold):
              reachedStepness = FALSE
              clusterStartDist = nextDist
              clusterOpened = TRUE
              currentClusterNumber++
       if(clusterOpened):
          if(nextDist > clusterStartDist):
              clusterOpened = FALSE
              lastPosIDX = -1
          else:
              if(nextDist > curDist):
                       lastPosIDX = i
                       reachedStepness = true
              if(reachedStepness AND (nextDist < curDist))
                       if(lastPosIDX NOT -1):
                             for(j in lastPosIDX to i)
                                 setCluster(i, -1)
                             clusterOpened = FALSE
                             lastPosIDX = -1
    if(lastPosIDX NOT -1)
       for(j in lastPosIDX to reachabilityDistances.length)
          setCluster(j,-1)
```

Listing 1 Algorithm to detect clusters based on steepness

The coloring may be switched to color the automatically detected clusters using our cluster detection algorithm (see Listing 1). Mouse-over events on both the scatter as well as the reachability plot show the corresponding food description and select the equivalent data point/bar in the other plot.

To evaluate if the inclusion of semantic information (d_{text}) yields an improved grouping we calculated the number of interruptions in the reachability plot for each food group from the USDA-NNDSR. Apparently the inclusion of the semantic information results only in a minor effect (see Figure 4).

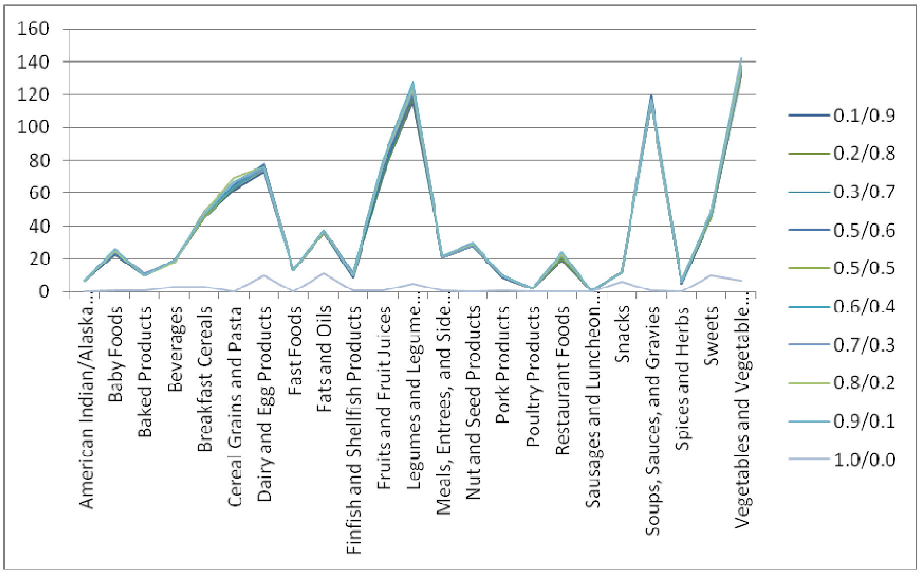


Figure 4 Interruptions of USDA-NNDSR food categories in the reachability plot for different distance datasets

4 Availability

The data visualizer is available via www.smarthealth.at/fmt-2014-food-visualizer.

5 Discussion

In this work we present an approach to used energy per volume ratios of foods for food categorization. We refer to the most prominent database for nutritional information and give users a new way to analyze the data. We are convinced that new categorizations of food based on our analysis tool will prove useful in practice. Future work includes testing the new categorizations in a practical setting (e.g. patients suffering from non-communicable diseases).

Funding source: This work was funded by the Austrian research promotion agency (FFG), project number 839076.

References

Achtert, E.; Kriegel, H. P. & Zimek, A. (2008): ELKI: a software system for evaluation of subspace clustering algorithms. In: *Scientific and Statistical Database Management*. Berlin/Heidelberg: Springer, pp. 580–585.

Ankerst, M.; Breunig, M. M.; Kriegel, H. P. & Sander, J. (1999): Optics: Ordering points to identify the clustering structure. In: *ACM Sigmod Record* 28 (2), 49–60.

Dansinger, M. L.; Gleason, J. A.; Griffith, J. L.; Selker, H. P. & Schaefer, E. J. (2005): Comparison of the Atkins, Ornish, Weight Watchers, and Zone diets for weight loss and heart disease risk reduction: a randomized trial. In: *Journal of the American Medical Association* 293 (1), 43–53.

Electronic Code of Federal Regulations (2014): Title 21: Food and Drugs, Part 101 – Food Labeling, Subpart A-General Provisions, §101.9 Nutrition labeling of food: http://www.ecfr.gov/cgi-bin/retrieveECFR?gp=1&SID=37b578dfb011937448ffadb559118b6f&ty=HTML&h=L&r=SECTION&n=se21.2.101_19.

Lowe, M. R.; Miller-Kovach, K.; Frye, N. & Phelan, S. (1999): An Initial Evaluation of a Commercial Weight Loss Program: Short-Term Effects on Weight, Eating Behavior, and Mood. In: *Obesity Research* 7 (1), 51–59.

R Core Team (2014): R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing: http://www.R-project.org.

US Department of Agriculture (2014): Composition of Foods Raw, Processed, Prepared, USDA National Nutrient Database for Standard Reference, Release 27, Documentation and User Guide: http://www.ars.usda.gov/sp2UserFiles/Place/12354500/Data/SR27/sr27_doc.pdf.

World Health Organization (WHO) (2004): Global Strategy on Diet, Physical Activity and Health: http://apps.who.int/iris/bitstream/10665/43035/1/9241592222_eng.pdf?ua=1.