Trends in Food Science & Technology 49 (2016) 1–13

Contents lists available at ScienceDirect

Trends in Food Science & Technology

journal homepage: http://www.journals.elsevier.com/trends-in-food-science-and-technology

Review

Data-driven recipe completion using machine learning methods

Marlies De Clercq*, Michiel Stock, Bernard De Baets, Willem Waegeman

KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, 9000, Ghent, Belgium

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form
8 September 2015
Accepted 30 November 2015
Available online 15 December 2015

Keywords:
Ingredient combinations
Recipe completion
Non-negative matrix factorization
Two-step regularized least squares
Recommender systems

ABSTRACT

Background: Completing recipes is a non-trivial task, as the success of ingredient combinations depends on a multitude of factors such as taste, smell and texture.
Scope and approach: In this article, we illustrate that machine learning methods can be applied for this purpose. Non-negative matrix factorization and two-step regularized least squares are presented as two alternative methods and their ability to build models to complete recipes is evaluated. The former method exploits information captured in existing recipes to complete a recipe, while the latter one is able to also incorporate information on flavor profiles of ingredients. The performance of the resulting models is evaluated on real-life data.
Key findings and conclusions: The two machine learning methods can be used to build models to complete a recipe. Both models are able to retrieve an eliminated ingredient from a recipe and the two-step RLS model is also capable of completing an ingredient set to create a complete recipe. By applying machine learning methods on existing recipes, it is not necessary to model the complexity of good ingredient combinations to be able to complete a recipe.

1. Introduction

Due to the increasing availability of data in online databases, data mining and machine learning methods are starting to play a prominent role in food consumption analytics and food preference modelling. Traditional methods such as clustering and classification methods are commonly applied to identify consumer patterns and consumer preferences – see e.g. Murray and Delahunty (2000); Westad, Hersleth, and Lea (2004); Bahamonde, Díez, Quevedo, Luaces, and del Coz (2007). More recent studies have introduced the use of more sophisticated methods, such as matrix factorization techniques and graph-based analysis tools, in order to identify consumer patterns based on surveys and check the theory of food pairing, amongst others – see. e.g. Guimet, Boqué, and Ferré (2006); Young, Fogel, and Hawkins (2006); Ahn, Ahnert, Bagrow, and Barabasi (2011); Zetlaoui, Feinberg, Verger, and Clémençon (2011). Moreover, the introduction of specialized machine learning methods in food consumer analytics has resulted in the development of a number of recipe recommender systems.

For example, Sobecki, Babiak, and Slanina (2006) developed a hybrid web-based recommender system, which can recommend certain recipes to different types of users. The system uses fuzzy reasoning techniques and bases its predictions on information contained in user profiles. Freyne and Berkovsky (2010) designed a personalized recipe recommender system that returns healthy recipes, with the aim of educating users in adopting and maintaining a healthier lifestyle. They compose recommendations by gathering ratings on individual ingredients, and combining those ratings into recipe scores, such that unknown ingredient and recipe scores can be predicted for a particular user. Forbes and Zhu (2011) showed that matrix factorization methods can yield a satisfactory predictive power for recipe recommendation, by connecting users with similar taste. The model follows a two-step approach and needs two inputs: a matrix containing the scores given to each recipe and a binary matrix containing the recipes and their ingredients. As a last example, van Pinxteren, Geleijnse, and Kamsteeg (2011) designed a recipe recommender system with the aim to deliver healthier recipes that fit into the daily routine of users. The authors constructed a measure that determines the similarity between recipes following a user-centered approach.

In the present article, we tackle the related problem of designing a recommender system for preparing a meal using some leftover ingredients in a refrigerator. One often has the feeling that by adding only one or two additional ingredients, a splendid dish

* Corresponding author.
E-mail addresses: mm.declercq@ugent.be (M. De Clercq), michiel.stock@ugent.be (M. Stock), bernard.debaets@ugent.be (B. De Baets), willem.waegeman@ugent.be (W. Waegeman).

http://dx.doi.org/10.1016/j.tifs.2015.11.010

could be created. The problem is deciding which ingredients to add to the grocery list to complete this dish. Solving this problem is a non-trivial task, since there is a multitude of factors that influence the flavor perception of an ingredient combination. The two most obvious factors are smell (Small and Prescott (2005); Smith and Margolskee (2006)) and taste (Prescott (2004); Stevenson and Boakes (2004); Moyer (2013)). Another factor is texture. The sweetness of honey, for instance, is determined by its viscosity (Cook, Hollowood, Linforth, and Taylor (2003); Auvray and Spence (2008)). A well-known example of how temperature can influence the flavor perception is the bitter taste of warm beer (Talavera, Ninomiya, Winkel, Voets, and Nilius (2007); Bakalar (2012); Bajec, Pickering, and DeCourville (2012)). A fifth influencing factor is color. Studies have shown that a more intense color leads to a higher flavor intensity (Morrot, Brochet, and Dubourdieu (2001); Zampini, Sanabria, Phillips, and Spence (2007)). Also the sound of foodstuffs can have an influence. Two examples are the crispness of potato chips and the snap of chocolate (Auvray and Spence (2008); Stevenson (2009)). These factors do not only influence the flavor perception, but also influence each other.

In spite of the difficulty of the task, already several solutions to the problem are available. A first solution is to look in a cook book or database for recipes that contain the remaining foodstuffs. As this activity can be quite time consuming, online search engines exist for retrieving such information automatically. These engines are connected to online databases that contain a large number of recipes. When given some ingredient names, they explore their database for recipes containing one or several of the given ingredients mentioned by the user. Some examples of search engines that work in that way are *supercook.com*, *myfridgefood.com* and *recipematcher.com*.

A second solution to find the desired additional ingredients is to use models for ingredient combinations. One example is the online model of Foodpairing® that suggests, for one ingredient, those ingredients that create tasteful combinations with it. The model is based on the theory of food pairing, which states that a good ingredient combination can be achieved when the combined ingredients have the same major flavor components. The model is created based on scientific flavor analysis of foodstuffs. Once all the flavor components of a foodstuff are analyzed, the major flavor components are identified. These are the components that will determine the smell of a foodstuff. These components are taken into account in the Foodpairing® model (Sense for Taste (2014)). The output of the model is visualized as a tree, with the ingredients of interest in the middle and the ingredients with similar flavors arranged around it. The closer an ingredient is to the center of the tree, the higher the number of shared major flavor components.

The above methodologies have a number of disadvantages. Existing methods based on searching in cook books and online databases are restricted to recommending existing recipes, whereas methods based on flavor components are likely to recommend ingredients with very similar flavors to one of the ingredients present and do not take into account other ingredients that are present in the recipe. As such, neither one of these two approaches can answer recipe completion questions such as "which type of meat can best be combined with all remaining vegetables" or "which herbs can best be used to spice up a meal". We present two new methods to build models that can address questions of that kind. The resulting models return, for a given set of ingredients, those ingredients that can best be combined with all the given ingredients.

Our solutions are based on data mining and machine learning methods. The first method we investigate is non-negative matrix factorization (Lee and Seung (1999)). This method will be limited to finding ingredient combinations that are already present in existing recipes. More precisely, the model using this method will not be able to find recipes that are not present in the data set, but it will focus on ingredient combinations that are often occurring in existing recipes. As a solution that is able to overcome this problem, two-step regularized least squares is introduced as a second machine learning method used to build a model for recipe completion. This model will not only suggest ingredient combinations based on those found in existing recipes, but also based on the presence of similar flavor components in the different ingredients. This second model will allow to suggest *new* ingredient combinations to complete a recipe.

We present a thorough evaluation and extensive experimental results that confirm the potential of our models. To train, validate and test the models, good data is required. The problem today is not the lack of data (recipe databases, flavor databases, consumption data, food composition data, etc.), but the fact that there is no clear vision on what can be learned from all this data. Next to this problem, there is the fact that not all the available data is equally reliable and useful, while some databases are incomplete or inaccurate. Unfortunately, a data-driven approach is only as good as the data used to learn from. For example, Varshney, Varshney, Wang, and Myers (2013) showed in a study on food pairing that two data sets containing the same type of information can result in conflicting conclusions. This problem could be solved by standardizing databases and by improving the quality of (existing) databases (e.g. Roe and Finglas, (2012)). In our experimental study, we have used the well-known and extensive data set of Ahn et al. (2011), containing data files about recipes, ingredients and flavor components.

The present article is organized as follows. First the data is described and appropriately transformed. In the next section the structure of the models is presented, after which non-negative matrix factorization and two-step regularized least squares are discussed in more detail. The performances of the two models are presented in a following section. Finally some conclusions and thoughts are formulated at the end of the article.

2. Materials

As mentioned before, the data used in the present article are taken from Ahn et al. (2011). This data set consists of five data files covering recipes, ingredients and flavor components. A first data file contains 56,498 recipes originating from eleven different cuisines. For each recipe the accompanying ingredients are listed, as well as the cuisine where the dish originates from. All these recipes together contain 381 different ingredients, ranging from almond to zucchini. Recipes containing less than three ingredients are eliminated in the present article, as these recipes contain no information on ingredient combinations necessary to complete (new) recipes. This reduces the number of recipes to 55,001. A second data file contains the names of 1,530 ingredients, including the 381 ingredients found in the first data file. Each ingredient is uniquely classified into an ingredient category (e.g. fruit, meat, vegetables, etc.). The third data file gives an overview of 1,107 flavor components found in foodstuffs. Each component is provided with its own chemical abstracts service (CAS) registry number. CAS numbers can be used to gather more information about a flavor component, like the chemical structure, synonyms, etc. A fourth data file connects each ingredient with its flavor components. From this data file, it can be concluded that the 381 ingredients, found in the recipe data, contain only 1,021 flavor components. Only these flavor components are hence taken into account. The fifth and final data file contains for 221,777 ingredient pairs, formed with the 1,530 ingredients from the second data file, the number of flavor components the two ingredients have in common. This data file is not