Investigating the Healthiness of Internet-Sourced Recipes

Implications for Meal Planning and Recommender Systems

Christoph Trattner
MODUL University Vienna & Know-Center,
Austria
christoph.trattner@modul.ac.at

David Elsweiler
University of Regensburg, Germany
david@elsweiler.co.uk

ABSTRACT

Food recommenders have the potential to positively influence the eating habits of users. To achieve this, however, we need to understand how healthy recommendations are and the factors which influence this. Focusing on two approaches from the literature (single item and daily meal plan recommendation) and utilizing a large Internet sourced dataset from Allrecipes.com, we show how algorithmic solutions relate to the healthiness of the underlying recipe collection. First, we analyze the healthiness of Allrecipes.com recipes using nutritional standards from the World Health Organisation and the United Kingdom Food Standards Agency. Second, we investigate user interaction patterns and how these relate to the healthiness of recipes. Third, we experiment with both recommendation approaches. Our results indicate that overall the recipes in the collection are quite unhealthy, but this varies across categories on the website. Users in general tend to interact most often with the least healthy recipes. Recommender algorithms tend to score popular items highly and thus on average promote unhealthy items. This can be tempered, however, with simple post-filtering approaches, which we show by experiment are better suited to some algorithms than others. Similarly, we show that the generation of meal plans can dramatically increase the number of healthy options open to users. One of the main findings is, nevertheless, that the utility of both approaches is strongly restricted by the recipe collection. Based on our findings we draw conclusions how researchers should attempt to make food recommendation systems promote healthy nutrition.

Keywords

online recipes; public health; recommender systems; meal planning

1. INTRODUCTION

Food recommenders are often touted as potential means to support healthy nutrition [13, 17]. The majority of the literature on food recommender systems, however, does not incorporate health or healthiness at all. The focus to date has primarily been on understanding and predicting meals users will like (e.g. [13, 17]), which does not necessarily equate with healthy nutrition. Indeed in many cases it will lead to the opposite; people who like fatty or calorie-laden meals will be recommended meals with exactly these properties [12].

Moreover, in the literature, it is common for recommendations to be made based on recipe databases collated via users of Internet food portals (e.g. [17]). It is unclear, though, if recipes sourced in this way are suitable for making healthy dietary recommendations. If we believe that the users of a food portal need dietary assistance, it may be a dangerous assumption to treat the recipes, which they themselves uploaded, as the basis for healthy recommendations.

Objective. This paper addresses both of these issues. We work towards integrating health into the food recommender system problem by first analyzing the healthiness of recipes sourced via the Internet to determine the suitability of crowd-sourced recipes for healthy nutrition. We use two widely accepted nutritional standards (from The World Health Organisation (WHO) [44] and the United Kingdom Food Standards Agency (FSA) [3]) to measure the healthiness of recipes from the current largest and most popular Internet food portal Allrecipes.com. Concretely, we look at different categories of recipes on the site and show how these might influence user decisions. In a second step, we use user interaction data with the recipes to shed light on the nutritional properties of recipes users prefer. Lastly, we investigate algorithmic approaches from the literature to see how these relate to healthy nutrition as defined by the same internationally recognised health organisations.

Outline. The structure of this paper is as follows: Section 2 reviews relevant related work in the field of recommender systems and beyond, which leads to the formulation of 5 research questions. Section 3 outlines the data set, which forms the basis of our analyses and experiments. Section 4 describes the metrics we use to measure the healthiness of recipes and meal plans. Section 5 presents our findings with each sub-section relating to a specific research question. A discussion of the findings and potential future research directions is given in Sections 6.

2. BACKGROUND & QUESTIONS

Relevant related research is collated in three main sub-sections: First, we review work evaluating the healthiness of Internet-sourced recipes. We continue to review research on recommender systems for food before finally summarizing work studying online food interaction patterns. All three domains contribute to the formulation of our research questions, which are listed at the close of the section.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052573

Table 1: Basic statistics of the Allrecipes.com dataset.

Total published recipes	60,983
Recipes containing nutrition information	58,263
Users with published recipes	25,037
Recipes rated/commented	46,713
Recipes bookmarked	58,194
Bookmarks	17,190,534
Ratings/comments	1,032,226
Users who provided ratings/comments	125,762
Users who provided bookmarks	155,769

Table 2: Distributions of Internet recipes in terms of WHO and FSA health scores.

	Total (Percentage)		Total (Percentage)
WHO score	Recipes n =58,263	FSA score	Recipes n =58,263
0	3319 (.06)	4	2309 (.04)
1	22,009 (.38)	5	4305 (.07)
2	17,403 (.30)	6	8012 (.14)
3	8977 (.15)	7	6834 (.12)
4	4211 (.07)	8	8613 (.15)
5	1767 (.03)	9	11,068 (.19)
6	498 (.01)	10	10,950 (.19)
7	79 (0)	11	5359 (.09)
		12	813 (.01)

Studies on the healthiness of Internet recipes. To our knowledge only two relevant publications have studied the healthiness recipes shared online. Schneider and colleagues investigated the nutritional properties of 96 recipes (entrees and main dishes) sourced via popular online food blogs [35]. The dishes were evaluated using dietary guidelines from the US Department of Agriculture and US Department of Health and Human Services. The analyzed recipes met energy recommendations but were excessive in saturated fat and sodium. A second study compared a sample of 2662 main-dish recipes from the online platform Allrecipes.com to a sample of 100 super-market ready meals and TV chef recipes [38]. Employing FSA and WHO health criteria, the Internet-sourced recipes were found to be the least healthy of the three samples. These findings suggest that Internet sourced recipes are not the healthiest, but offer little insight into what this means for food recommendation.

Studies on food recommenders. Food recommender systems aim to algorithmically suggest meals or recipes to users based on the user's preferences or past behaviour [13]. Freyne and Berkovsky [13] proposed a hybrid algorithm that considers recipe content (e.g., ingredients) and collaborative filtering into a recommender model. Teng et al. [36] on the other hand suggested the use of complement and substitution networks to generate highly accurate predictions. Harvey et al. [17] carried out a long-term study to analyze factors that influence people's food choices. This work provides the first clues regarding the importance of healthiness in the recommendation process. Amongst other factors found to influence ratings, two groups of users were identified, one preferring healthy recipes, whereas a second, larger group did not care about health and typically preferred less healthy meals. More recently health aspects have been considered in the recommendation process by, for example, targeting health care patients [10]. Two algorithmic approaches to incorporating health that have been reported are 1) to modify predictions by incorporating calorie counts into the recommendation algorithm [14] and 2) to use recommendations as a basis for deriving daily meal plans [11]. The idea here is to recommend the user recipes they will like, but combine them in such a way as to achieve balanced plans, which adhere to nutritional guidelines. This idea has yet to be subjected to any rigorous evaluation. Other than that worth mention here is a recent preliminary study conducted by Achananuparp and Weber [8], who propose a novel method for food substitutions, that could be potentially used in health-aware recommender systems. Again, the idea has yet to be subjected to any rigorous evaluation.

Studies on online food interactions patterns. The way people interact with recipes online can give clues about their food preferences and eating habits. Kusmierczyk et al. and Trattner et al. analyzed data from the German community platform Kochbar.de and found clear seasonal and weekly trends in online food recipe production, both in terms of nutritional value (fat, proteins, carbohydrates, and calories) [23, 40] and in terms of ingredient combinations and experimentation [22]. Similar patterns were observed by Wagner et al. [42] and West et al. [43]. West and colleagues also found correlations between recipes accessed via search engines and incidence of diet-related illness, which resemble findings reported recently by Said & Bellogin [33], De Coudhury et al. [9] and Abbar et al. [7, 26] in the context of Allrecipes.com, Instagram and Twitter respectively. Rokicki et al. [30] investigated differences in nutritional values between user recipes created by different user groups finding, for example, that recipes from females are, on average, richer in carbohydrates. The carbohydrate content of recipes seems to decrease with the age of the user mirroring the advice given by most nutrition advice centers. Finally, Wagner & Aiello [41] and Rokicki et al. [31] studied gender differences in eating preferences in the context of the online platform Flickr and Kochbar.de. Cultural differences in terms online cooking were also recently studied by Ahn et al. or Kim et al. [21], investigating the online recipe portals such as cookpad.com, Allrecipes.com and recipesource.com. However, these works do not provide an insight on how recipe preferences relate to the healthiness of a recipe.

Summary. The outlined research reveals 1) we know little about the healthiness of online recipes or their suitability for healthy food recommendation, 2) the way people interact with recipes online can give clues about food preferences, but it is unknown how this relates to healthiness and 3) knowledge of preferences can be used to improve recommendations, but only preliminary work has been performed to test two proposed strategies for healthy food recommender systems: plans and single-item recommendation incorporating calorie counts. How these strategies relate to the healthiness of the collection is also an open question. Based on the summarized literature, we identify the following research questions:

- **RQ1:** How healthy are Internet-Sourced recipes with respect to recognized standards?
- **RQ2:** How do user interactions such as ratings, comments or social bookmarks people apply to recipes relate to the healthiness of recipe content?
- **RQ3:** How healthy are the recipes recommended by standard recommendation algorithms when applied to the food recommendation problem?
- **RQ4:** Can we improve standard recommender algorithms in terms of making the recommendations they offer more healthy?
- **RQ5:** How easy is it to combine recipes in the form of meal plans in a healthy manner?

Table 3: Nutritional content (Energy, fat, saturated fat, sugar and sodium) per 100g of Internet recipes created by users in Allrecipes.com in each category, user interactions (comment sentiment, number of bookmarks, rating and number of ratings) and user health perception (1=unhealthy to 7=healthy) sorted by FSA score (4=healthy to 12=unhealthy). Furthermore, we show the simulated FSA front of package label (green, amber and red) for an average recipe to visually highlight differences between categories. A Kruskal-Wallis test performed on each column reveals there are statistically significant differences between the categories ($p < .001$).

		Mean											
			FSA front of package label				User Interactions					Health scores	
Category	n	Energy (kCal)	Fat (grams)	Sat. Fat (grams)	Sugar (grams)	Sodium (grams)	Comment Sentiment	Num Bookmarks	Rating	Num Ratings	User Health Perception†	WHO score	FSA score‡
Desserts	11,317↑	331.48↑	16.27 ↑	7.27 ↑	27.92 ↑	0.21 ↓	1.67	298.59↓	4.27	19.35	$2.06^{(0)}$	1.61	$9.64^{(1)}$
Ingredients	2039	265.06↑	14.13 ↑	5.84 ↑	16.44 ↑	0.36 ↑	1.92↑	1913.21↑	4.57↑	133.66↑	$4.28^{(-15)}$	1.59	$9.06^{(2)}$
Dinner	1033↓	166.61	9.07	3.44	2.59 ↓	0.35	1.94↑	2553.92↑	4.53↑	163.28↑	$4.31^{(-15)}$	1.41	$8.43^{(3)}$
Holidays and events	11,185	218.42↑	11.33 ↑	4.52 ↑	12.62 ↑	0.28	1.76	526.6↑	4.39	31.81	$2.66^{(+1)}$	1.87	$8.38^{(4)}$
Trusted brands	1744	200.45	10.06	4.08 ↑	8.73	0.32	1.77	111.02↓	4.37	6.57↓	$3.13^{(7)}$	1.83	$8.2^{(5)}$
Bread	2972	261.86↑	9.95	3.53	12.72 ↑	0.35 ↑	1.7	438.66	4.29	32.37↑	$3.63^{(-4)}$	2.42	$8.18^{(6)}$
Meat and poultry	12,672↑	151.97	8.46	3.09	2.62	0.33	1.74	465.88	4.3	26.79	$3.47^{(-2)}$	1.62	$8.17^{(7)}$
Breakfast and brunch	2167	188.8	9.26	3.56	7.82	0.28	1.69	377.25	4.31	22.86	$4.16^{(-6)}$	2.11	$8.09^{(8)}$
Main dish	13,188↑	159.51	8.36	3.08	2.48 ↓	0.31	1.73	438.92	4.27	25.59	$4.22^{(-7)}$	1.77	$8.09^{(9)}$
Appetizers and snacks	4162	226.67↑	15.73 ↑	5.79 ↑	4.8	0.44 ↑	1.74	428.86	4.35	25.4	$3.03^{(+4)}$	1.82	$8.08^{(10)}$
US recipes	3556	185.89	9.76	3.52	8.3	0.36 ↑	1.65↓	313.67	4.32	16.1↓	$2.19^{(+9)}$	1.92	$8.08^{(11)}$
Grilling	1682↓	156.72	8.74	2.77	4.83	0.54 ↑	1.83↑	481.01	4.41↑	22.68	$2.84^{(+8)}$	1.64	$8^{(12)}$
Allrecipes magazine	842↓	190.79	10.08 ↑	3.84	9.27	0.33	1.86↑	1952.1↑	4.54↑	142.78↑	$4.22^{(-2)}$	2	$7.94^{(13)}$
Everyday cooking	22,657↑	187	9.69	3.71	8.66	0.28	1.73	506.92	4.32	31.74	$4.47^{(-5)}$	2	$7.97^{(14)}$
Quick and easy	1955	167.82	8.65	3.23	2.39 ↓	0.32	1.7	404.72	4.25↓	23.55	$3.25^{(+7)}$	1.83	$7.86^{(15)}$
Pasta and noodles	2692	186.21	8.62	3.28	2.79	0.27	1.68	388.21	4.21↓	22.53	$3.84^{(+5)}$	2.31	$7.82^{(16)}$
Fruits and vegetables	19,574↑	171.44	8.7	3.25	9.06	0.24 ↓	1.73	373.59	4.32	21.85	$6.34^{(-9)}$	2.15	$7.76^{(17)}$
World cuisine	7444	178.05	9.05	3.26	7.46	0.29	1.68	361.72	4.28	19.53	$4.59^{(-3)}$	2.16	$7.68^{(18)}$
Lunch	693↓	158.36	9.1	2.78	3.11	0.32	1.94↑	515.8	4.6↑	26.54	$3.94^{(+6)}$	2.07	$7.63^{(19)}$
Slow cooker	1283↓	121.26↓	5.66 ↓	2.17 ↓	3.67	0.3	1.6↓	709.98↑	4.18↓	37.16↑	$5.19^{(-2)}$	1.89	$7.6^{(20)}$
Seafood	3237	157.6	8.94	3.05	1.79 ↓	0.32	1.75	298.29↓	4.31	16.95↓	$5.50^{(-2)}$	1.9	$7.46^{(21)}$
Salad	3031	146.84	9	1.93 ↓	4.48	0.24	1.78	247.46↓	4.36	13.17↓	$6.00^{(-3)}$	2.33	$7.22^{(22)}$
Vegetarian	4889	159.09	8.47	3.01	5.95	0.26	1.66↓	417.68	4.22↓	23.87	$5.50^{(-1)}$	2.58	$7.15^{(23)}$
Side dish	4006	128.99↓	6.64 ↓	2.69	3.71	0.24	1.71	324.4	4.3	19.1	$3.84^{(-12)}$	2.58	$6.97^{(24)}$
Soups stews and chili	3605	82.93↓	3.89 ↓	1.59 ↓	1.65 ↓	0.22 ↓	1.69	323.19	4.32	20.12	$4.56^{(+5)}$	2.29	$6.87^{(25)}$
Drinks	1801	86.37↓	1.5 ↓	0.82 ↓	10.22 ↑	0.03 ↓	1.57↓	126.26↓	4.36	6.51↓	$2.88^{(+21)}$	2.51	$6.01^{(26)}$
Healthy	3175	107.83↓	2.34 ↓	0.56 ↓	6.77	0.2 ↓	1.65↓	340.03	4.21↓	17.97	$6.53^{(0)}$	3.43	$5.6^{(27)}$
All recipes	58,263	204.87	10.58	4.10	10.55	.31	1.70	295.05	4.29	17.72	4.10	1.94	8.13

Note: Top-5 values in respect to macro nutr. content (i.e. Fiber, Sodium, Fat,...) and user interactions marked with ↑, bottom-5 in the corresponding column highlighted with ↓.
† Superscripts denote differences in ranking when compared to the FSA ranking of the actual category. ‡ Superscripts denote category ranking in respect to the FSA score.

3. DATASET

To address these questions we obtained recipe and nutritional data from the Web by implementing a standard Web crawler. Between 20^{th} and 24^{th} of July 2015, the crawler collected 60,983 recipes published between the years 2000 and 2015 on the Allrecipes.com website. We focus only on recipes that have been published on the main site and ignore personal recipes, which are often incomplete and do not provide nutrition information. Allrecipes.com was chosen for two main reasons. First, at the time of writing, it claims to be the world's largest food-focused social network. The site has a community of 40 million users accessing 3 billion recipes annually across 24 countries [4]. Second, the site has been associated with positive press coverage, claiming that "...diabetics, coeliac and even those specifically wanting to increase their fibre intake - are all catered for" [5]. Positive press combined with government health campaigns promoting home-cooking (e.g. [2]) may persuade members of the public that cooking recipes sourced from the Internet is an approach likely to improve their diet, this despite no systematic study having comprehensively assessed the nutritional content of online recipes or the technology used to access them.

In addition to comments, bookmarks, ratings, and user profiles, the following information was collected for each recipe: year of publication, the recommended number of servings; and total energy (kCal), protein (g), carbohydrate (g), sugar (g), sodium (g), fat (g) saturated fat (g), and fibre (g) content. The nutritional meta-data was available via Allrecipes.com and collected during the main crawl. Allrecipes.com estimates the nutritional content for an uploaded recipe by matching the contained ingredients with those in the ESHA research database [6]. Table 1 provides an overview of the basic statistics of the dataset. .

4. MEASURING HEALTHINESS

Throughout our analyses we make use of two internationally recognized standards for measuring the healthiness of meals and meal plans: The World Health Organization (WHO) guidelines [3] and the UK FSA "traffic light" system for labeling food [44].

The WHO has defined 15 ranges of macro-nutrients which should be considered in a daily meal plan. We follow the approach of Howard et al. [18] who chose the 7 most important (i.e. proteins, carbohydrates, sugars, sodium, fats, saturated fats, and fibers) and their corresponding ranges to determine a so-called WHO health score. The scale ranges from 0 - 7 (0 meaning none of the WHO ranges are fulfilled and 7 meaning all ranges are met). A recipe or meal plan with a WHO score of 7 is interpreted as being very healthy whereas a score of 0 is seen as very unhealthy.

A similar approach is taken to derive a FSA traffic light labeling system score. The FSA score relates only to 4 macro-nutrients (sugar, sodium, fat and saturated fat). The scale is green (healthy),

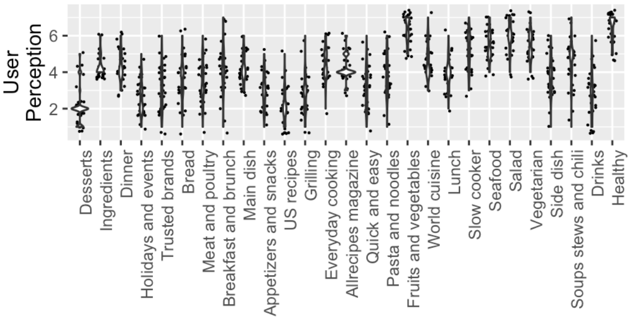

Figure 1: Violin plot shows how user perception of the healthiness (1=unhealthy to 7=healthy) varies across Allrecipes.com categories (sorted by highest FSA score (left) to lowest (right)).

amber and red (unhealthy). In order to derive a single metric we follow the procedure of Sacks et al. [32] who first assign an integer value to each color (green=1, amber=2 and red=3) then sum the scores for each macro-nutrient resulting in a final range from 4 (very healthy recipe) to 12 (very unhealthy recipe).

5. RESULTS

The following sub-sections provide answers to the above listed research questions.

5.1 RQ1: Determine the Healthiness of Internet Recipes

Table 2 presents the FSA and WHO score distributions over the full collection. The analyses do not suggest the recipes to be particularly healthy. 3319 (5.7%) recipes failed to meet any of the WHO guidelines. Only 79 (0.14%) meet all of the criteria. The majority of recipes meet only 1 or 2 guidelines (67.6%). In terms of the FSA criteria, few recipes receive all green (4%) or all red scores (1%). As shown in the last row of Table 3, on average the recipes in the dataset receive a red-score for fat and saturated fat, and a medium score for sodium content. Sugar-content, however, receives a green score on average.

There are 27 main categories of recipe on the Allrecipes.com website. These include types of meal (e.g. main dish, dinner, breakfast and brunch etc.), as well as characteristics of dishes (e.g. quick and easy, slow-cooker, vegetarian and healthy-recipes). Table 3 depicts the average nutritional properties across these categories showing that the healthiness of recipes in different categories varies greatly. Predictably, "dessert recipes" are the least healthy whereas those in the "healthy recipes" category are the most healthy. Less predictable results include that recipes in the "quick and easy" category are low sugar, but high in fat. It seems that recipes in the "main meal" and "dinner" categories are less healthy than "sides" and "lunches" category. This begs the question of whether it is better for users to combine such smaller dishes in their diet. On average recipes in the "vegetarian" category were determined as healthy, with no FSA criteria being assigned a red label. In summary, the analyses show that based on the FSA and WHO criteria, there are healthy recipes in the collection, however, overall, the recipes can be considered to be relatively unhealthy.

To determine whether users are aware of these nutritional differences between categories we performed an additional user study. 32 participants (34.3% female) recruited via social-media rated each category on a 7-point scale (1=unhealthy to 7=healthy).

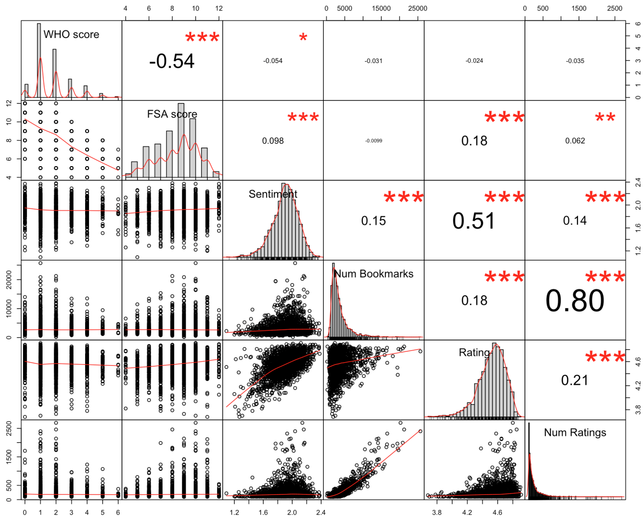

Figure 2: Correlation matrix (spearman) depicting how WHO and FSA scores correlate with sentiment, num. of bookmarks, ratings and num. of ratings. Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

The results are shown in Table 3 (column User Health Perception) with changes in rank being shown in superscript. Positive changes mean categories were ranked as healthier than estimated by FSA score. Some differences were observed between the healthiness rating provided by participants and the metrics we calculated for each category, but there was evidence of some overlap. A spearman rank correlation analysis shows weak correlation with the WHO score ($rho = .33$, $p < .001$, $n = 864$) and a medium correlation for the FSA score ($rho = -.42$, $p < .001$, $n = 864$). The rank of certain categories, such as "desserts" and "healthy" were predicted exactly, while others such as "vegetarian" and "holiday and events" were also very close. Others categories were on average very poorly estimated, including "drinks", which was ranked as much healthier than its FSA score and "sides", which contained much healthier recipes than the participants believed. While participants showed high-agreement in judgments for some categories (e.g. "healthy" and "seafood") (see Figure 1), the judgments for many other categories were much more varied. This is confirmed by an overall Fleiss' Kappa Inter-rater agreement score of $\kappa = .165$ ($z = 42$, $p < .001$).

Thus, on average the recipes on Allrecipes.com are judged to be relatively unhealthy, although recipes rated as very healthy can also be found. Some categories of recipes are much healthier than others although not all users are able to judge this effectively.

5.2 RQ2: Investigating User Interaction

From the literature we know that user interaction data informs on user preferences, context information and other external behaviour. This information forms the basis of the recommendation process as we show later in the paper. To determine whether it also provides insight into healthiness, we examine four different means by which users can interact with recipes. We look at the recipes users saved to their favorites list (bookmarked), the ratings users applied to recipes (on a 5-point scale), the number of comments left on recipes and the sentiment scores for comments (ranging from -4 to +4)[1]. To reduce effects noise and to obtain enough data evidence for the vari-

[1] Comment sentiment was determined via the popular SentiStrength framework (http://sentistrength.wlv.ac.uk/) [37].

Table 4: Predicting WHO and FSA scores employing rating, sentiment, number of ratings, bookmarks and category features using ordinal logit models. Only best models (performing a step-wise analysis) with corresponding coefficients and odd ratios are presented.

	Dependent variable											
	WHO score (0=unhealthy to 7=healthy)						FSA score (4=healthy to 12=healthy)					
Model	(1_{who})		(2_{who})		(3_{who})		(1_{fsa})		(2_{fsa})		(3_{fsa})	
Coef.	β	OR†	β	OR†	β	OR†	β	OR†	β	OR†	β	OR†
Sentiment	$-.413^*$	.662			$-.461^*$	.631					.435	1.546
Num Bookmarks (log)					$.269^*$	1.309	$-.667^{***}$	.513			$-.378^{***}$	.685
Rating							1.651^{***}	5.214			1.115^{***}	3.050
Num ratings (log)	$-.102$	.903			$-.310^{**}$	.734	$.673^{***}$	1.960			$.376^{***}$	1.457
Category‡			2.373^{***}	10.734	2.325^{***}	10.229			-2.234^{***}	.107	-2.360^{***}	.094
Observations	1963		1963		1963		1963		1963		1963	
Log Lik	-2856.794		-2590.669		-2586.256		-3780.543		-3382.027		-3361.525	
AIC	5729.587		5223.338		5220.511		7583.086		6808.055		6775.05	
McKelvey & Zavoina R^2	.003		.248		.252		.044		.356		.370	

Note: $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$

† Odd ratios. ‡ Categories have been collapsed and only most sign. Coef. and ORs are shown, which is the "Healthy" recipe category for the WHO models and "Drinks" for the FSA models.

Table 5: Distributions of user (filtered to at least $k \geq 20$ recipes) and recipe profiles (filtered to at least $k \geq 100$ user interactions = ratings) according to the WHO and FSA health scores.

WHO score	Total (Percentage)		FSA score	Total (Percentage)	
	Users ($k \geq 20$) n =4791	Recipes ($k \geq 100$) n =1963		Users ($k \geq 20$) n =4791	Recipes ($k \geq 100$) n =1963
0	0 (.00)	152 (.08)	4	0 (.00)	24 (.01)
1	1120 (.23)	852 (.43)	5	0 (.00)	103 (.05)
2	3634 (.76)	556 (.28)	6	0 (.00)	203 (.10)
3	37 (.01)	212 (.11)	7	56 (.01)	220 (.11)
4	0 (.00)	135 (.07)	8	1835 (.38)	306 (.16)
5	0 (.00)	46 (.02)	9	2767 (.58)	488 (.25)
6	0 (.00)	10 (.01)	10	133 (.03)	387 (.20)
7	0 (.00)	0 (.00)	11	0 (.00)	194 (.10)
			12	0 (.00)	38 (.02)

ables avg. rating and sentiment, we only consider recipes that have been rated and commented on at least 100 times. Ideally the recipes bookmarked most often, rated highest and assigned the most positive comments would also be the healthiest. To establish whether this is indeed the case we performed a correlation analysis using pairwise spearman rank correlations.

Figure 2 presents the results and first of all shows that there is a sign. negative medium correlation ($rho = -.54, p < .001$) between the FSA and WHO scores, which could be expected as both health scores should report more or less the same (though on different macro nutrition), but with opposite scales. Furthermore, we find that there are significant correlations between the FSA score and number of ratings ($rho = .062, p < .01$), the rating applied ($rho = .18, p < .001$) and the sentiment score on comments ($rho = .098, p < .001$). Only comment sentiment is significantly correlated with the WHO score ($rho = -.054, p < .05$).

The signs of these correlation coefficients all suggest that the popular and highly-rated recipes are the ones which are the least healthy. Table 4 provides further insight into the relation between interaction data and the healthiness of recipes by presenting 6 ordinal logit models which predict the WHO and FSA scores of recipes based on the users interaction data and the category the recipe was published to. The models were created using a step-wise search approach based on the Akaike Information Criterion (AIC).

The models show that interactive features improves the fit to the data (see (1_{who}) and (1_{fsa})) compared to a null model (=intercept-only model) suggesting that these features offer complementary information (Likelihood ratio tests: (0_{who}) vs (1_{who}); $\chi^2(3) = 6.5$, $p = 0.03$; (0_{fsa}) vs (1_{fsa}); $\chi^2(3) = 92.29, p = 0$). An even better fit can be achieved by using the category information (discussed in Section 5.1) as a predictive feature (see also Table 3). The category information actually offers far more explanatory power than the interactive features (see (2_{who}) and (2_{fsa})), but combining with the interaction features further improves the fit significantly for both FSA and WHO scores (Likelihood ratio tests: (2_{who}) vs (3_{who}); $\chi^2(3) = 8.83$, $p < .032$; (2_{fsa}) vs (3_{fsa}); $\chi^2(4) = 41.00$, $p < .001$; parallel slopes assumption does hold for WHO and FSA score models employing Harrell's graphical method [16]). What is also shown in Table 4 is that the signs of the coefficients for the interaction features of the WHO models are in general negative and positive for the FSA models, which are in line with the results obtained in the correlation analysis as presented in Figure 2. In summary, we can attain information regarding the healthiness of a recipe, both from the categories to which it is assigned and by how users interact with it. The recipes interacted with most often and rated higher tend to be less healthy, which is worrying as these are the recipes most likely to be cooked an eaten.

5.3 RQ3: Analyzing Recommendations

Next we turn our attention to recipe recommendation, investigating how the recommendations provided by commonly applied algorithms relate to health. As algorithms typically promote popular items and items with high ratings and we now know that these tend to be less healthy, we suspected that the recommended items would also be unhealthy.

To test our assumption, we ran a series of experiments evaluating the performance of 9 prominent recommender algorithms on the rating data[2] using the LibRec[3] framework. The algorithms tested are: Random item ranking (our baseline), Most Popular item ranking (MostPop), user- and item-based collaborative filtering (denoted as UserKNN and ItemKNN) [34], Bayesian Personalized Ranking (BPR) [28], Sparse Linear Methods (SLIM) [27], Weighted

[2] We also run all experiments presented below for bookmarking and sentiment data available. For space reasons we can only present the rating data experiments. In general, however, the trends are the same using all proxies. Algorithms show same ranking patterns, with LDA standing out and Random being the worst approach. The only marginal difference between the experiments reported here and the bookmark and sentiment experiments is that in general algorithms perform slightly better in the rating setting, showing an improvement of 1-2%.

[3] http://www.librec.net/

Table 6: Recommender accuracy sorted by nDCG and recommender accuracy post-filtered by FSA scores. The mean WHO and FSA scores of the top-5 recommended recipes are also reported along with the differences in terms of WHO and FSA scores (algorithms sorted by Δ FSA) between recommended recipes and recipes rated by the users. A negative Δ FSA score and a positive Δ WHO score indicates that the recommended list is healthier than the recipes rated by the user. The highest ranking scores are obtained by the LDA approach, while the opposite is observed for the Random approach. All Δ scores are statistically significant at $p < .001$ employing a two-sample t-test. Pairwise comparison employing a two sample t-test shows that all algorithms produce sign. healthier recommendation list when investigating the FSA/WHO and Δ FSA/WHO scores and when applying a health score post-filtering function.

	Mean (n =4791)									
							FSA front of package label			
	MAP@5	nDCG@5	WHO score	FSA score	Δ WHO	Δ FSA	Fat (g)	Sat. Fat (g)	Sugar (g)	Sodium (g)
LDA	**.0175**	**.0395**	1.554	9.110	-.137***	.498***	8.70	3.73	8.73	0.32
WRMF	.0160	.0365	1.496	9.114	-.196***	.503***	9.50	3.89	8.84	0.34
AR	.0149	.0343	1.550	9.206	-.141***	.595***	9.27	4.12	10.50	0.25
SLIM	.0143	.0326	1.643	8.907	-.048***	.295***	9.27	3.82	7.91	0.33
BPR	.0141	.0325	1.432	9.252	-.259***	.641***	8.69	3.82	7.83	0.29
MostPop	.0126	.0294	1.537	9.004	-.154***	.393***	9.02	3.94	10.01	0.23
UserKNN	.0100	.024	1.583	8.985	-.108***	.372***	8.96	3.73	7.98	0.31
ItemKNN	.0073	.0178	1.660	8.652	-.032***	.041***	8.59	3.51	6.03	0.31
Random	.0011	.0029	**1.750**	**8.486**	**.059*****	**-.126*****	8.74	3.49	5.71	0.30
	FSA score post-filtered ($score_{u,i,fsa}$)									
LDA	**.0137**	**.0321**	2.170	7.323	.479***	-1.288***	6.51	2.42	4.03	0.29
WRMF	.0131	.0303	2.140	7.361	.449***	-1.250***	6.48	2.30	4.75	0.31
SLIM	.0109	.0248	2.384	7.008	.692***	-1.604***	6.20	2.56	2.59	0.24
AR	.0100	.0238	2.600	6.984	.909***	-1.627***	5.64	1.94	3.95	0.28
MostPop	.0096	.0228	2.542	7.334	.851***	-1.278***	5.37	2.02	2.46	0.24
BPR	.0086	.0205	2.783	6.722	1.092***	-1.889***	6.42	2.30	4.95	0.26
UserKNN	.0069	.0168	2.486	6.722	.795***	-1.891***	6.88	2.73	3.33	0.33
ItemKNN	.0044	.0109	2.703	6.124	1.012***	-2.488***	5.15	1.79	3.51	0.25
Random	.0009	.0022	**3.228**	**4.305**	**1.537*****	**-4.306*****	1.59	0.43	1.45	0.09

Note: $^{***}p < .001$

matrix factorization (WRMF) [19], Association Rules (AR) [20] and Latent Dirichlet Allocation (LDA) [15]. We used 5-fold cross validation as protocol for all the experiments and report the recommendation performance results employing MAP@5 and nDCG@5 as performance metrics [29]; thus we focus on a ranking task aiming to predict the 5 recipes users would rate highest. To determine the healthiness of this list, we again report the mean WHO and FSA scores. All algorithms and appropriate parameters were tuned omitting the hold-out data. To reduce data sparsity issues, a well-known issue in collaborative filtering-based methods [29], we applied a p-core filter approach using only user profiles with at least 20 rating interactions and recipes that have been rated at least 100 times by the users, resulting in a final dataset comprising $n = 4791$ user profiles and $n = 1963$ recipe profiles. More detailed statistics of the filtered dataset are provided in Table 5.

The results of this experiment are shown in the top half of Table 6. In terms of recommendation accuracy factorization approaches, such as LDA or WRMF perform the best, whereas these are the amongst the worst performing algorithms in terms of health scores. Overall, the recommendations generated were not particularly healthy with all algorithms achieving an average WHO score of < 1.8 and FSA score of > 7.8. The best performing approach in terms of health was to recommend recipes at random, which naturally achieved poor results in terms recommendation accuracy. Thus there is a trade-off between giving users what they like and what is healthy.

Examining the delta scores for FSA and WHO, which communicate the differences between the health scores for the recipes used to train the recommendation algorithm and those for the recipes recommended shows that, with the exception of random approach, the difference was always negative for WHO and positive for FSA. In other words the recommended recipes were unhealthier than the positive training cases provided by the user. This means that in general, due to the way they work, standard recommender algorithms implicitly promote unhealthy recommendations.

5.4 RQ4: Generating More Healthy Recommendations

To establish whether we can alter recommendation algorithms to make the recipes they suggest more healthy and in particular to investigate the potential different algorithms have to address the trade-off described above, we evaluate a simple initial solution to the problem that tries to improve the healthiness of the recommended items while preserving the recommender accuracy.

In a first step we performed a correlation analysis between recommender accuracy estimates nDCG and MAP and the FSA and WHO scores (see Table 7). Generally, as expected, the two accuracy metrics (nDCG and MAP) are negatively correlated with WHO and positively correlated with the FSA score. The MostPop

Table 7: Pearson correlations (= rho) between MAP and nDCG and FSA and WHO health scores (on user level) for individual algorithms. As shown, in general, there is a sign. positive correlation between the FSA score and MAP/nDCG measure and negative correlation between the WHO and MAP/nDCG metric.

	nDCG (n =4791)		MAP (n =4791)	
	WHO score	FSA score	WHO score	FSA score
	rho		rho	
Random	-.02	.00	-.02	.00
ItemKNN	.05**	.06***	.04**	.04*
SLIM	-.04**	.14***	-.02	.15***
UserKNN	-.10**	.19***	-.06***	.17***
MostPop	-.59***	.19***	-.52***	.09***
LDA	-.05***	.06***	-.06***	.09***
WRMF	.01	.05***	-.01	.09***
AR	-.09***	.00	-.06***	.03*
BPR	-.18	-.02	-.13***	.01
All	-.15***	.14***	-.13***	.15***

Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

approach – again confirming our suspicions given the results above – shows the strongest correlation. The other algorithms show far weaker correlations hinting that re-ranking items according to their health profiles might work without the same impact on the user preferences and the nDCG and MAP scores.

To test these interpretations, we compare the performance of the algorithms with a simple post-filtering procedure, where each item (recipe) is re-weighted according to a scoring function that could be e.g. linear or of an exp. nature. Post-filtering has been shown to work well in several scenarios in the past. For example in combination with collaborative filtering, the approach works better than matrix-factorization methods using context information directly in the model [39, 25]. To post-filter items in our scenario we apply a simple scoring function which re-weights the scores of a recipe for a particular user based on the WHO or inverse FSA score of the recipe, see:

$$score_{u,i,who} = score_{u,i} \cdot (who_i + 1) \qquad (1)$$

$$score_{u,i,fsa} = score_{u,i} \cdot (16 - fsa_i - 4 + 1) \qquad (2)$$

We also tried other methods, such as linear combinations as discussed in [12], but this offered rather poor performance very close to a random baseline. Our method is parameter free, scalable and can be applied to any existing recommender method without changing the internal properties of the method. As an initial approach it receives solid results (see bottom half of Table 6)[4]. The LDA-based approach provides the most accurate recommendations, while the random approach performs worst. All methods perform significantly better ($p < .001$, pairwise comparison employing a two-sample t-test) when looking at the mean WHO, FSA and ΔWHO and ΔFSA scores compared to their unfiltered derivatives. Although the recommender accuracy drops in all cases ($p < .001$, pairwise comparison employing a two-sample t-test), some of the recommender approaches (e.g. LDA and WRMF) still provide higher accuracy estimates and better health scores when compared to unfiltered algorithms such as MostPop, User or ItemKNN. With respect to individual macro-nutrients, the post-filtered results improve across the board, but in terms of the traffic-light classification the best results are for fat, which are transformed from amber and red scores to all green and for sugar, which improve from mostly red to amber and green. Few classification improvements were achieved for saturated fat or sodium.

These results highlight that 1) it is possible to balance and perhaps optimise the trade-off between recommendation accuracy and the healthiness of recommendations and 2) some recommendation algorithms may be more or less suitable to this process. Nevertheless the results also show that 3) while the approach shows potential benefit and future work should try to optimise the trade-off, the method by itself will not lead to healthy nutrition - at least not with this collection. The post-filtered results with the highest values show that the best FSA and WHO scores were 4.305 and 3.228 respectively and are associated with extremely poor recommendation accuracy. These represent the best health values which can be achieved using an individual item recommendation approach, indicating that complementary ideas are necessary.

5.5 RQ5: Generating Meal Plans

A second approach in the literature to incorporating health aspects in the recommendation process is to combine recommendations in daily meal plans. The idea here is that it is okay to recommend users items they like, even recipes considered unhealthy in isolation, as long as they can be assessed as healthy in terms of a balanced daily meal plan. Elsweiler and Harvey [11] showed by experiment that the approach had potential using a small test collection. We test their idea on the Allrecipes.com dataset. Taking an approach similar to that of Elsweiler and Harvey, we create meal plans derived from recipes from three categories ("breakfasts", "lunches" and "dinners"). A full search is then performed over all recipes to find every combination in the sequence [breakfast, lunch, dinner]. Here we only use the recipes explicitly labeled as one of these three categories ($n = 3893$) and not the full dataset analyzed above where we are unsure of the type of recipe involved.

The starting point for the planning algorithm (i.e. the recipes of each type and health score) is given in Table 8, highlighting just how difficult it is to find plans consisting of recipes with high health scores. For example, in the subset of recipes available to the planner there is a particular lack of recipes achieving a WHO score $\geq$6 and not a single dinner meets 6 or more criteria. More recipes achieve the highest FSA scores, but this remains a very low percentage of the recipes overall.

As we wanted to test the approach generally and not for specific users, instead of calculating specific target nutritional intakes for each user as was done in [11], we applied the WHO and FSA scores used above as evaluation criteria. We moreover applied an additional WHO recommendation, which recommends a healthy typical daily diet should consist of at least 2000 kCal per day with approximately 20% of these coming from snacks and drinks [1]. Thus for a meal plan to be valid it needs to consist of at least 1600 kCal. The results of the full-search over all 1,551,288,123 combinations revealed 141,259,632 meal plans containing at least 1600 kCal. The results are presented in Table 9.

One extremely positive finding is that the planning approach increases the number of options available meeting all 7 WHO criteria. Whereas only 4 recipes meet all 7 criteria individually (see Table 8), the search uncovered over 339 times as many (1358) meal plans with a WHO score of 7, which is more than 17 times the number of individual recipes with a maximum WHO score in the entire collection (see Table 2). Over 27,000 plans received a WHO score of 6 compared to 55 recipes individually. Thus, it seems that, when considering the WHO metric, the meal plan approach offers an ampli-

[4] For space reasons we only present the results for the FSA post-filtering function, but experiments confirm the same trends for the WHO post-filter.

Table 8: Distributions of recipes in the breakfast, lunch and dinner categories (All) and at the same time in the "healthy" category (Healthy).

	Total (Percentage)							Total (Percentage)					
	Breakfast		Lunch		Dinner			Breakfast		Lunch		Dinner	
WHO score	(All) n =2167	(Healthy) n =214	(All) n =693	(Healthy) n =50	(All) n =1033	(Healthy) n =45	FSA score	(All) n =2167	(Healthy) n =214	(All) n =693	(Healthy) n =50	(All) n =1033	(Healthy) n =45
0	91 (.04)	0 (.00)	29 (.04)	0 (.00)	99 (.10)	0 (.00)	4	44 (.02)	20 (.09)	29 (.04)	14 (.28)	6 (.01)	3 (.07)
1	829 (.38)	1 (.00)	259 (.37)	2 (.04)	577 (.56)	8 (.18)	5	80 (.04)	17 (.08)	63 (.09)	13 (.26)	35 (.03)	12 (.27)
2	527 (.24)	21 (.10)	196 (.28)	5 (.10)	239 (.23)	6 (.13)	6	348 (.16)	138 (.64)	119 (.17)	15 (.30)	96 (.09)	16 (.36)
3	367 (.17)	80 (.37)	105 (.15)	18 (.36)	81 (.08)	19 (.42)	7	313 (.14)	24 (.11)	121 (.17)	6 (.12)	152 (.15)	11 (.24)
4	215 (.10)	59 (.28)	65 (.09)	10 (.20)	25 (.02)	8 (.18)	8	388 (.18)	11 (.05)	110 (.16)	0 (.00)	204 (.20)	2 (.04)
5	93 (.04)	32 (.15)	25 (.04)	10 (.20)	12 (.01)	4 (.09)	9	489 (.23)	4 (.02)	121 (.17)	2 (.04)	264 (.26)	1 (.02)
6	41 (.02)	19 (.09)	14 (.02)	5 (.10)	0 (.00)	0 (.00)	10	406 (.19)	0 (.00)	113 (.16)	0 (.00)	213 (.21)	0 (.00)
7	4 (.00)	2 (.01)	0 (.00)	0 (.00)	0 (.00)	0 (.00)	11	80 (.04)	0 (.00)	9 (.01)	0 (.00)	23 (.02)	0 (.00)
							12	19 (.01)	0 (.00)	8 (.01)	0 (.00)	40 (.04)	0 (.00)

Table 9: Distributions in respect to WHO and FSA scores of meal plans generated based on all (breakfast, lunch and dinner) recipes (All) and recipes at the same time in the healthy category (Healthy). Only meal plans are presented that meet the 1600kCal per day limit.

	Total (Percentage)			Total (Percentage)	
WHO score	(All) n =141,259,632	(Healthy) n =108	FSA score	(All) n =141,259,632	(Healthy) n =108
0	19,423,450 (.14)	0 (0)	4	0 (0)	0 (0)
1	96,843,099 (.69)	2 (.02)	5	0 (0)	0 (0)
2	21,222,221 (.15)	19 (.18)	6	156 (0)	7 (.06)
3	3,038,201 (.02)	34 (.31)	7	0 (0)	0 (0)
4	572,126 (0)	33 (.31)	8	130,273 (0)	24 (.22)
5	132,020 (0)	16 (.15)	9	35,943 (0)	0 (0)
6	27,157 (0)	3 (.03)	10	32,982,905 (.23)	32 (.3)
7	1358 (0)	1 (.01)	11	79542 (0)	0 (0)
			12	108,030,813 (.76)	45 (.42)

fication function increasing the options open to users. We note that the effect is not replicated with the FSA-metric is applied.

A second clear outcome of the experiment is, however, just how difficult it is to generate healthy meal plans using recipes from Allrecipes.com. The majority of possible plans created (77%) had a WHO score of 1 or less and 71% of plans had an FSA score of 8 or more. Only 1% of plans received a WHO score higher than 3 and over 72% of plans had an FSA score of 10 or more. Moreover, these plans were created without taking any kind of user personalisation into account – filtering combinations by user preferences would restrict the number of possible healthy plans further still.

To establish the effect of healthier recipes on the planning process, we repeated the search process, but restricted the starting set of candidate recipes to the breakfasts, lunches and dinners, which also feature in the "healthy" category ($n = 309$) as described above. These results are also shown in Table 9.

Considering only recipes in the "healthy" pool indeed results in a smaller proportion of plans receiving the poorest WHO (<2) and FSA (>10) scores. However, only a tiny number of plans can be made overall and many of the possible plans are not particularly healthy (>80% have a WHO score $\leq$4 and 72% an FSA score of $\geq$10). This indicates that even if a user were to eat only recipes from the "healthy" category, which we showed to be healthier than the others, it does not necessarily equate with healthy nutrition. We temper this observation by noting that because many of the meals in the "healthy" category contain relatively little energy ($mean =$ 107.83 kCal).

In this section we studied the utility of algorithmically generating daily meal plans as a recommendation strategy. The approach does seem to offer utility as it can increase the options open to users with high WHO scores. The main finding, however, was that generating plans, which meet WHO and FSA criteria is challenging using the Allrecipes.com collection with the majority of the possible combinations only meeting few criteria or none at all. This means that users have little chance of creating healthy plans without support. The task becomes a little easier when the recipes are restricted to those in the healthy category with the proportions of healthy plans increasing. However, this strategy would require an enormous pool of recipes in order to find healthy plans, which meet user food preferences.

6. DISCUSSION

When taken together the main findings from the analyses described above are as follows:

- Only a small percentage of Allrecipes.com recipes can be considered healthy according to WHO and FSA guidelines.
- The "healthiness" of recipes varies across categories, but even recipes in the "healthy recipes" category can be misleading.
- Users are to some extent able to judge how healthy categories will be, but often disagree.
- Interaction data reveals that people are most positive about the unhealthy recipes i.e. the recipes, which do worst according to the nutritional assessment are those bookmarked most often, rated highest, have the most comments and comments with highest sentiment.
- Current state-of-the-art recommender algorithms in general produce unhealthy recommendations. However, when post-filtering and re-ranking recipes according to their healthiness scores (WHO and FSA) in a simple multiplicative manner reveals that healthiness of recommendations from standard algorithms can be improved.
- Combining the recipes into plans is not straightforward. Only a minority of plans meet health guidelines. However, more healthy plans exist than healthy recipes, thus increasing the options open to users.

These findings demonstrate that both of the two main approaches from the literature (the recommendation of individual recipes and the generation of meal plans) offer benefit and should be developed and evaluated further. That being said a common theme across all of the experiments we performed was that the utility of both single item recommendation and meal-planing algorithms is severely limited by the recipe pool available.

Despite including recipes, which can be considered “healthy” according to the criteria published by health bodies, the overall picture painted by the analyses is an unhealthy one. It seems, therefore, that the assumption made regularly in the literature that Internet-sourced recipes can be used for healthy food recommender systems is indeed a dangerous one.

Additional findings worthy of discussion relate to user aspects. We showed that while there was a weak correlation between user “healthiness estimates” for categories, some were judged very inaccurately and considerable disagreement was observed across judges. Moreover, the interaction data suggest that in general Allrecipes.com users are drawn to unhealthy recipes. These findings underline the scale of the challenge of algorithmically deriving healthy food choices, which users will actually like and eat.

Limitations, unanswered questions and future research.

An important thing to bear in mind when interpreting our results is that they relate only to Internet sourced recipes from one site albeit the largest food portal on the Internet - Allrecipes.com. The site is primarily used by users from the United States and repeating the analyses with data from sites hosted in other countries may result in different outcomes. We plan to source other datasets and repeat our analyses.

Our analyses showed that algorithmic solutions to single-item recommendation and meal planning offer potential benefit and should be further examined. In terms of the trade-off between accuracy and healthiness our experiments barely scratched the surface of what can be explored and a thorough algorithmic evaluation is necessary. It would also be interesting to perform user studies to establish exactly when users notice that recommendation accuracy is being sacrificed in favour of healthiness. Algorithmically, improving prediction accuracy, for example, by incorporating context or category information would improve results overall. Similarly, rather than optimizing for general healthiness metrics generally as we have done here, it might be interesting to optimize for specific macro-nutrients because users can have special dietary needs.

In our meal plan experiments we learned that meal plans with 3 “healthy” recipes were restricted due to the low energy content of these meals. Future work could consider more complicated meal combinations to see if larger number of smaller, healthier recipes could be an effective means of solving this problem. This becomes a complex algorithmic problem as the number of combinations becomes extremely large.

Moreover, there is much to learn regarding user perception of recipe healthiness and how this relates to the way recipes are presented. If users are not able to distinguish between healthy and unhealthy recipes then why do they interact more with unhealthy ones? Perhaps there are other biases in the way these recipes are presented, organised or accessed, which leads to this outcome. These are all aspects, which should be investigated in the future.

Finally, as the recipe collection seems to be a bottleneck in terms of achievable health scores researchers may want to think of ways to address this issue. Perhaps by helping users to publish more healthy recipes via ingredient substitution suggestions [8] when recipes are uploaded or automatically generating healthier versions of recipes as alternatives [24].

7. REFERENCES

[1] Fsa nutrient and food based guidelines for uk institutions. available at http://www.food.gov.uk/sites/default/files/multimedia/pdfs/nutrientinstitution.pdf. last accessed on 20.6.2016. 2007.

[2] Usda. cook more often at home. available at http://www.choosemyplate.gov/weight-management-calories/weight-management/better-choices/cook-home.html. last accessed on 20.6.2016. 2011.

[3] Fsa. guide to creating a front of pack (fop) nutrition label for pre-packed products sold through retail outlets. available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/300886/2902158_FoP_Nutrition_2014.pdf. last accessed on 27.6.2016. 2014.

[4] Allrecipe.com press report. available at http://press.allrecipes.com/. last accessed on 20.6.2016. 2016.

[5] Allrecipe.co.uk press report. available at http://allrecipes.co.uk/news.aspx. last accessed on 20.6.2016. 2016.

[6] Esha, nutrition labeling software. available at http://www.esha.com/. last accessed on 20.6.2016. 2016.

[7] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through twitter. In *Proc. of CHI'15*.

[8] P. Achananuparp and I. Weber. Extracting food substitutes from food diary via distributional similarity. *arXiv preprint arXiv:1607.08807*, 2016.

[9] M. De Choudhury and S. S. Sharma. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proc. of CSCW '16*.

[10] T. De Pessemier, S. Dooms, and L. Martens. A food recommender for patients in a care facility. In *Proc. of RecSys'13*, pages 209–212. ACM.

[11] D. Elsweiler and M. Harvey. Towards automatic meal plan recommendations for balanced nutrition. In *Proc. of RecSys'15*, pages 313–316. ACM.

[12] D. Elsweiler, M. Harvey, B. Ludwig, and A. Said. Bringing the "healthy" into food recommenders. In *Proc. of DRMS'15.*, pages 33–36.

[13] J. Freyne and S. Berkovsky. Recommending food: Reasoning on recipes and ingredients. In *Proc. of UMAP'10*, pages 381–386.

[14] M. Ge, F. Ricci, and D. Massimo. Health-aware food recommender system. In *Proc. of RecSys '15*, pages 333–334.

[15] T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. 2002.

[16] F. Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[17] M. Harvey, B. Ludwig, and D. Elsweiler. Learning user tastes: a first step to generating healthy meal plans? In *Proc. of LIFESTYLE'12*, page 18.

[18] S. Howard, J. Adams, M. White, et al. Nutritional content of supermarket ready meals and recipes by television chefs in the united kingdom: cross sectional study. *BMJ*, 345, 2012.

[19] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Proc. of ICDM'08*, pages 263–272. Ieee.

[20] C. Kim and J. Kim. A recommendation algorithm using multi-level association rules. In *Proc. of WI'03*, pages 524–527. IEEE.

[21] K.-J. Kim and C.-H. Chung. Tell me what you eat, and i will tell you where you come from: A data science approach for global recipe data on the web. *IEEE Access*, 4:8199–8211, 2016.

[22] T. Kusmierczyk, C. Trattner, and K. Nørvåg. Temporal patterns in online food innovation. In *Proc. of WWW'15 Companion*.

[23] T. Kusmierczyk, C. Trattner, and K. Nørvåg. Temporality in online food recipe consumption and production. In *Proc. of WWW'15*.

[24] T. Kusmierczyk, C. Trattner, and K. Nørvåg. Understanding and predicting online food recipe production patterns. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, pages 243–248. ACM, 2016.

[25] S. Larrain, C. Trattner, D. Parra, E. Graells-Garrido, and K. Nørvåg. Good times bad times: A study on recency effects in collaborative filtering for social tagging. In *Proc. of RecSys'15*, pages 269–272.

[26] Y. Mejova, H. Haddadi, A. Noulas, and I. Weber. # foodporn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 51–58. ACM, 2015.

[27] X. Ning and G. Karypis. Slim: Sparse linear methods for top-n recommender systems. In *Proc. of ICDM'11*, pages 497–506. IEEE.

[28] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proc. of UIAI'09*, pages 452–461. AUAI Press.

[29] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[30] M. Rokicki, E. Herder, and E. Demidova. What's on my plate: Towards recommending recipe variations for diabetes patients. *Proc. of UMAP'15 LBRS*, 2015.

[31] M. Rokicki, E. Herder, T. Kusmierczyk, and C. Trattner. Plate and prejudice: Gender differences in online cooking. In *Proc. of UMAP'16*, pages 207–215.

[32] G. Sacks, M. Rayner, and B. Swinburn. Impact of front-of-pack 'traffic-light'nutrition labelling on consumer food purchases in the uk. *Health promotion international*, 24(4):344–352, 2009.

[33] A. Said and A. Bellogín. You are what you eat! tracking health through recipe interactions. In *Proc. of RSWeb'14*.

[34] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proc. of WWW'01*, pages 285–295. ACM.

[35] E. P. Schneider, E. E. McGovern, C. L. Lynch, and L. S. Brown. Do food blogs serve as a source of nutritionally balanced recipes? an analysis of 6 popular food blogs. *Journal of nutrition education and behavior*, 45(6):696–700, 2013.

[36] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic. Recipe recommendation using ingredient networks. In *Proc. of WebSci'12*, pages 298–307.

[37] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *JASIST*, 63(1):163–173, 2012.

[38] C. Trattner and D. Elsweiler. Estimating the heathiness of internet recipes: A cross sectional study. *Frontiers in Public Health*, 2017.

[39] C. Trattner, D. Kowald, P. Seitlinger, T. Ley, and S. Kopeinik. Modeling activation processes in human memory to predict the use of tags in social bookmarking systems. *J. Web Science*, 2(1):1–16, 2016.

[40] C. Trattner, T. Kusmierczyk, and K. Nørvåg. FOODWEB - studying food consumption and production patterns on the web. *ERCIM News*, 2016(104), 2016.

[41] C. Wagner and L. M. Aiello. Men eat on mars, women on venus? an empirical study of food-images. In *Proc. of WebSci'15 Posters*.

[42] C. Wagner, P. Singer, and M. Strohmaier. The nature and evolution of online food preferences. *EPJ Data Science*, 3(1):1–22, 2014.

[43] R. West, R. W. White, and E. Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proc. of WWW'13*, pages 1399–1410.

[44] J. Who and F. E. Consultation. Diet, nutrition and the prevention of chronic diseases. *World Health Organ Tech Rep Ser*, 916(i-viii), 2003.