You are What You Eat! Tracking Health Through Recipe Interactions

Alan Said
TU-Delft
The Netherlands
alansaid@acm.org

Alejandro Bellogín
Universidad Autónoma de Madrid
Spain
alejandro.bellogin@uam.es

ABSTRACT

On today's World Wide Web, social recommender systems have become a commodity regardless of application domain. Even tangible items such as food and clothes have become social. Together with a seemingly endless amount of personalization and recommender systems ranging from movies, music, or consumer products, *recipe recommender systems* are attracting many users looking for inspiration on the next thing to purchase or cook. There is however a conceptual difference between recommending consumer goods for leisure and entertainment, and recommending food. What people eat has a *direct effect on their health*, an aspect commonly overlooked in the context of recommendation.

In this work, we present an early analysis of users' interactions with recipes (ratings) on the online social network Allrecipes.com. We compare the interaction patterns of users from locations known to have poor health to users from locations known to have good health in order to identify whether there is an observable difference between the two populations.

Our results point to a statistically significant difference between the healthy and unhealthy groups, a difference that could potentially be used to create health-conscious, personalized, recommendation services to aid people in their daily lives.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services - Commercial Services; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - Information filtering; H.1.2 [**Models and Principles**]: User/Machine Systems - Human Factors; K.4.1 [**Computers and Society**]: Public Policy Issues - Computer-related Health Issues

General Terms

Human Factors; Experimentation; Design

Keywords

Personalization; Food Recommendation; Health; Human-Data Interaction; Recommender Systems; Persuasion; Social Web

Proceedings of the 6th Workshop on Recommender Systems and the Social Web (RSWeb 2014), collocated with ACM RecSys 2014, 10/06/2014, Foster City, CA, USA.
.

1. INTRODUCTION

Today, Internet users turn to the Web for help with the planning and selection of many daily tasks; whether what music to listen to (Spotify), what consumer products to purchase (Amazon), what movies to watch (Netflix), or what food to prepare (Allrecipes). Consumers put a considerable amount of trust into systems which are able to simplify their information needs, no matter the type of information (or products) sought for. Often, these online services implement persuasion systems telling the users to buy, listen to, watch, or even eat items or products that their peers have interacted with. It should however be noted that there is a distinct conceptual difference in recommending a piece of information to be consumed online, e.g. a news article or a song, and a tangible object, e.g. a computer or a car. Among the differences between the types of objects, we find aspects such as consumption cost (in terms of money, time, effort), the expected longevity of a product (a music track lasting a few minutes, a book lasting a week, a car lasting several years), etc. These aspects need to be accounted for when creating a personalized experience, whether for an *online* consumption case, or for a *real-world* product.

In turn, when recommending food and recipes, there is an additional dimension of the recommendation that needs to be considered: the health aspect of what is being recommended to a specific user. A personalization system which has a (more or less) direct effect on the user's daily life and health, such as a recipe recommender, needs to be aware of the potential outcome of the recommendation, not only in terms of increased business value for the vendor and the general utility as experienced by the consumer, but also of the well-being of the consuming user.

It is because of the above stated aspect that we, in this paper, focus on health aspects involved in personalizing users' experiences in a food-related online social network. We do so by taking into account the general health in the area where the user lives. By using data from County Health Rankings & Roadmaps[1] in combination with data from the recipe-focused online social network Allrecipes[2] we are able to show that there is a significant difference in consumption patterns between users from counties with a high health ranking and users from counties with a low health ranking. Our motivation is that these differences can be used to identify users with higher health risks, even in cases where the geographical location is not known.

The main contribution of our work is to show a significant correlation between recipe usage on an online social network and the reported health in users' geographic locations.

[1] www.countyhealthrankings.org
[2] www.allrecipes.com

2. RELATED WORK

Over the last decade, a massive body of work on multimedia recommender systems has been accumulated, e.g. movies [1] music [4], online news [9], and practically any other type of consumer products [2]. Food recommendation on the other hand, which also has been an online phenomenon for a long time, has only recently started gaining attraction from information system and personalization researchers and practitioners, e.g. improving the food preparation competence of cooks [14], dinner planning for groups [3], educating potential cooks on healthy foods [7] or diversifying the meals served in care facilities [5].

When personalizing the culinary experience, it is important to be aware of the conceptual difference between recommending a movie to watch or a song to listen to, compared to recommending a dish to eat or cook. The movies one watches and songs one listens to have no direct effect on the health of the subject receiving recommendations. Recommending food on the other hand, as mentioned in Section 1, means that the recommendation will indeed have an effect on the user's health, either by simply proposing the user to eat something unhealthy directly, or, by attempting to altering a user's (long term) food habits – which might remain even after the user is no longer using the service. However, there exists only a limited body of work on food recommendation and personalization from a health-oriented aspect, e.g. Hsiao and Chang [12] show that by aiding in planning meals it is possible to improve the health of a system's users. Some research approaches food recommendation from the perspective of diet and exercise [8], attempting to understand the users' reasoning around recipes. More recently, Harvey et al. [10, 11] reported on a study attempting to identify the factors that affect the ratings given to recipes in order to leverage this information in a recipe recommender system able to recommend recipes which are not only nutritional, but also well-liked by the users.

In this work, we base our finding on geographical areas with good or bad health, inspired by the line of research known as *Health Geography* [13]. Here, Dummer showed that "Geography and health are intrinsically linked" [6]. With this in mind, we attempt to find whether it is possible to use concepts from information management and human-computer interaction to alleviate potential health effects in online recommendation services even when the location of the user is not known.

3. RECIPES & HEALTH DATA

To perform our analysis, we scraped the recipe-related social network Allrecipes.com. In this process, we collected user profiles, recipes, ingredients, *recipe boxes* (users collect and rate their recipes in virtual recipe boxes making them easily accessible at later points in time), social connections, and demographic information on users (location, interests, hobbies, etc.). This data collection[3] was performed during October 2013, and resulted in a dataset containing information on more than 170 thousand users, 54 thousand recipes, $8,400$ ingredients, and 17 million recipe box assignments (which we refer to as ratings[4]).

Having collected the data, we used health rankings by county from County Health Rankings to identify users living in healthy and unhealthy counties. Our health focus was specifically on obesity, i.e. the percentage of adults suffering from obesity in each county.

[3]The scripts used to scrape the data from the Allrecipes website are available at `github.com/alansaid/RecipeCrawler`

[4]Even though users can rate the recipes they put in their recipe boxes (if they wish), in the scope of this paper we have only analyzed the binary relationships between users and recipes.

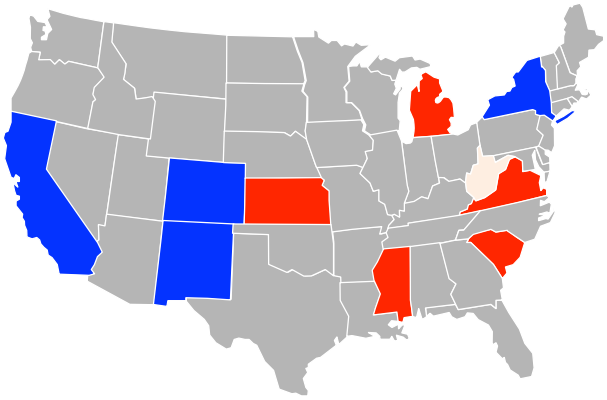

Figure 1: Map showing the US states where the analyzed counties lie. Blue counties indicate low adult obesity, red counties indicate high obesity. Note that two of the counties (Boulder, La Plata) with the lowest obesity are in Colorado, thus the figure only shows four blue states.

The health ranking dataset contains data for more than $3,400$ US counties, including the percentage of obese adults.

The dataset collected from Allrecipes does not contain the counties where users live in. In order to connect users to counties, we used a mapping of $42,000$ US cities to $3,200$ US counties[5]. This allowed us to link the recipe and health datasets to each other. It should be noted that users of the Allrecipes social network do not have to state their hometown, and when they choose to do so, this is done in free text. The implication is that it is not possible to automatically map all users to counties, e.g. some users state made up cities, or local slang names (Chicagoland for Chicago, The Big Apple for New York, etc.), or simply misspell the name of their hometown. Additionally, large cities (e.g. Dallas, TX) may be composed of several counties, making the mapping of these cities onto distinct counties problematic unless additional information is available or manual mapping is performed. Furthermore, the counties in the county health ranking dataset and the city-to-county mapping dataset do not overlap perfectly, as noted above the county health data contains $3,400$ counties whereas the county mapping data contains $3,200$. However, with some manual tuning (replacing e.g. Hollywoodland with Hollywood, The Big Apple with New York City, etc.) we were able to infer the counties for the majority of the users.

4. MAPPING UNHEALTHY INGREDIENTS TO HEALTH DATA

In order to analyze whether it is indeed possible to use the county health ranking data in combination with food-oriented websites, e.g. Allrecipes, we focused on a relatively small number of healthy and unhealthy counties.

As a first step, we identified how often a certain ingredient is used by users in a certain county. This was accomplished by mapping each recipe onto its composing ingredients, and correspondingly mapping all ratings given by users (per county) on the recipes onto the ingredients of the recipes. This process war repeated for the one hundred and ten most used ingredients in each county. Following this, we calculated the percentage of how often an ingredient was used in average in the counties with low obesity and high obesity separately. This information allowed us to identify the five

[5]`www.farinspace.com/us-cities-and-state-sql-dump`

Table 1: The counties used in the analysis and the data available for each county, the top five (Table 1a) are counties with the lowest percentage of adults suffering from obesity, the bottom five (Table 1b) are counties with the highest percentage of adults suffering from obesity. Note that there are many power users with several hundred to several thousand rated recipes in their recipe boxes. Also note that the total number of recipes has been excluded as the individual recipes are not distinct across rows.

(a) Statistics for counties with low obesity percentage.

State	County	Adult obesity	Users	Ratings	Recipes
New Mexico	Santa Fe	14%	26	3009	2721
Colorado	Boulder	15%	99	9938	6614
New York	New York	15%	384	32468	14118
California	Marin	15%	12	570	537
Colorado	La Plata	16%	16	2439	2069
Total			537	48424	

(b) Statistics for counties with high obesity percentage.

State	County	Adult obesity	Users	Ratings	Recipes
Mississippi	Lowndes	37%	11	827	783
Kansas	Wyandotte	38%	49	6924	5235
South Carolina	Berkeley	38%	159	12637	7539
Virginia	Portsmouth	39%	18	1512	1400
Michigan	Saginaw	40%	33	1315	1224
Total			149	46430	

Table 2: The twenty most commonly used ingredients and their popularity as a percentage of how often they appear in counties with high (↑) and low (↓) obesity. The ingredients are sorted by the percentage of times they appear in recipes stored by cooks in counties with high obesity. Note, for instance, the difference between usage of olive oil and garlic vs. dairy products (milk, cheddar and cream cheeses) between the county types.

No.		Salt	Butter	Sugar	Eggs	Flour	Onions	Garlic	Water	Pepper	Milk
1-10	↑ **Obesity**	51.04%	33.72%	30.67%	27.25%	26.14%	23.93%	22.79%	21.96%	20.65%	14.96%
	↓ **Obesity**	55.30%	32.92%	31.01%	26.77%	25.68%	24.86%	27.31%	21.54%	21.42%	13.23%
No.		Vanilla	Olive Oil	Brown Sugar	Chicken	Cinnamon	Parmesan	Baking Soda	Veg. Oil	Cheddar Ch.	Cream Ch.
11-20	↑ **Obesity**	14.85%	14.07%	12.54%	10.20%	9.81%	7.96%	7.89%	7.29%	6.81%	6.79%
	↓ **Obesity**	14.52%	18.04%	12.56%	8.70%	10.00%	8.25%	8.75%	7.41%	5.35%	5.21%

most obese and five least obese counties with available ingredient data. Due to the mapping procedure and dataset described in the previous section, the five counties with the lowest percentage of obese adults selected were within the top 15 of the least obese counties. Similarly, the counties with the highest percentage of obesity were within the top 100 of the most obese counties. The top counties together with statistics for each are shown in Table 1.

It should be noted that the geographic distribution of the counties is not limited to an isolated geographical location within the US, instead the counties are spread throughout the country, as shown in Fig. 1. This should further strengthen the health aspect of the analysis, while minimizing potential effects of local food trends found in isolated geographical locations [13].

5. ANALYSIS & RESULTS

For each group of counties, i.e. with high and low obesity percentage, we identified the top 110 most popularly used ingredients in both types of counties, i.e. the top intersecting ingredients used by users in both types of counties. Table 2 shows the 20 ingredients used most often in counties with high (↑) obesity and the corresponding percentage in counties with low (↓) obesity. Having this information, we performed a statistical significance analysis (t-test) on the vectors containing the percentages of how often the ingredients were used in both type of counties (the same ingredients appearing in the same places in both vectors). The justification of this is that, if the ingredients were in fact used differently in the two types of counties, we should be able to distinguish between high and low risk users independent of their geographical location. Thus ensuring that high/low-risk users can be identified by their online recipe interaction patterns.

The obtained p-value from the t-test ($p < 0.05$) confirms that the ingredient usage in counties with high obesity is in fact different from that of counties with low obesity. The implication of this is that high-risk/low-risk users can be identified simply by their recipe interactions in an online social network. This information can in turn be used to personalize a food recommendation system based on the recorded interactions of a user.

6. DISCUSSION

In the previous sections, we have described our analysis of a health-related dataset and an analysis of a real-world recipe-focused online social network. Our results point to that it is possible to identify users from high-risk (poor health) areas just from their recipe interactions. This suggests that, should a recommendation system be employed, it can be tailored to not only provide high-quality recipes to the user, but also take into consideration the potential health aspects of the user. The health effects can be mitigated by either filtering out recipes which can be deemed unhealthy, or to create personalized recipes – by altering the doses of certain ingredients – and still fulfilling the users' expectations. This needs however be done in such a way as to not lower the usability and quality of the system, as perceived by the user. A personalization approach of this type would serve as an insurance that the service would not be the cause of, or aiding to, any detrimental effects on the users'

health. Given the increasing quality of recommender systems, a system being conscious of the (inferred) health of its users appears as a plausible next step.

We are aware of the limits of our analysis, e.g. only analyzing the binary connections between a recipe and an ingredient – not taking into consideration the amount of the ingredient used. Nevertheless, we believe our results to be indicative of what can be attained when using the ingredient amount as well. This is currently the focus of our ongoing work, however, the ambiguous and non-standardized unit and ingredient declaration in recipes, e.g. *one cucumber*, *half a cup* of sugar, *one glass* of water, *two crackers*, etc., makes this a non-trivial task.

It should be noted that the results obtained in our analysis are the result of early work, we do however believe that this is a feasible approach to proactively care for the users of similar food- or otherwise health-oriented services. As mentioned in Section 1, there is a conceptual difference between recommending an entertainment-focused item (song, movie) compared to domains where the personalization system has a direct effect on the user's health.

7. CONCLUSION & FUTURE WORK

In this work, we have analyzed a recipe dataset and combined it with data reporting health aspects in US counties. We have identified counties that suffer from poor health (large percentage of adults suffering from obesity) and found that there exist statistically significant differences in how users from poor health counties interact with recipes compared to users from counties with good health (low percentage of adults suffering from obesity). Our work suggests a potential approach to health-oriented recommender systems which takes into account the possible adverse effects on a user, based on demographic information as well as through information on the recorded interactions (ratings) with the system.

As for future (and current) work paths, we are currently investigating whether there are other user-related features that also correlate to health aspects, e.g. inferring health through stated interests and hobbies. Similarly, we intend to investigate whether the social ties (follower/followee relationships) between users, a concept that has been proven to be useful in personalization and recommendation approaches in other domains, hold similar health-related information. Additionally, we plan to study whether the nutritional aspects of ingredients can help in identifying health-oriented aspects in individual users.

8. ACKNOWLEDGMENTS

This work was in part carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no.246016.

The authors would like to thank Arjen P. de Vries and Jacco van Ossenbruggen from CWI for feedback during the work resulting in this paper.

9. REFERENCES

[1] X. Amatriain and J. Basilico. Netflix recommendations: Beyond the 5 stars (part 1) – the netflix tech blog. `http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html` (retrieved May 12, 2012), April 2012.

[2] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

[3] S. Berkovsky and J. Freyne. Group-based recipe recommendations: Analysis of data aggregation strategies. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 111–118, New York, NY, USA, 2010. ACM.

[4] Ò. Celma. *Music Recommendation and Discovery in the Long Tail*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2008.

[5] T. De Pessemier, S. Dooms, and L. Martens. A food recommender for patients in a care facility. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 209–212, New York, NY, USA, 2013. ACM.

[6] T. J. Dummer. Health geography: supporting public health policy and planning. *Canadian Medical Association Journal*, 178(9):1177–1180, 2008.

[7] J. Freyne and S. Berkovsky. Intelligent food planning: Personalized recipe recommendation. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, pages 321–324, New York, NY, USA, 2010. ACM.

[8] J. Freyne, S. Berkovsky, and G. Smith. Recipe recommendation: Accuracy and reasoning. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, UMAP'11, pages 99–110, Berlin, Heidelberg, 2011. Springer-Verlag.

[9] F. Garcin and B. Faltings. Pen recsys: A personalized news recommender systems framework. In *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge*, NRS '13, pages 3–9, New York, NY, USA, 2013. ACM.

[10] M. Harvey, B. Ludwig, and D. Elsweiler. Learning user tastes: a first step to generating healthy meal plans? In *Proceedings of the ECIR Workshop on Searching4Fun*, Searching4Fun '12', 2012.

[11] M. Harvey, B. Ludwig, and D. Elsweiler. You are what you eat: Learning user tastes for rating prediction. In *Proceedings of the 20th International Symposium on String Processing and Information Retrieval*, SPIRE, pages 153–164. Springer, 2013.

[12] J.-H. Hsiao and H. Chang. Smartdiet: A personal diet consultant for healthy meal planning. In *Proceedings of the 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems*, CBMS '10, pages 421–425, Washington, DC, USA, 2010. IEEE Computer Society.

[13] G. Moon. Health geography. In R. Kitchin and N. Thrift, editors, *International Encyclopedia of Human Geography*, volume 5, pages 35–55. Elsevier, July 2009.

[14] J. Wagner, G. Geleijnse, and A. van Halteren. Guidance and support for healthy food preparation in an augmented kitchen. In *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, CaRR '11, pages 47–50, New York, NY, USA, 2011. ACM.