Mirosława LASEK
Karolina KULIGOWSKA
Małgorzata PIWOWARSKA
Monika POTOCZNA[1]

TEXT MINING APPLICATIONS IN E-MARKETING: ANALYSIS OF COOKING RECIPE TEXTS

Abstract: This paper provides a novel insight into e-marketing research, focusing on text mining methods that companies employ in order to identify websites in which they could post information about their products, or use for advertisement and product placement purposes. The paper focuses on the food industry and therefore we analyse cooking recipe texts available on one of the many Internet sites with recipes. Our analysis consists of clusterization processes, classification models and concept linking associations. Each of these techniques contributes to the selection of marketing tools applied to analysis of websites. Clustering allows to identify groups of recipes that use specific ingredients produced by specific companies. Classification assigns newly added recipes to a suitable category, which afterwards may be used to prepare advertising materials appropriate for selected category. Concept linking helps to discover associations between products and cooking accessories.

Keywords: text mining, e-marketing, cooking recipes, text analysis

1. Introduction

The growing popularity of websites which publish cooking recipe texts, food guides and culinary blogs creates new opportunities for the marketing divisions of food sector companies and for producers of kitchen accessories, tools and utensils. In order to meet this trend, we propose to apply text mining methods to explore culinary websites. Our analysis was performed using SAS Text Miner 4.2 and SAS Enterprise Miner 5.3 software.

The article is organized as follows: in section 2, we present the source of cooking recipe texts analysed in this paper. In section 3, we present clustering methods used in order to determine the ingredients commonly used in cooking as well as the kitchen accessories, and then we group cooking recipe texts according to similar supply requirements. The classification analysis of recipe texts is discussed in section 4, where we compared results of various models of classification in order to determine the best model that assigns a recipe for adequate category. In section 5, we present analysis of linkages that occur among terms contained within recipes. Finally, in section 6 we present conclusions we have reached and indicate lines of further research.

2. Data source

Cooking recipe texts are easily available on the Internet, in various forms, including cookbooks, cooking encyclopedias and online culinary games [Badra, 2008, p.224], [Shidochi, 2009, p.1]. The cooking recipe texts used in our analysis come from: http://www.razzledazzlerecipes.com. This website is a rich source of culinary ideas; it offers recipes for dishes with specific ingredients, such as chicken, as well as ones for special occasion dishes, e.g. Thanksgiving dinner. Such a comprehensive database of recipes allows one to identify the most frequently used ingredients and kitchen accessories.

[1] Mirosława Lasek, Prof., Karolina Kuligowska, PhD, Małgorzata Piwowarska, BA, Monika Potoczna, BA - Department of Information Systems and Economic Analysis, Faculty of Economic Sciences, University of Warsaw

3. Clustering

In the preliminary step, we introduced parsing - a method using natural language processing techniques - to extract the information hidden in the above-mentioned database [Blansché, 2010, p.195]. We identified 78,614 terms occurring in recipe texts. Among the most popular were: *recipe* (5615 recipes) and *cup* (a cup is a common cooking measurement in the USA; it appears in 4328 recipes). Additionally, many of the recipes were associated with baking (the term *bake* appears in 2911 recipes). Afterwards, clustering analysis was performed three times, and each time we modified the program settings to get the best interpretable clusters, most useful from the marketing research point of view. Due to the length of the recipe texts, we chose Singular Value Decomposition as a dimension reduction method.

The first clustering was performed using Entropy as a term weighting function, which assigns greater weight to expressions that occur rarely [Nigam, 1999, p.2]. The number of terms describing a cluster was set at 5. As a result, we obtained 14 clusters (Fig. 1), which were not so clearly divided into topics and did not allow us to identify the main themes of each cluster except for cluster number 2 (with keywords: *salad, green, macaroni salad recipe, salad recipe*, indicating recipes for salads), and cluster number 5 (described by the words: *taco, Mexican, Mexico, salad*, indicating recipes for Mexican dishes).

Fig. 1. First clustering results (Entropy)

1	+ portion, no, + site, without, + reproduce
2	+ salad, green, + macaroni salad recipe, + potato, + salad recipe
3	cream, source, cream, + ice, online
4	+ bread, copyright, + cup, + cup, + home
5	+ taco, mexican, mexico, + salad, mexican
6	+ sugar, + cup, + minute, + egg, + recipe
7	+ add, + minute, with, in, dazzle
8	+ ham, easter, online, source, + easter recipe
9	+ grill, + tablespoon, + minute, with, dazzle
10	into, with, in, + home, recipes
11	+ tablespoon, everyday, + teaspoon, copyright, + minute
12	recipes, copyright, + recipe, + home, razzle
13	+ serving, + minute, + home, + cup, with
14	+ halloween recipe, + not, halloween, + candy, source

Source: own elaboration.

Because of the unsatisfactory results of the first clustering, we decided to clear the recipe database and remove redundant terms which did not carry any significant information [Solka, 2008, p.96]. Therefore we removed such terms as *recipe*, and parts of speech such as auxiliary verbs, conjunctions, articles, interjections, particles, prepositions, pronouns, abbreviations, adjectives, numerals, nouns, and deverbal adverbs. However, we left relevant parts of speech, such as adjectival adverbs (e.g. referring to the cooking time), nouns (in order to outline food ingredients and kitchen tools), and verbs (e.g. the characteristic verb *to bake* that is used in cake recipes). We used Mutual Information as a weighting method for terms. This method allows one to check how the distribution of documents containing a specific term suits the division into group categories [Albright, 2001, p.2]. The second clustering did not bring the expected results (Fig. 2), and therefore we continued to work on clearing the text data.

Fig. 2. Second clustering results (Mutual Information)

1	+ egg, + sugar, + bake, + mix, + cup
2	+ dish, + onion, + taco, + serve, + copyright
3	+ cook, + onion, + pepper, + serve, + be
4	+ mix, + bake, + flour, + sugar, + egg
5	+ teaspoon, + copyright, + bake, + flour, + bread recipe
6	+ tablespoon, + teaspoon, + copyright, + cook, + minute
7	+ pepper, + copyright, + home, + recipe, + cook
8	source, + forget, + not, + candy, halloween
9	+ tablespoon, + pepper, + grill, + cook, + serve
10	+ stir, + not, + sugar, + recipe, + cup
11	razzle, dazzle, + ham, + easter recipe, source
12	+ chop, + serve, + slice, + bake, + bowl
13	+ have, + need, + use, as, + do

Source: own elaboration.

Before the subsequent clustering process, we removed a number of words denoting measures (e.g. *inch*), noun groups and individual terms that did not add significant information to clustering, which resulted in the creation of a stop list containing 11360 irrelevant terms. We also decided to change the weight of the terms on the Global Frequency Times Inverse Document Frequency (GF-IDF) function, which assigns the greatest importance to terms which are rarely found in the database. This procedure is consistent with the Zipf law, according to which the frequency of a term in a document is inversely proportional to its rank [Reincke, 2003, p.2]. The number of terms describing the cluster was set at 10.

As a result we obtained seven clusters (Fig. 3) that represent the most common types of recipes appearing on the website. The first cluster comprises Easter food recipes (hence the words *Easter* and *egg*). The second cluster contains recipes related to Halloween (the words *Halloween* and *candy*). The third cluster focuses on recipes for salads (the words *salad* and *coleslaw*). The fourth and fifth clusters include recipes for cakes (the words *oven* and *bake*). The key terms suggest that in the fourth cluster there are recipes for dough, while the fifth cluster contains sponge cake recipes that require whisking eggs (the words *beat* and *egg*). Although the terms describing the sixth cluster do not indicate any recipes of specific type, the seventh cluster points to recipes containing chicken.

Fig. 3. Third clustering results (GF-IDF)

1	easter, + ham, source, + place, + make, + pan, + cover, + hour, + egg, + top
2	+ dessert, source, + candy, + appetizer, + dinner, + site, halloween, + make, + place, + stir
3	macaroni, + potato, + drain, + gelatin, + dress, + salad, + onion, coleslaw, + pepper, + fruit
4	+ oven, + stir, + cup, + sugar, + butter, + egg, + bowl, + mixture, + flour, + bake
5	+ flour, + oven, + cream, + beat, + top, + sugar, + bake, + mix, + cool, + egg
6	+ recipe, + site, c, + sugar, + top, + mix, + make, + place, + butter, + bake
7	+ oil, + heat, + hour, + chicken, + cook, + pepper, garlic, + cover, + sauce, + onion

Source: own elaboration.

To sum up, we used three methods as term weights: Entropy, Mutual Information and Global Frequency Times Inverse Document Frequency. The best results were obtained for Global Frequency Times Inverse Document Frequency clustering with exactly 10 terms describing the cluster. Based on the obtained and identified clusters, we can see that the website publishes mainly recipes for cakes, salads and dishes with chicken.

4. Classification

For more advanced analysis of our database of recipes we decided to use classification models. Application of those models allowed us to check whether the recipes posted on the website *razzledazzle.com* were written in a manner which is accessible to automatic text processing tools, and whether their topic scope could be easily identified by classifying recipes into predefined categories [Ide, 2010, p.242]. The reason is that while preparing a meal, one needs to know what ingredients are required to prepare soup, cake, or pancakes, and whether the dish is suitable for e.g. vegetarians.

For each of the three binary variables (*pancakes*, *meatballs*, *vegan dishes*) describing the belonging of a recipe to appropriate group, adequate classification models were formed. For this purpose, we chose three types of models available in the Enterprise Miner software: Neural Network (NN), Decision Tree (DT) and Memory-Based Reasoning (MBR). In section 4.a we present full results of the procedure followed by the category group of recipes for *pancakes*, along with charts and test set results. For the other two categories, i.e. *meatballs* and *vegan dishes* in sections 4.b and 4.c respectively, we present results obtained after carrying out the same steps as for pancakes, which lead to the emergence of the best classifier for all the three groups of recipe texts.

a) Pancakes

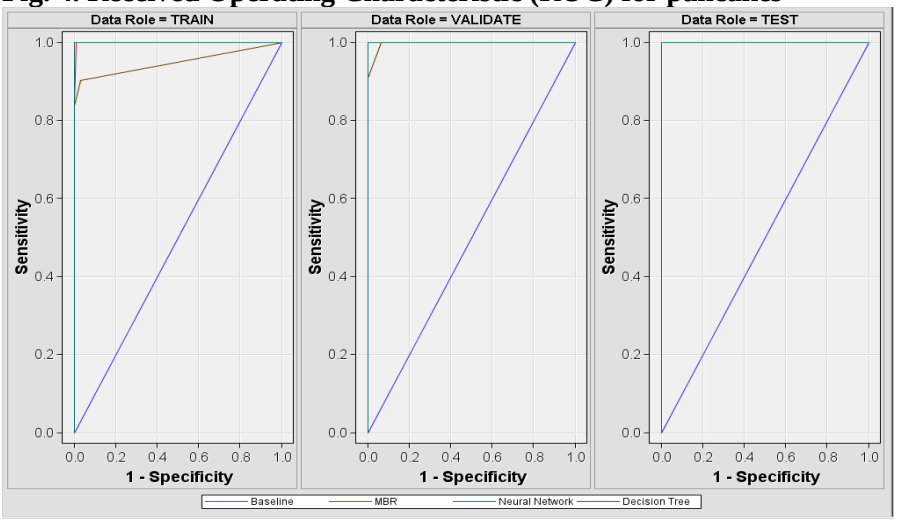
While analyzing the classification models for the variable *pancake*, the database of recipes was divided into training, validation and testing sets in the following proportions 60%, 20%, and 20% respectively. The results presented on ROC[2] charts (Fig. 4) clearly show the superiority of the NN model for each set. For this model, the ROC curves are closest to the upper left corner, which indicates an error-free classification of recipes into appropriate groups. These observations are confirmed by the values of Misclassification Rate shown in Table 1. The first column of the table entitled "Selected model" indicates best-fit model (Y=yes) selected automatically by the SAS software. Note the full accuracy of the classification model in the test set, which is the most important factor in the evaluation of the classifier[3].

Table 1. Values of Misclassification Rates in each classification model (pancakes)

Selected model	Model	Misclassification Rates in each set		
		TRAIN	VALIDATION	TEST
Y	Neural Network	0,000	0,000	0,000
	Decision Tree	0,041	0,024	0,000
	MBR	0,008	0,048	0,000

Source: own elaboration.

Fig. 4. Received Operating Characteristic (ROC) for pancakes



Source: own elaboration.

The results of the MBR model are also presented in Confusion Matrix, as shown in Table 2. The test set contained 42 observations (12 recipes for pancakes with the value of 1 for the variable *pancake* and 30 others with the value 0). The recipes were assigned as follows: 71.43% of recipes (i.e. 30 of 42) were classified as belonging to the pancakes recipes group (100% correct classification rate), while 28.57% of recipes were classified as not belonging to the pancakes recipes group (again, 100% correct classification rate). The total measure of the precision of the MBR model in a test set is 100%. However, despite such a good result, this model was not assessed as the best. The NN model also revealed 100% of the well-assigned recipes in the training and validation sets, which the MBR model does not exhibit. Therefore, given full results, we consider the NN model as the best classification model in this case.

[2] ROC – Received Operating Characteristic plots TP (True Positive rate) versus FP (False Positive rate), showing a ratio of correctly classified positive observations and negatives classified as positive observations [Blockeel, 2011, p.5].
[3] Performance of a classifier in test set is the main factor in choosing the best model, because we can observe the efficiency of a classifier in cases unseen before, which is the very goal of creating a model.

Table 2. Confusion Matrix for Memory-Based Reasoning classification model (pancakes)

PANCAKE I_PANCAKE(Into: PANCAKE)

Liczebność / Procent / Proc. wier. / Proc. kol.	0	1	Razem
0	30	0	30
	71.43	0.00	71.43
	100.00	0.00	
	100.00	0.00	
1	0	12	12
	0.00	28.57	28.57
	0.00	100.00	
	0.00	100.00	
Razem	30	12	42
	71.43	28.57	100.00

Source: own elaboration.

b) Meatballs

In the case of the classification of meatballs recipes, the Memory-Based Reasoning model presents a definite advantage. Having examined ROC charts, Confusion Matrix and Misclassification Rates (Table 3), we consider the MBR model as the best classification model.

Table 3. Values of Misclassification Rates in each classification model (meatballs)

Selected model	Model	Misclassification Rates in each set		
		TRAIN	VALIDATION	TEST
	Neural Network	0,000	0,146	0,163
	Decision Tree	0,033	0,024	0,000
Y	MBR	0,000	0,000	0,000

Source: own elaboration.

The test set contained 43 recipes, 17 of which belonged to the group of recipes for meatballs, while the remaining 26 were pancakes and vegan dishes recipes. The MBR model perfectly assigned this collection of recipes to appropriate groups: 39.53% of the recipes without any errors were classified as recipes for meatballs, and 60.47% as not belonging to that group. Again, we obtained a 100% correct classification rate in the MBR model, which turned out to be best suited for the analyzed categories. If we assigned "points" for best classification models, we could say that until now the score is 1:1:0 (denoting MBR: NN: DT respectively), with same number of points for NN and MBR.

c) Vegan dishes

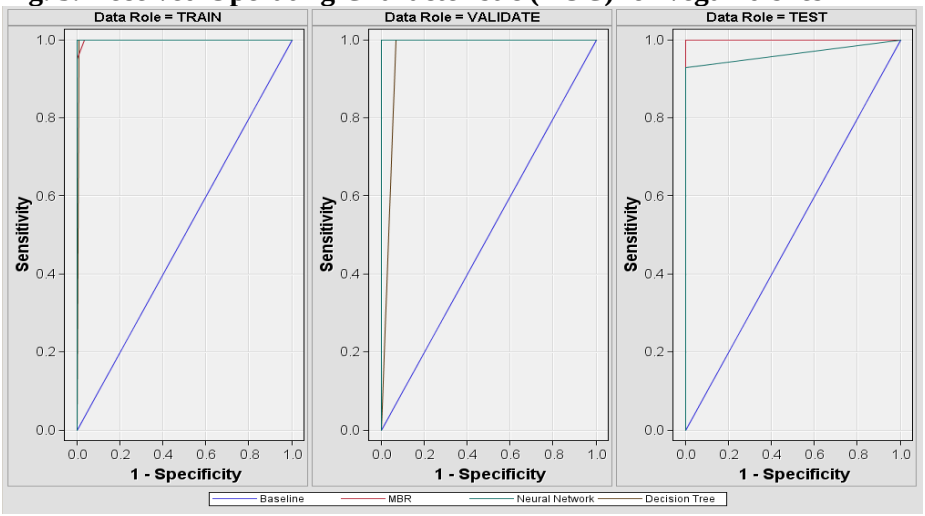
For this group of recipes, ROC charts are worth a closer look (Fig. 5). They show clearly that the leading classification model of recipes from test set is the MBR model. The ROC curve for this model is in the optimal position. Also the Decision Tree model shows the classification accuracy in the test set, but other results, including analysis of the Misclassification Rate and the best model selected by the SAS program (Table 4), indicate conclusively the MBR model as one with the best classification properties for this category of data.

Table 4. Values of Misclassification Rates in each classification model (vegan dishes)

Selected model	Model	Misclassification Rates in each set		
		TRAIN	VALIDATION	TEST
	Neural Network	0,000	0,000	0,023
	Decision Tree	0,008	0,048	0,000
Y	MBR	0,016	0,000	0,000

Source: own elaboration.

Fig. 5. Received Operating Characteristic (ROC) for vegan dishes



Source: own elaboration.

The MBR model reveals 100% correct classification rate in the classification of vegan recipes. All the 14 recipes for vegan dishes (i.e. 32.65% of the 43 recipes in the test set) were assigned correctly, and the remaining recipes (pancakes and meatballs) were classified as not fitting this category. The measure of precision is 100%. Thus again we have a situation where the MBR model turns out to be well suited for cooking text data and it classifies recipes from the test set with 100% correct classification rate.

To sum up the chapter on classification models, the final score of best classification model analysis is 2:1:0, with 2 points on account of the MBR model, 1 point for NN and no points for the DT model. Summary of these results is presented in Table 5.

Table 5. Comparison of classification models

Category of classification	**PANCAKES**	**MEATBALLS**	**VEGAN DISHES**
Selected model	Neural Networks	MBR	MBR
Best fit (decision Y =yes)	+	+	+
Misclassification Rate in test set	0,0000	0,0000	0,0000
Classification precision confirmed by Confusion Matrix (concerns only MBR)	-	100%	100%

Model	Total points
Decision Tree	0
Neural Networks	1
Memory-Based Reasoning	2

MBR model turned out to be the best classification model!

Source: own elaboration.

In the analysis presented above, all recipes were properly classified into appropriate category groups. What was the reason for such a good result? The recipe database on which we made analysis included 207 recipes, which is quite a small number compared to over 14,000 recipes that may be found on the whole website. Therefore, it can be seen that the sample size determined database characteristics and led to enhanced accuracy of basic statistical measures [Baayen, 2001, p.5]. Enterprise Miner copes better with smaller number of documents in the database. For comparison, if we perform the experimental additional clustering (using Entropy as a term weighting function) of the database limited to only 83 recipes, we can immediately see three clusters that correspond with the three categories of recipes: cluster 1 contains vegan dishes recipes, cluster 2 – meatball recipes, and cluster 3 – pancake recipes (Fig. 6).

Fig. 6. Clustering results with database limited to 83 recipes

1	+ vegan recipe, vegetarian, vegan, vegan, + tablespoon	21	0.2530120481...	0.1579433...
2	+ ball, + meatball, + meat, + meat ball recipe, meat	34	0.4096385542...	0.1431953...
3	+ pancake, flour, + cup, in, back	28	0.3373493975...	0.1503122...

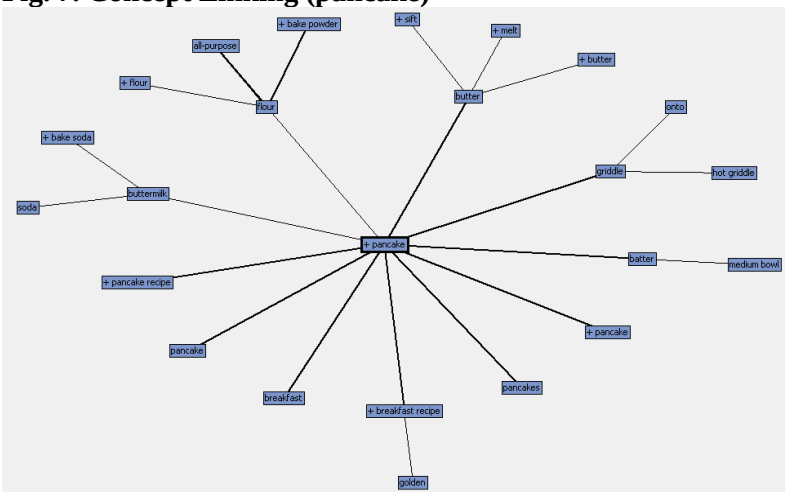
Source: own elaboration.

5. Linking of concepts

Concept Linking graphs with key terms visualization for a group of text documents offer a simple but, at the same time, very useful tool. Recipes linkage may be interesting for a particular manufacturer in order to discover associations among terms within a collection of recipes. To illustrate this method, we performed concept linking analysis for the words from the limited database (Fig. 6), and we chose the most common terms occurring in the recipes database, i.e. *pancake*, *meatball*, *vegan dishes*, and *bake*.

Based on the Concept Linking chart for the term *pancake* (Fig. 7) we can conclude that pancakes are a breakfast dish (*breakfast recipe*), and in order to prepare them one needs to use flour (*flour*), buttermilk (*buttermilk*), melted butter (*butter - melt*), a pan (*griddle*) and a bowl (*bowl*).

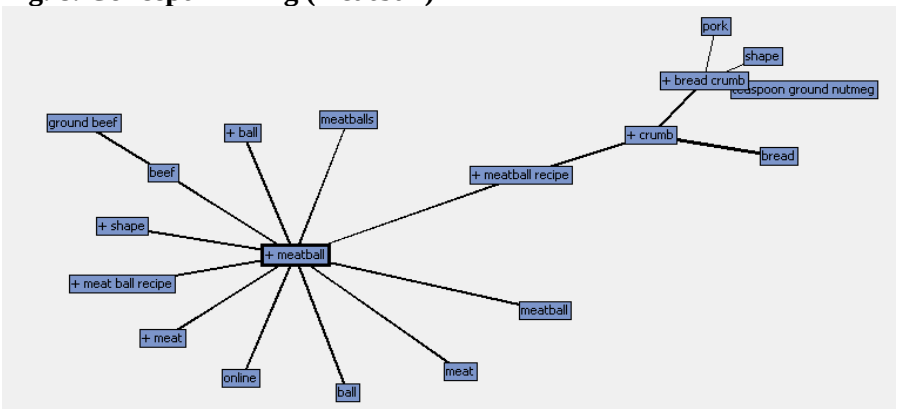
Fig. 7. Concept Linking (pancake)



Source: own elaboration.

In order to o prepare meatballs (Fig. 8), one needs meat (*meat*) of various kinds, for example beef (*beef*) or pork (*pork*). The chosen meat is shaped (*shape*) into round forms (*ball*) and then coated in bread crumbs (*crumb*).

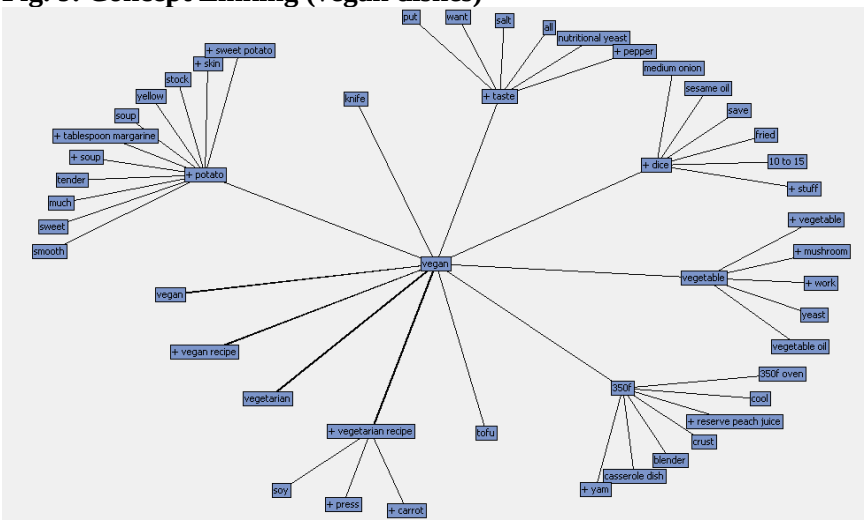
Fig. 8. Concept Linking (meatball)



Source: own elaboration.

Vegan dishes (Fig. 9) are associated with potatoes (*potato*), other vegetables (*vegetable*) and tofu (*tofu*). Among kitchen equipment associated with the term 'vegan', we can find only a knife (*knife*), which corresponds with cutting into cubes (*dice*).

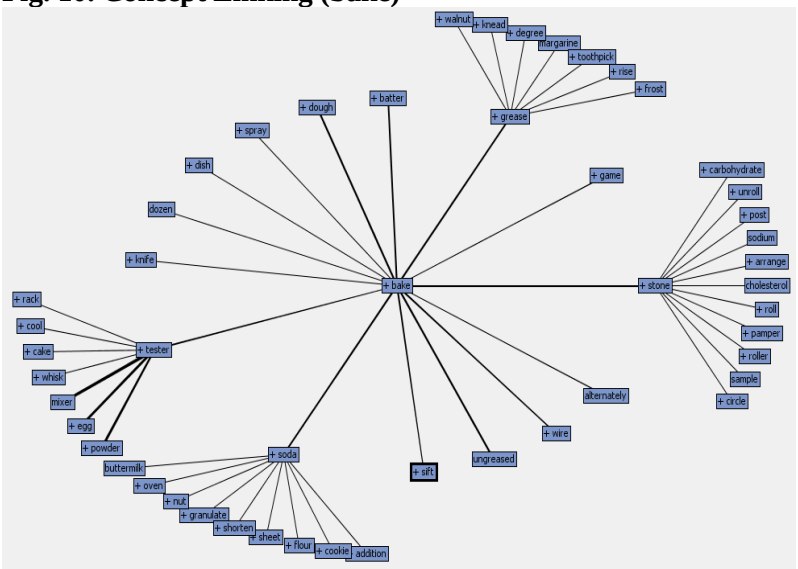
Fig. 9. Concept Linking (vegan dishes)



Source: own elaboration.

One of the first associations with the term *bake* (Fig. 10) is a cake; not surprisingly this term is associated with roast (*bake - dough*). For baking one often uses baking soda (*bake - soda*), while fat (*grease*) is used both for baking cakes and roasting meat. Baking is also linked to kneading dough (*knead*) and toothpicks (*toothpick*), which are used to check if cake or meat is ready for consumption. Summing up the tree linkings, in order to start baking one needs soda (*soda*), baking powder (*powder*), eggs (*egg*), additives such as nuts (*walnut*), and oven (*oven*) heated to a suitable temperature (*degree*).

Fig. 10. Concept Linking (bake)



Source: own elaboration.

6. Conclusions

In this article we presented methods of textual data analysis useful in e-marketing research for food industry. In order to examine a database of cooking recipes, at first we used clustering. The best results were obtained using Global Frequency Times Inverse Document Frequency as the term weighting method. The results comprised seven clusters, which indicated the main topics of recipes, including salads, cakes recipes, dishes with chicken and special occasion dishes.

The second part of our analysis was devoted to classification models. For each of the three selected types of recipes (pancakes, meatballs and vegan dishes) we used the Neural Network (NN), the Decision Tree (DT) and the Memory-Based Reasoning (MBR) models to select the best model for assigning recipes to appropriate category. In two cases (meatballs and vegan dishes) the MBR model proved to be the best, for pancake recipes the NN model exhibited the best fit. The DT model did not have the best-fit characteristics in any case. In each selected model, the correct classification rate in the test set was 100%, therefore we assume that the recipes are framed in an accessible way, easy to identify their affiliation to specific topic categories, which increases the reliability of targeting ads to the right audience.

By applying the Linking Concept analysis, we succeeded in identifying cooking ingredients (such as buttermilk, flour, eggs), kitchen accessories that one needs for cooking (for example knife, oven, griddle), as well as specific products for preparation of foods, like tofu for vegan dishes or baking powder for cakes. Among many websites that contain recipes, manufacturers of cakes and baked goods could be interested in placing advertising information about their products only on targeted websites.

To conclude, the analysis of textual data, such as cooking recipes, constitutes a useful tool for marketing departments of companies - in this case ones from the food sector and kitchen accessories branches. This may, for instance, reduce the costs of advertising, because it allows firms to reach out to clients who are interested in a particular type of product and persuade them to choose and buy specific products.

Literature

1. Albright R., Cox J.A., Daly K., *Skinning the Cat: Comparing Alternative Text Mining Algorithms for Categorization*, SAS Institute, Chicago, 2001
2. Baayen R.H., *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht, Boston, London, 2001
3. Badra F., Bendaoud R., Bentebitel R, Champin P.A., Cojan J., Cordier A., Després S, Jean-Daubias S., Lieber J., Meilender T., Mille A., Nauer E., Napoli A., Toussaint Y., *TAAABLE: Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking*, 9th European Conference on Case-Based Reasoning, Trier, Germany, 2008
4. Blansché A., Cojan J., Dufour-Lussier V., Lieber J., Molli P., Nauer E., Skaf-Molli H., Toussaint Y., *Taaable 3: Adaptation of ingredient quantities and of textual preparations*, 18th International Conference on Case-Based Reasoning, Alessandria, Piemonte, Italy, 2010
5. Blockeel H., *Machine Learning and Inductive Inference*, course text, K.U. Leuven, Acco, 2011
6. Ide I., Shidochi Y., Nakamura Y., Deguchi D., Takahashi T., Murase H, *Multimedia supplementation to a cooking recipe text for facilitating its understanding to inexperienced users*, 2010 IEEE International Symposium on Multimedia, Nagoya, Japan, 2010
7. Nigam K., Lafferty J., McCallum J., *Using Maximum Entropy for Text Classification*, Workshop on Machine Learning for Information Filtering, Stockholm, Sweden, 1999
8. Reincke U. (ed.), *Profiling and classification of scientific documents with SAS Text Miner*, The third Knowledge Discovery Workshop, Karlsruhe, Germany, 2003
9. Shidochi Y., Takahashi T., Ide I., *Finding Replaceable Materials in Cooking Recipe Texts Considering Characteristic Cooking Actions*, CEA'09, Beijing, China 2009
10. Solka J. L., *Text Data Mining: Theory and Methods*, Statistics Surveys No. 2, 2008