Temporal Patterns in Online Food Innovation

Tomasz Kusmierczyk
NTNU
Trondheim, Norway
tomaszku@idi.ntnu.no

Christoph Trattner
NTNU
Trondheim, Norway
chritrat@idi.ntnu.no

Kjetil Nørvåg
NTNU
Trondheim, Norway
noervaag@idi.ntnu.no

ABSTRACT

Since innovation plays an important role in the context of food, as evident in how successful chefs, restaurants or cuisines in general evolve over time, we were interested in exploring this dimension from a more virtual perspective. In particular, the paper presents results of a study that was conducted in the context of a large-scale German online food community forum to explore another important dimension of online food recipe production, namely known as online food innovation. The study shows interesting findings and temporal patterns in terms of how online food recipe innovation takes place.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

online food recipes; temporal patterns; innovation

1. INTRODUCTION

Food community Websites have gained tremendously in popularity over the past few years with millions of users accessing and creating new innovative recipes every day. Although recent research in this area has shown how people consume content from those community platforms, to find important correlations between, e.g., food consumption and health related issues [5, 10], little is known from the perspective of the producer. To fill this gap in the literature, we have recently started a research project that tries to deal with that issue by analyzing not only online food recipe consumption but also production patterns in time [4], since we believe that online food consumption is heavily influenced by what we as a community produce and share with others.

In addition to a high correlation between online food recipe consumption and production patterns, we also found that recipe production is heavily influenced by time and there are significant differences over the course of the week and year. Furthermore, we found that not all types of recipes are similarly interesting to the users over time. For instance, we could identify that vegetarian recipes stay significantly less interesting in time than, for example, allergy free recipes. In this paper, we study another important dimension of online food recipe production known as "food innovation". Similar to our previous study, we put a particular emphasis on the temporal dynamics and study the effect in the context a large online food community Website.

Findings. Based on an extensive set of experiments we find a number of interesting patterns:

- Although the number of known ingredients remains relatively low and constant, the community is able to continuously combine them to form a number of new and innovative recipes over time. Hence, after an initial phase where the innovation factor is decreasing we find that the innovation factor is not only stabilizing at a surprisingly high level but is slowly growing in time.
- The food innovation factor depends on the season of the year and to some smaller extent also on the day of the week.
- We find significant differences in terms of the innovation factor between recipes from different categories that cannot be explained by the number of recipes produced in the category.
- The temporal profiles of the users meaningfully vary, e.g., some users are more successful in innovating over time than others.
- Geographical origin is the most important factor, significantly more than age, gender or number of friends to drive the users' innovation factors.

Contributions. Overall, our contributions can be summarized as follows:

- The mining, introduction and provision of a novel large-scale dataset to study online food recipe consumption and production patterns.
- The in-depth study and a set of interesting findings on the temporal dynamics in online food innovation.

The rest of the paper is structured as follows: In Section 2 we highlight relevant related work in the field. Section 3 gives an introduction to our dataset and the crawling

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741700.

methodology we applied to obtain the data. Section 4 introduces the methodology we choose to analyze the data. Section 5 presents the results of our experiments, and finally Section 6 concludes the paper and gives directions for future work.

2. RELATED WORK

Studying online food recipe consumption and production patterns is a relatively new strand of research and only a few related studies exist so far in this context. While previous research in this area was mostly conducted on a small scale, through, e.g., online or telephone surveys [7], computational approaches follow a more radical approach applying statistical and data-mining techniques to sources such as online community platforms or social media, in order to study these patterns on a larger scale.

The first significant large-scale research effort in this context was a study conducted by Ahn et al. [2]. In their work they mined and analyzed a large-scale online food community platform with over 10^6 community created-recipes called cookpad.com[1] to unveil patterns on how recipes are created in a global sense in terms of their flavor. Among other things they found significant difference among countries in how recipes in terms of their ingredients are created. An interesting follow-up study was then performed by Teng et al. [6] in the context of the online food recipe platform allrecipes.com[2], applying a similar approach as Ahn et al. to train a statistical model that is able to recommend recipes to users (see also [8] for a recently published related study in the context of recommender systems).

There are also computational approaches to study patterns in terms of how we consume food online. The most prominent work in this context is a study conducted by West et al. [10]. In their work, they analyzed log-files of users accessing recipes in the online food community website allrecipes.com. Among other things, they find significant seasonal trends in what people prefer over the course over the year and in regions in the US. Furthermore, they find a correlation between online food consumption patterns and heart disease in certain regions of the US. An interesting follow-up study was the work of Wagner et al. [9] who investigated the dynamics of online food consumption in an online food community forum located in Europe. Similar to the findings of West et al., they find clear temporal patterns in how online food recipes are consumed over the course of a week and in terms of the people's locality, highlighting significant differences between the US and the European community. Similar to the work previous mentioned is the study of Said and Bellogin [5], who showed a significant correlation between obesity cases in the US and online food consumption and recipe production in the context of the online food community website allrecipes.com; a similar result that was also recently achieved by Abbar et al. in the context of Twitter [1].

Despite all these previous efforts, research on studying online food recipe consumption and production patterns is still at an early stage. Especially on the producer side there is still much to explore, which was also the motivation for our recent research effort looking at the temporal dynamics of online food recipe consumption and production [4]. Contrary to all previously mentioned works, the aim of this study is to shed light on the temporal dynamics in online food innovation – a dimension that has to the best of our knowledge not been studied yet.

[1]http://cookpad.com

[2]http://allrecipes.com

3. DATASET

Our work relies on a dataset obtained from the German online food community website kochbar.de[3]; one of the largest of its kind in Europe. The dataset was collected via a simple Web-crawling approach between 2014-11-22 and 2014-12-05, and is available online[4]. The dataset covers more than 400 thousand recipes published in the years 2008-2014 (170 recipes are published on average per day). For each recipe, information about ingredients and preparation is provided. Recipes are labeled with about 230 categories of 7 classes (here we focus on 4) and were rated by almost 200 thousand users providing 7 million ratings (3300 ratings on average per day). Among these 200 thousand users, almost 5 thousand are active producers and published 10 or more recipes. The reason for choosing the kochbar.de community platform over other popular alternatives such as allrecipes.com are manyfold, but mainly stem from the observation that kochbar.de features a richer set of meta-data than other popular and related online food communities. For instance, there is not only detailed information available about what types of ingredients are used or how many calories, fat, carbohydrates, etc. a recipe has, but also detailed user profiles with explicit friendship relations, group information, rating information, and so on.

To investigate our dataset in terms of online food recipe innovation, several pre-processing steps were necessary to clean the data. Since we rely on the ingredients used by the users in their recipes as the main entity to calculate innovation, the following procedure was applied: in kochbar.de ingredients of recipes are lists of arbitrary strings which are defined as free-form text by the users. As a consequence, word variants, misspellings, etc. lead to a disambiguation problem. Hence, as a first step, we split conjunctions such as 'salt and pepper'. Then, each alternative of two ingredients was replaced with the name of the more popular option, e.g., 'butter or margarine' was replaced with 'butter'. After that, we filtered out stop-words, special characters, amounts and units, and words describing the preparation process, e.g., 'cooked', 'washed'. Finally, we replaced ingredients occurring less than 200 times with more popular variants. To each of such ingredient names we tried to match with others starting from the most popular ones, e.g., the ingredient 'the glass of salted water' was replaced with 'water' and 'salt'. Sometimes no variants were matched. In this special case, ingredient names occurring less than 100 times were simply discarded. This procedure reduced the initial number of over 334 thousand ingredients in our dataset to 2208.

4. METHODOLOGY

In this section, we describe our experimental methodology, i.e., the metrics used, the analyses performed and the experiments conducted.

[3]http://kochbar.de

[4]Download link available on request.

4.1 Studying Community Patterns

To investigate the temporal patterns in terms of innovation from the perspective of the community we rely on two different measures. The first measure we employ is entropy, which allows us to study from information-theoretic point of view how the community evolves in terms of online food innovation. The second measure we use, is a similarity-based measure that tries to capture what is actually new in the community at each particular point of time.

The *entropy* is a measure to capture the complexity of a distribution in terms of random variable X, defined as [3]:

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

Entropy expresses the amount of information required to remember the signal, i.e., average number of bits needed to encode samples drawn from X. It increases with the number of possible values x and when the distribution becomes more uniform. We apply entropy to measure the complexity of distribution of ingredients i in the system.

The measure expressing the amount of information needed to represent one variable Y knowing already the value of the other variable X is known as *conditional entropy*. It is defined as [3]:

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) log_2 p(y|x)$$

Conditional entropy is a relative measure which tells us how many bits we need on average to encode a sample of Y knowing value of X. $H(Y|X) = 0$ means that Y is fully determined by X. $H(Y|X) = H(Y)$ on the other hand means that Y is independent from X. Similarly, *mutual information* [3]:

$$I(X,Y) = H(Y) - H(Y|X)$$

is the most popular measure of dependence between two variables, e.g., $I(X,Y) = 0$ means independence. We apply conditional entropy and mutual information to measure how well a single ingredient $i \in r$ determines recipe r, i.e., what amount of the information from r is in i.

Another measure, as already highlighted before, is the *Innovation Factor*. It is defined as follows:

$$IF(r) = 1 - \max_{r' \prec r} sim(r, r')$$

The operator $\prec$ means temporal precedence and the function $sim \in [0,1]$ measures similarity between two recipes. Consequently, the *Innovation Factor* for r returns the difference between the recipe r and the most similar one available before r publication time. Although similarity sim can be measured in an arbitrary way we represent recipes r, r' as combinations of ingredients i and therefore *Jaccard's Index* is the most natural choice:

$$sim(r, r') = JI(r, r') = \frac{|\{i : i \in r \wedge i \in r'\}|}{|\{i : i \in r \vee i \in r'\}|}$$

4.2 Studying User Patterns

In order to study food recipe innovation patterns on the level of the user we rely on the *Innovation Factor* measure. The mean value of the measure in the context of a user u is calculated as:

$$IF(u) = \frac{1}{|recipes(u)|} \sum_{r \in recipes(u)} IF(r)$$

where $recipes(u)$ stands for all recipes published by u. Other statistics such as medians are expressed in an analogous way.

To analyze more extensively trends in the users' innovation factors we model their temporal profiles with a linear regression model. Hence, the time since the user registration expressed in number of weeks is used as the independent variable (horizontal axis) of each data point. Consequently, the Innovation Factor is used as the response variable (vertical axis). The intercept informs then about the initial level of users' innovation and the slope provides information about increase or decrease ratio.

To finally determine the factors related to innovation we apply *Information Gain*. The measure finds weights of attributes X, e.g., $X = user\ age$, basing on their correlation with a special attribute Y (in our case Innovation Factor: $Y = IF$).

5. RESULTS

The following two subsections present the results of the experiments that have been performed on our dataset.

5.1 Community Patterns

General Patterns. Figure 1 presents the results of the evolution of the community in terms of food recipe innovation over time. As depicted, the number of known ingredients (blue line left plot) in our dataset saturates very fast at the final level of around 2000. A similar behavior can be observed for entropy of ingredients. The plot saturates at the level of 9 bits. If the distribution of ingredients was uniform, 11 bits would have been used. The discrepancy of 2 bits represents the degree of diversity in ingredients probabilities (how far the probabilities are from a uniform distribution).

Although the number of ingredients after the initial phase remains almost constant, the number of recipes and the number of ingredients combinations continuously increase in the considered time period. What is more, both curves grow at the same rate and are indistinguishable on the plot. Users seem to be continuously able to combine the limited number of ingredients in an innovative way. This finding is supported also by conditional entropy curve that is growing accordingly and mutual information curve that remains constant after initial phase. On the other hand, growth ratio is decreasing and both curves are saturating. If the trend will be preserved in the future, it seems that the number of combinations will not exceed 0.5 million (about 19 bits) and the community in terms of users' activity will be degrading.

Similar to the first analysis, we analyzed the temporal dynamics of the community exploring the Innovation Factor measure. The results of this analysis are presented in Figure 2. Two phases of community development can be observed. First is the initial phase with a strong decline in the food innovation factor – lasting for around two years (up to the year 2010). Then the second phase lasting for almost 5 years. Although the observed differences in the slope are much smaller than in the initial phase, a steady increase in terms of food innovative is noticeable; both means and medians grow. Figure 3 presents the distribution of innovation of recipes published after 2010-01-01 (the second phase). Two facts can be observed. First, the distribution is significantly shifted toward high values. Mean and median are surprisingly high showing high motivation of users to be innovative. Second, some small fraction of recipes have IF close or equal

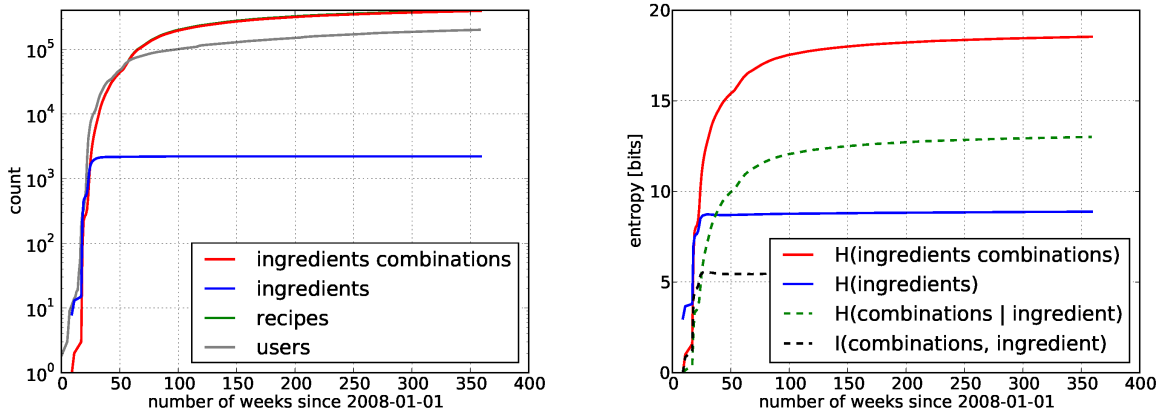

Figure 1: Temporal patterns in terms of number of users, recipes produced, ingredients, and combinations of ingredients used. Although the number of known ingredients saturates very fast (at around 2000 ingredients), users are able to contribute with new combinations for almost every new recipe created in time.

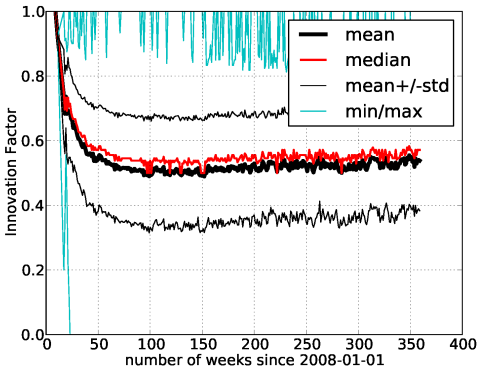

Figure 2: Food innovation over time: after an initial phase of natural innovation decrease the innovation factor stabilizes and shows a steady but slow growth.

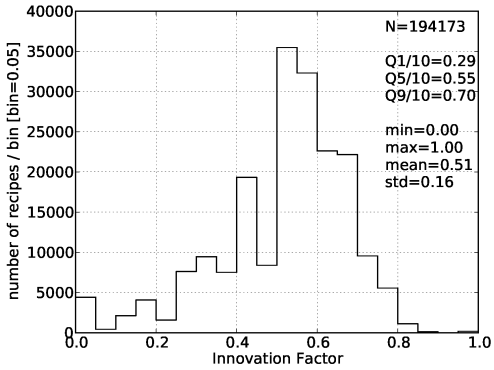

Figure 3: Distribution of innovation of recipes published after 2010-01-01. A shift towards higher values and a long tail are observed.

to 0. One hypothesis is that these recipes are innovative in a way not captured by our measure. An alternative explanation is that there are situations when it is beneficial to reintroduce recipes, e.g., if recipe was forgotten by the community. The question remains open for future research.

Seasonal Patterns. Figure 4 presents seasonal and weekly patterns in the innovation factor of recipes published by the community. Although the differences in general are small there is a clear observable pattern in the course of a year. Two bursts of innovation are visible. The first occurs just after the new year, in January and February. The second one after the summer time in September. Both can be explained by the natural assumption that people starting in the new year or returning back to work after holidays are more creative or eager to start something new. However, temporal patterns on weekly bases are less pronounced, although showing a peak on Thursdays and growth around Mondays.

Categorical Patterns. Finally, Figure 5 highlights the results obtained from our analysis in the context of the top-20 most popular categories assigned to the recipes in our dataset. As shown, online food innovation varies considerable among the investigated categories, showing that, e.g., meat-related dishes being innovated over time significantly more than other categories. To confirm that differences between category innovation arise from their internal characteristics and not by the number of recipes produced (see thick gray lines in Figure 5) in that category, we measured Spearman's rank correlation between those two variables, showing no statistically significant correlation ($\rho = -0.083$, $p = 0.487$). This finding confirms our hypothesis that categorical innovation is not driven by the amount of recipes produced in that category, but rather by other factors.

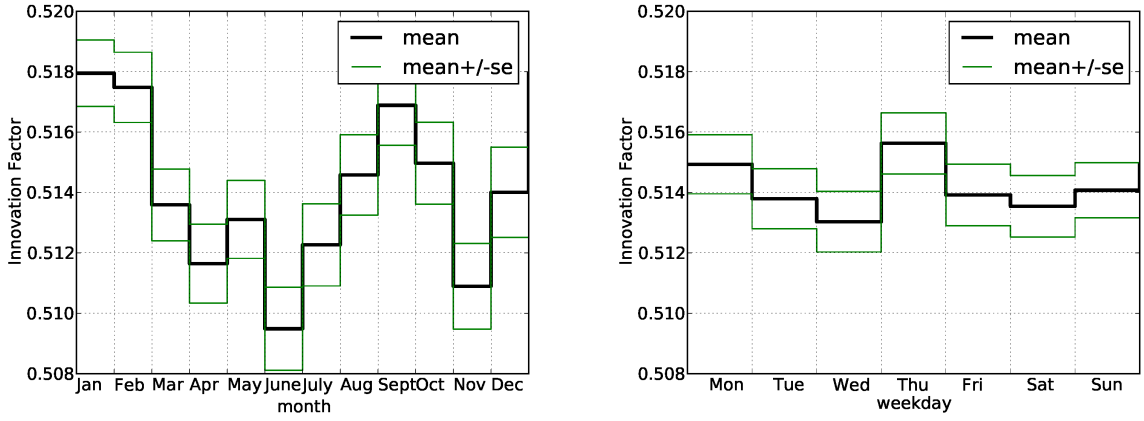

Figure 4: Seasonal and weekly trends in food innovation (for recipes published after 2010-01-01). Seasonal bursts in January and September as well as weekly bursts are observed on Thursdays.

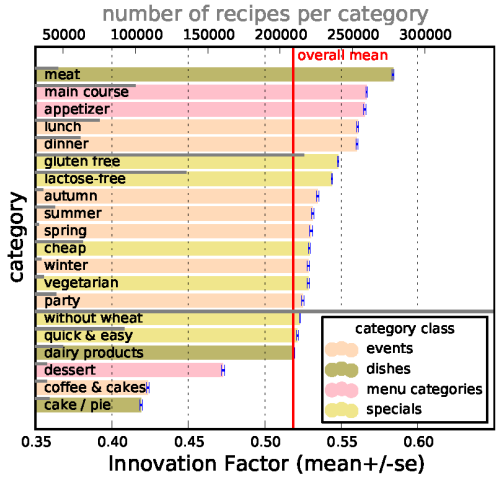

Figure 5: Average innovation and production rate (thick gray line) among top-20 most used food recipe categories (for recipes published after 2010-01-01). Significant discrepancies between different categories, e.g., cakes and meat-based meals, are shown.

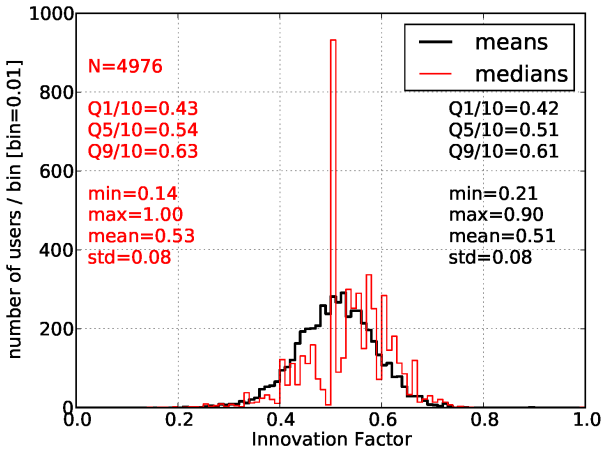

Figure 6: Food innovation distributions among users (min. number of recipes per user set to 10). Although means follow a normal distribution, medians reveal potential irregularities.

5.2 User Patterns

After studying the temporal dynamics in terms of online food innovation on a more global level, we were also interested in studying this effect on a more local level by investigating the users' temporal profiles. To obtain reliable results we filtered out those users having less than 10 recipes published in total, which resulted in almost 5000 users left in our dataset.

General Patterns. Figure 6 highlights the results of our first analysis, where we tried to understand how innovative in general the users are in our dataset. The plot shows distributions of mean and median values of Innovation Factor on the level of the users. In general we find that the means fit nicely Gaussian distribution with a peak at 0.51 which is in line with the results obtained from our previous analysis on recipe level. Interestingly the median distribution draws a different picture, showing that two separate clusters of users exist (smaller and greater than 0.5 in Innovation Factor), i.e., there is a group of users which is more innovative than the other.

To further study this phenomena we a fitted a linear regression model to each of the user's profiles. The results of this experiment are presented on Figure 7. Each user is represented with a dot in the plot. On the horizontal axis the slope and on the vertical axis the intercept of linear function are represented. Time was measured in weeks. What becomes apparent from the plot is that most of the users' innovation factors do not change over time, which means they typically stay still on a certain innovation level. However, there are also clear outliers in the plot showing for some of the users a clear upward or downward trend in time, show-

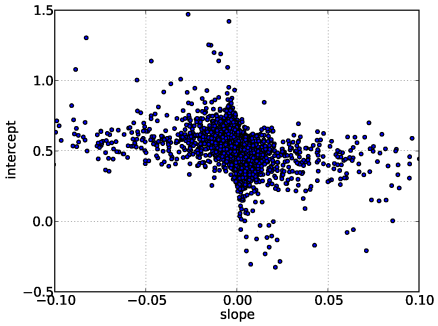

Figure 7: Results of fitting linear regression to users' food innovation temporal profiles (min. number of recipes per user set to 10). We observe that most of the user profiles do not change in time, although there are outliers illustrating that in general change in food innovation over time is possible.

feature	Info. Gain
city	0.0508
#ratings (total)	0.0209
gender	0.0184
#comments (total)	0.0161
#friends	0.0161
guestbook size	0.0133
#recipes	0.0123
#ratings (average)	0.0115
age	0.0008

Table 1: Information Gain values of features in comparison to Innovation Factor. As shown, geographical localization is confirmed as the strongest feature among others such as age or gender.

ing evidence that innovation in time in a positive or negative sense is in general possible.

Factorial Patterns. To further understand the factors that drive the users innovation we measured *Information Gain* between mean values of Innovation Factor and features characterizing users. Although at this moment we do not predict values of *IF* yet, the analysis of relations between features is the first step towards this task. Decreasing values of *Information Gain* and corresponding features are shown in Table 1. Among all features presented in the table we found that geographical localization is the most valuable one compared to the others such as number of ratings, gender, number of friends, etc. Although the results of real prediction experiments are missing and there are for sure further interesting features to study that might improve the model, it is interesting to see that among the ones we proposed, locality was the most important one.

6. CONCLUSIONS & FUTURE WORK

In this paper we explored the extent to which innovation takes place in the context of online food recipe production. Although several recent studies have applied computational methods to study online food recipe consumption and production patterns, to the best of our knowledge we are the first showing how online food communities innovate over time in terms of the recipes they produce. Based on a number of experiments we find interesting patterns such as a steady growth of the innovation factor over time. Another interesting finding is that there are significant seasonal (in the course over a year) and less significant trends over the course of a week how innovation takes place. Also worth highlighting are differences on category level, which means that some online food recipe classes show higher innovation factor than others. Finally, we find on the user level that some users are significantly more innovative than others and if we want to tell how innovation can be predicted, among features such as social relations, productivity, or age and gender, the users locality (i.e., where she lives) is the best feature to predict the innovation factor.

Future Work. Although this work already contains a set of interesting findings in the context of online food recipe innovation, we believe there are several other dimensions worth exploring. First, there is the question of what are the factors that drive online food recipe innovation. Although our work already explored several interesting features in this context, other features such as the users' diversity, creativity or network effect would be interesting to investigate. Also, the locality feature should be further explored and features such as size of the city would be interesting to investigate. Finally, we want to investigate the extent to which recipes change over time and especially why some of them seem to stay more interesting over time than others.

Acknowledgments: This work was carried out during the tenure of an ERCIM "Alain Bensoussan" fellowship program by the second author.

7. REFERENCES

[1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through twitter. In *Proc. of CHI'15*, 2015.

[2] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási. Flavor network and the principles of food pairing. *Scientific reports*, 1, 2011.

[3] E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *Proc. of HT '08*, 2008.

[4] T. Kusmierczyk, C. Trattner, and K. Nørvåg. Temporality in online food recipe consumption and production. In *Proc. of WWW'15*, 2015.

[5] A. Said and A. Bellogín. You are what you eat! tracking health through recipe interactions. In *Proc. of RSWeb'14*, 2014.

[6] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic. Recipe recommendation using ingredient networks. In *Proc. of WebSci'12*, 2012.

[7] D. Tilman and M. Clark. Global diets link environmental sustainability and human health. *Nature*, 515(7528):518–522, 2014.

[8] M. Trevisiol, L. Chiarandini, and R. Baeza-Yates. Buon appetito-recommending personalized menus. In *Proc. of HT'14*, 2014.

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

[10] 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*, 2013.