Ingredients matching in bakery products

Tome Eftimov [1,2], and Barbara Koroušič Seljak [1]
[1] Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000, Ljubljana, Slovenia
{tome.eftimov, barbara.korousic}@ijs.si

ABSTRACT
In this paper, we present the analytical results of the ingredients matching in bakery products. We collected recipes from a free recipes web site and the main goal was to find association rules between the recipes' ingredients. For this purpose we applied an *Apriori algorithm* and various visualization techniques to represent the discovered association rules. The paper covers: data extraction, data preprocessing, association rules and visualization of the results during this work.

Keywords
association rules, text mining, ingredients matching

1. INTRODUCTION
The aim of the analysis presented in this paper was to find potentially interesting and relevant relations between the recipes' ingredients.
As our target data, we selected bakery recipes in English and focused on exploring relations between ingredients that occur in the bakery recipes.
First, we collected the data from a free Internet data source [1]. Afterwards, we preprocessed it in the form needed for the analysis. Then we looked for association rules and finished by representing discovered results and possible future work.

2. DATA
The data we used is a collection of 1,900 bakery recipes written in English, and we collected it using HTML parser to extract the information from a free recipes web site [1].
We considered the names of the ingredients for each recipe, while the quantity-unit pair associated with the ingredient was ignored as our goal was analysing only the relations between the ingredients.
Before the analysis, we preprocessed our target data. Because the data contained many adjectives that are associated with the cooking process (e.g. sliced, mashed), we removed them. We also located synonyms that appear in the data (e.g. pumpkin puree, pumpkin) and mapped them in the form required for the analysis. After cleaning the data, the preprocessed data was transformed into a document-text matrix and after that into a transactional matrix that is the form needed for our analysis. At the end, our transformed data contained 1,900 transactions (rows) and for each transaction we needed to consider the presence of 542 ingredients (columns).

For the cleaning process and the mapping of the synonyms we applied some regular expressions using the R programming language. The summary of the basic statistics of our data shows that the data set is rather sparse with a density just above 1.65%. The ingredient "salt" is the most popular and the average transaction contains less than 9 ingredients.

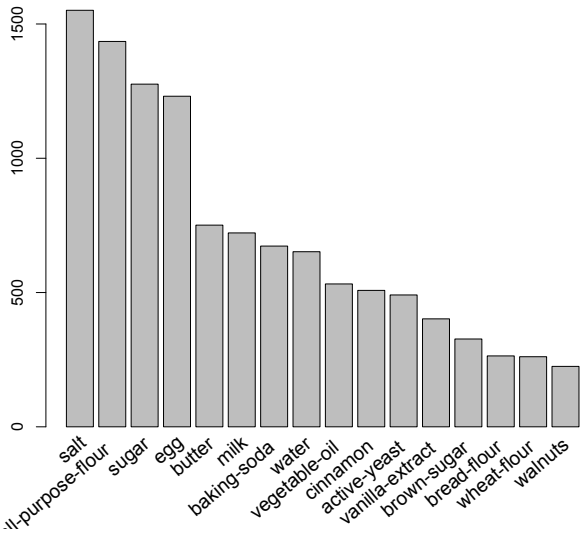

Figure 1: The most frequently used ingredients

In Figure 1, we can see that the ingredients "all-purpose flour", "egg", "salt" and "sugar" are most frequently used and because the probability of the presence of these ingredients in a bread recipe is very high, we rejected them for the analysis and focused upon the relations between other ingredients. After excluding the above mentioned most frequently used ingredients, our data set contained 1,900 transactions, each having 538 ingredients. The data set is rather sparse with density just above 1.13% and the average transaction contains less than 7 ingredients.

3. METHODS
Finding potentially interesting and relevant relations between the ingredients in bakery products is a task of the descriptive data mining method, known as the association rules mining [6]. In our case, having an association rule

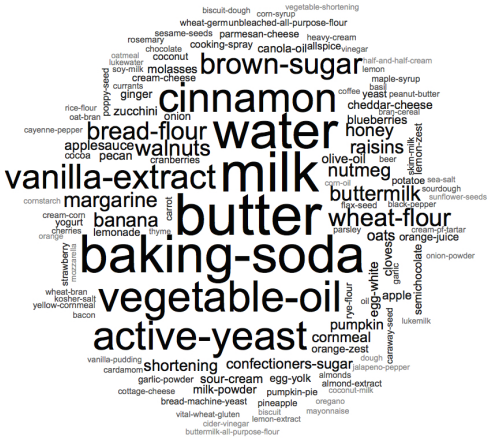

Figure 2: A wordcloud of the ingredients

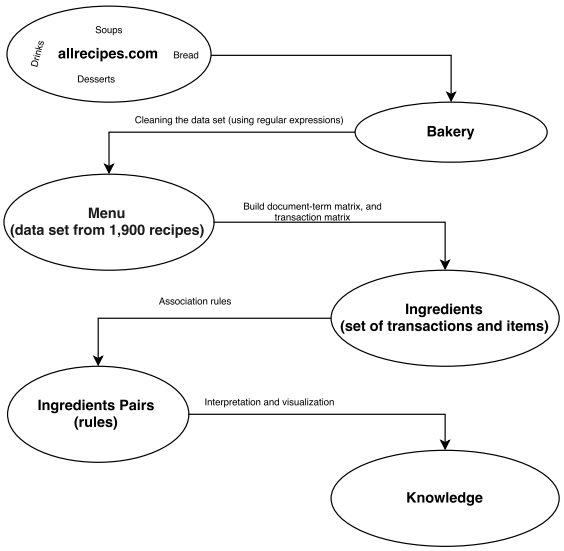

Figure 3: The knowledge discovery process

X→Y, where X and Y are sets of ingredients, the intuitive meaning of such a rule is that a recipe that contains all ingredients from X also tends to contain all ingredients from Y. The sets of ingredients X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule, respectively.

Because usually the number of such rules is huge, the space of all possible association rules needs to be reduced and for this purpose two criteria are used, support and confidence of the association rule.

Support of an association rule is the ratio of the number of recipes that have true values for all ingredients in X and Y and the number of recipes in our database. The confidence is the ratio between the number of recipes that have true values for all ingredients in X and Y and the number of recipes that have true values for all ingredients in X. Another measure that we used is a lift which tells us how many more times the ingredients in X and Y occur together then it would be expected if the sets of ingredients (X and Y) were statistically independent.

The whole knowledge discovery process is represented in Figure 3.

4. EVALUATION

There are several association rules algorithms and in our analysis we used the basic algorithm known as Apriori [3] and its implementation from the package "arules" in R [5].

After we imported the data into R, we used the Apriori algorithm to find the association rules and we tried it out for different values of the minimum support and minimum confidence. At the end, we decided to fix the support on 0.005, which means that at minimum 10 recipes will contain the ingredient and the confidence on 0.75. The number of discovered rules using these parameters is 1,235. Because some rules are redundant, which provide little or no extra information when some other rules are in the result, we pruned them and at the end we have 594 rules. The top 15 rules with respect to the lift measure are given in Table 1.

Because the number of the discovered association rules is huge and it is not recommended to go through all of them, we used some visualization techniques, which are implemented in the R's package "arulesViz" [4]. For visualization of our result we used graph-based visualization, parallel coordinates plots and grouped matrix-based visualization.

In Figure 4, we present the graph-based visualization with ingredients and rules as vertices for our top 10 rules with respect to the lift measure. Here the rules are the vertices, the size of the vertex is the support of the rule, while the color of the vertex is the lift of the rule. We can see how the rules are composed of individual ingredients and how they share ingredients. For example, we can see that if the recipe contains "garlic powder" and "milk" also tends to contain "cheddar-cheese". The graph-based visualization is an efficient technique to represent analytical results to people who are unfamiliar with data mining as from the graph they can see the relation between ingredients.

Another visualization suitable for people without knowledge on data mining is the parallel coordinate plot. In Figure 6, we present the parallel coordinate plot of our top 30 rules with respect to the lift. The width of the arrows gives the support and the intensity of the color presents the confidence. On the x-axis are represented the position in the rule, i.e., first ingredient, second ingredient, etc., while the arrow is used for the consequent.

In Figure 7, we have presented the grouped matrix-based visualization using a ballon plot with antecedent groups as columns and consequents as rows. The color of the ballon is the aggregated lift in the group, while the size of the ballon is the aggregated support. The aggregated lift is decreasing top down and from left to right, and the most interesting group is on the top left corner. The group of most interesting rules contains 5 rules, which contain "caraway seed" and 3 other ingredients in the antecedent and "rye flour" in the consequent. Another interesting group contains 2 rules,

	LHS	RHS	support	confidence	lift
1	{bread-flour, caraway-seed}	{rye-flour}	0.006	0.928	45238
2	{active-yeast, caraway-seed}	{rye-flour}	0.008	0.888	43304
3	{caraway-seed, water}	{rye-flour}	0.008	0.888	43304
4	{cranberries, orange-juice}	{orange-zest}	0.005	0.846	30333
5	{orange-juice, walnuts}	{orange-zest}	0.005	0.833	29874
6	{baking-soda, cinnamon, molasses}	{ginger}	0.006	0.800	24126
7	{garlic-powder, milk}	{cheddar-cheese}	0.005	0.769	21181
8	{cream-cheese, milk, vanilla-extract}	{confectioners-sugar}	0.005	0.909	16142
9	{baking-soda, cinnamon, nutmeg, water}	{pumpkin}	0.007	0.823	14901
10	{baking-soda, nutmeg, water}	{pumpkin}	0.007	0.789	14285
11	{butter, cream-cheese, milk}	{confectioners-sugar}	0.005	0.785	13951
12	{cinnamon, pumpkin-pie}	{pumpkin}	0.005	0.769	13919
13	{allspice, water}	{pumpkin}	0.005	0.769	13919
14	{pumpkin-pie, vegetable-oil}	{pumpkin}	0.006	0.764	13837
15	{bread-flour, butter, water, wheat-flour}	{honey}	0.005	0.833	10021

Table 1: The top 15 rules

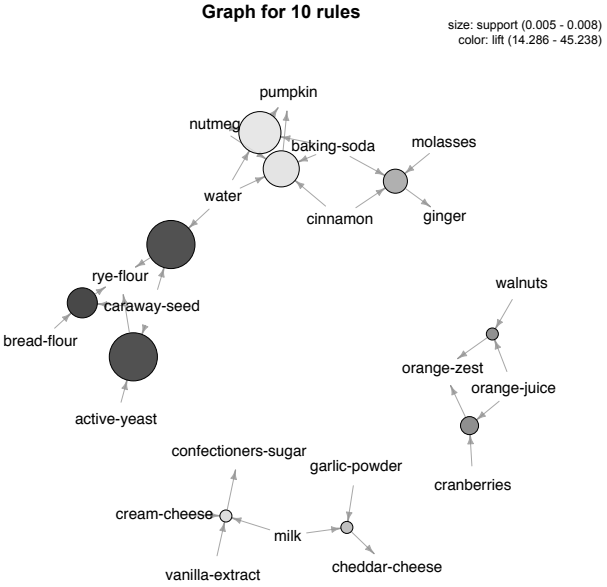

Figure 4: Graph-based visualization with ingredients and rules as vertices for top 10 rules

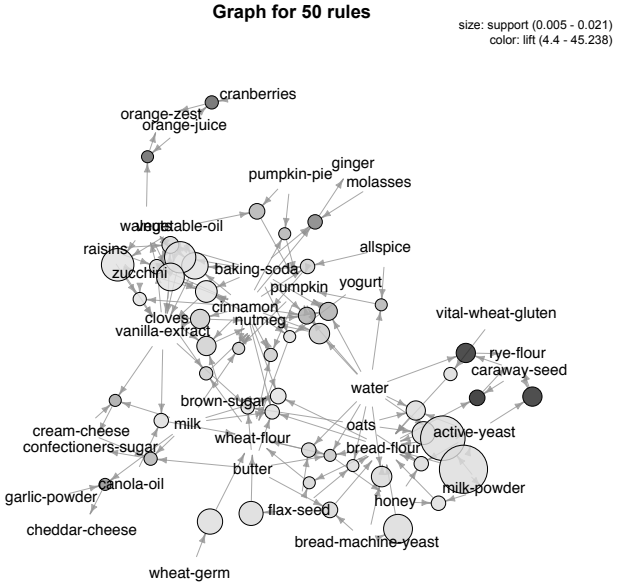

Figure 5: Graph-based visualization with ingredients and rules as vertices for top 50 rules

which contain "orange-juice" and 2 other ingredients in the antecedent and "orange-zest" in the consequent.

5. CONCLUSION

We analyzed 1,900 bakery recipes and found some interesting relations between the ingredients of the recipes. Some of the discovered rules are intuitively known, for example if the recipe contains "yeast" also tends to contain "water", if the recipe contains "apple" also tends to contain "cinnamon". We also found some unexpected combinations of the ingredients that occur in bakery recipes, for example the recipe that contains "baking-soda", "cinnamon" and "molasses" also tends to contain "ginger", the recipe that contains "baking soda", "nutmeg" and "water" also tends to contain "pumpkin". This analysis allows us to see how the ingredients are combined in bakery recipes. The information is very important for food compilers who need to collect analytical data for food items frequently used in national dietary surveys based on foods and recipes.

In the future, we would like to analyze these combinations in order to determine the nutritional properties for different values of quantity-unit pair for each ingredient and to discover for which values of the quantity-unit pair of each ingredient in the combination is good in the meaning of healthy diet. Also, to compare these relations with the relations provided by Foodpairing® that suggests, for one ingredient, those ingredients that create tasteful combinations with the given ingredient [2].

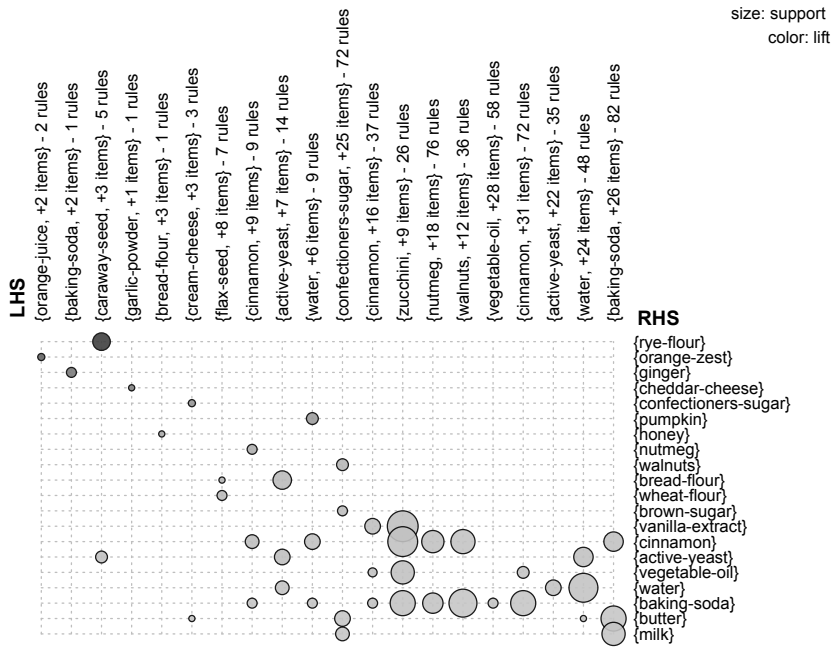

Figure 7: Grouped matrix-based visualization

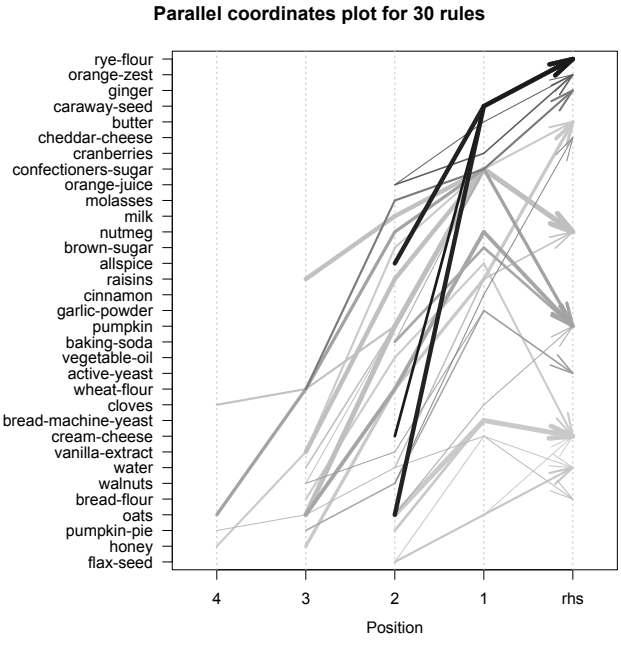

Figure 6: Parallel coordinate plot for top 30 rules

Acknowledgments

This work was supported by the project ISO-FOOD, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 621329 (2014-2019).

References

[1] Data source. http://allrecipes.com/Recipes/Bread/Main.aspx. Accessed: 2014-10-30.

[2] Foodpairing. https://www.foodpairing.com. Accessed: 2015-09-10.

[3] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[4] Hahsler, M., and Chelluboina, S. arulesviz: Visualizing association rules and frequent itemsets. *R package version 0.1-5* (2012).

[5] Hahsler, M., Grün, B., and Hornik, K. Introduction to arules–mining association rules and frequent item sets. *SIGKDD Explor* (2007).

[6] Tan, P.-N., and Kumar, V. Chapter 6. association analysis: Basic concepts and algorithms. *Introduction to Data Mining. Addison-Wesley. ISBN 321321367* (2005).