Nudging Grocery Shoppers to Make Healthier Choices

Elizabeth Wayman
PARC, a Xerox Company
800 Phillips Road
Webster, NY 14580, USA
ElizabethDWayman@gmail.com

Sriganesh Madhvanath
PARC, a Xerox Company
800 Phillips Road
Webster, NY 14580, USA
smadhvan@parc.com

ABSTRACT
Despite the rampant increase in obesity rates and concomitant increases in rates of mortality from heart disease, cancer and diabetes, getting the general public to adopt a healthy diet has proven to be challenging for a variety of reasons. In this paper, we describe Foodle, a research project aimed at providing automated, personalized and goal-driven dietary guidance to users based on their grocery receipt data, by leveraging the availability of digital receipts for grocery store purchases. We discuss challenges faced, the current state of the project, and directions for future work.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords
Nutrition; Diet; Recommender Systems

1. INTRODUCTION
Obesity is a health problem that has assumed epidemic proportions not only in the US but also in large parts of the world, and has been related to increased risk of serious health problems, including diabetes, heart disease, stroke, arthritis and some cancers. Obesity is a consequence of a number of factors [7] but clearly an important one is the poor choice of foods. Despite several public health attempts over the years to specify guidelines for an "ideal diet" and communicate it in simple terms (e.g. the "food pyramid" [1], Dietary Reference Intake (DRIs) [4]), getting the general public to adopt a healthy diet has proven difficult.

In this paper we describe Foodle, a research project aimed at bridging what we see as a vital gap in the present landscape of available tools and aids to promote healthier diets. Foodle is based on the analysis of users' grocery purchase data that is now increasingly available for example, through "shoppers club" accounts. The distinguishing objectives of Foodle are (i) *automation*, to address the tedium and inaccuracy of self-reporting; (ii) *personalization*, to address issues of information overload and relevance; and (iii) *goal-directedness*, to motivate users to achieve dietary goals.

2. RELATED WORK
There are a vast number of nutritional scoring systems for foods [2] accessible either through websites or prominently displayed on food packages in stores, that attempt to relieve the burden of understanding package food labels. For example, Heart Check, developed by the American Heart Association, guides consumers in choosing heart healthy products. The glycemic index, a relative measure of the carbohydrate content of a food [11], is used by diabetics. The Healthy Eating Index (HEI) [12] assesses diet quality by measuring consumption of the various food groups. Aggregate Nutrient Density Index (ANDI) from Whole Foods Markets, ShopWell.com, NutritionFacts.org, Kraft's Sensible Solutions and NuVal are all examples of privately promoted scoring systems developed in conjunction with nutritionists and physicians. While well intentioned, these scoring systems are inconsistent and occasionally of questionable accuracy [2]. While some nutrition scoring sites attempt personalization by requesting data about users' current diets and dietary restrictions, they are tedious to use and do not convey to users the changes they must make to their current diets in practical terms. A personalized diet plan requires the services of a dietician, who may not be accessible or affordable to most individuals and families.

Food recommendations are a relatively nascent area of research. Phanich et al [8] describe a food recommendation system for diabetics that uses clustering to group foods categorized by nutritionists to recommend diabetes friendly alternatives to a food item specified by a user. Other systems attempt to match users to foods they may like. A content-based recommendation system that recommends foods by computing a cosine similarity score between foods and a user profile is described in [9]. The user rates randomly chosen ingredients of foods served at various types of restaurants to build a profile. M. A. El-Dosuky et al [5] propose a semantic food recommendation framework based on matching foods to user supplied profiles from an ontology of foods classified into categories. The system incorporates nutrition heuristics to bias recommendations toward healthy eating patterns.

Work on linking grocery receipt data with nutrition databases is very recent [3] and has been driven primarily by the desire to understand nutritional intake from a public health perspective. Tracking supermarket grocery purchases has been shown to be good proxy for understanding dietary habits and much better than relying on self-reporting [10].

3. THE FOODLE PROJECT
The objective of Foodle was to explore the feasibility of providing personalized food recommendations from a direct understanding of user nutrition. It differs from previous research and existing tools and aids for food recommendations in its emphasis on *automation*, *personalization* and *goal-directedness*. Automation refers to relying on an analysis of the user's grocery receipts instead of self-reporting to understand the user's current diet. Digital receipts are increasingly common, especially in association with customer loyalty programs, and provide a better estimate of a family's diet than any other available method. Personalization refers to food recommendations based on the user's actual purchases, which may be further customized using a user profile containing information such as height, weight, age,

RecSys '15, September 16 - 20, 2015, Vienna, Austria.

ACM 978-1-4503-3692-5/15/09 $15.00
DOI: http://dx.doi.org/10.1145/2792838.2799669

activity level, dietary restrictions and food preferences. Finally, unlike other efforts, Foodle's recommendations are not meant to be solely foods that have high nutritional scores and are generally "good for you", but suggestions that will gradually move the user towards a nutritional goal, such as the USDA Dietary Reference Intake values [4].

As a first step in this exploration, and in order to study the various technical and research challenges involved, we have implemented a prototype, comprised of a user interface and a set of web services. The user interface enables users to visualize their current state of nutrition, set dietary goals, receive food recommendations, and track their progress towards their goals as they implement the recommendations and make other dietary changes. These are supported by a set of web services that ingest grocery receipts, compute the user's state of nutrition from receipt data, support interactive visualization of the current state and historical trends, evaluate the current nutritional state against the user's goals, and compute specific food recommendations. Key aspects of the prototype are detailed below.

3.1 User Interface

Fig. 1 shows the web user interface of the initial Foodle prototype. Once logged in, the user is provided access to several charts that summarize her state of nutrition. The most important of these is a "scorecard" that shows, in the form of bar charts, how the nutrient content of the user's food purchases compares with the recommended amounts. Nutrient content is computed for a temporal window, e.g. the most recent 60 days. The scorecard is divided into three charts corresponding to meets target, excess, and insufficient nutrients. The values are presented as percentages of the recommended daily amounts.

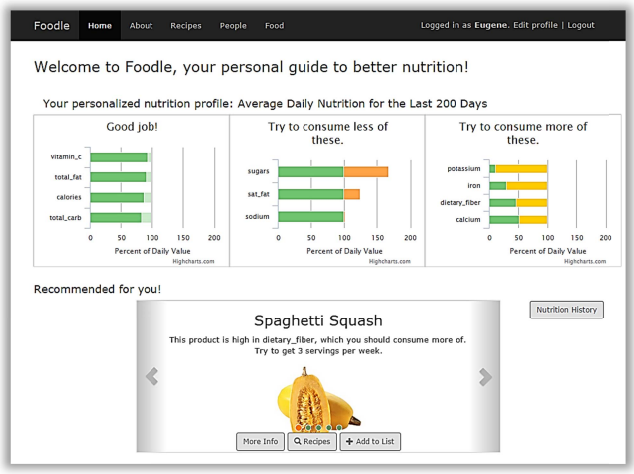

Figure 1. Foodle user interface

The excess or shortfall in nutrient content of food purchases is shown using different colors. Hovering over a bar provides the actual data for that nutrient, as well as the 2-3 food items (and their portions) that substantially contributed that nutrient. The central idea of the visualization is to allow the user to focus on what she needs to work on, as opposed to overwhelming her with information.

Also provided are recommendations of foods that will help the user fill the gaps in the nutritional content of her purchases, as well as a suggestion of how many servings per week are needed (Fig. 1). These recommendations are shown on a carousel that the user can click to advance. For each recommendation, the user is able to see a more complete nutritional description by clicking on the 'More Info' button just below the item. This information includes the percent Daily Value per serving of each of the nutrients contained in the item. The 'Recipes' button takes the user to a recipe site showing results for a search on that item. The 'Add to list' button conveniently adds the item to the user's digital grocery list, accessible through the user profile.

The user interface also provides a Nutrition History chart showing a detailed historical view of the scorecard by plotting nutrient content of food purchases as percentages of their recommended quantities against the time axis (Fig. 2).

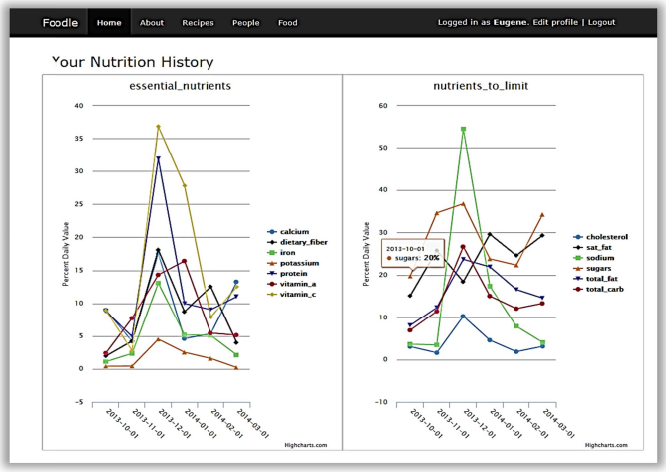

Figure 2. Nutrition History

The chart includes all nutrients currently shown on food labels and groups them into essential nutrients or nutrient to limit. Nutrient values are estimated using overlapping sliding windows of 60 days. The main purpose of this chart is to show the user how she is progressing towards her final goal, and to help identify trends in nutrient content. Hovering over a nutrient in the chart shows detailed information for that nutrient.

3.2 Databases

3.2.1 Receipt Database

The receipt data used for our prototype corresponds to grocery purchases at a local grocery chain that supports a shopper loyalty card. The chain's website allows its customers to view their own grocery receipts and download them in electronic form. The data for the prototype was donated by 15 volunteers and de-identified. The data contains purchase date, product name, department, UPC, quantity, and price information for each product purchased over a period ranging from 6 months to 1.5 years. Products from non-food departments, such as pet food and paper products were excluded from the analysis. The total number of food product purchases across all users was 16123, and the total number of distinct products was 5120.

3.2.2 Nutrition Databases

One of the significant challenges faced in this research was finding nutrition information for the products in the receipt database. The prototype eventually used nutrition data from three sources: (i) a commercial database with an open access API, (ii) nutrition datasheets available from the grocery chain, and (iii) a USDA public database. The preferred source for nutrition data was the commercial database, since it was indexed by UPC and allowed us to unambiguously identify a product based on UPC.

However, while the database claimed to provide nutrition information for as many as 700,000 products, the coverage of food products from our receipt database was poor. One of the contributing factors was the significant fraction of house brand products purchased by our users. In addition, some products were missing nutrition information.

The next source of nutrition data was nutrition datasheets provided by the grocery chain on its website. While these datasheets provided nutrition data for almost 4000 products carried by the chain, not all products carried by the chain were covered, and only about 9% of the products that were covered were identified by UPC. We therefore also explored the use of a USDA public database, the Food and Nutrition Database for Dietary Studies (FNDDS) [6]. This database contains nutrient information for 64 nutrients for over 8,000 food descriptions. However, this database also has no UPC information and the food descriptions are in a form (*<base product>*, *<comma separated modifiers>*) that made matching product names difficult and ambiguous. For example, in the FNDDS, there are 31 distinct entries for "carrots," including "carrots, raw," "carrots, cooked, from fresh," and "carrots, canned, low sodium." However, the grocery store receipts show product descriptions such as "[store name] cleaned and cut triple washed peeled baby carrots," "[store name] carrots, organic," and "[store name] carrots, sliced, no salt added." When a product from the receipt database was not matched by UPC in either the commercial or the grocery chain database, we used string-matching algorithms and heuristics to match product names between receipts and all three databases, and this proved especially valuable for house brand products.

In addition to these food nutrition databases, we used a database containing the recommended daily values of consumption, or dietary reference intake values (DRI) of certain nutrients, available from the USDA [4]. This is a set of tables that captures average recommended consumption amounts of over 20 nutrients, determined by the Food and Nutrition Board, Institute of Medicine, National Academies. These may be customized based on the age, gender and weight of the individual.

3.2.3 Recommendable Foods Database

This database identifies the set of food items from which to select recommendations to users. There are several choices for this. For the prototype, we restricted the recommendations to the food items listed in the nutrition datasheets provided by the grocery chain, since (i) we wish to recommend items that are easily accessible to users, and (ii) the datasheets provided good coverage of produce and other healthy foods. The commercial database, on the other hand, focused on packaged foods, many of which may not have been accessible to our users.

3.3 Nudging Recommendations

3.3.1 Nutritional Gap

We chose to use the nutritional gap between the current state of nutrition and a target state (corresponding to a dietary goal) to guide food recommendations as the initial research approach. The prototype used the DRI values as the dietary goal. The current nutritional intake was estimated over a sliding temporal window such as the most recent 30, 60 or 90 days, by aggregating the nutritional contents of different food products purchased during this window (obtained from the nutrition databases), and scaled by the quantity purchased. Scaling is complicated by the fact that the purchased quantity reported on the receipt varies from mass units (e.g. for produce items) to number (for packaged goods), whereas nutrition databases typically report nutrition per *serving*, and the definition of a serving typically varies both across food products within a database, as well as across databases.

The current nutrition is represented as a vector in *n*-space, where *n* is the number of nutrients. For the prototype, we used a core set of 12 nutrients commonly found on food package labels (and therefore found in most nutrition databases), including, for example, calcium, sodium, and protein. The target nutrition is similarly represented as a vector in this space. This vector uses the DRI values, scaled for the number of days in the period under consideration. The nutritional gap is represented as a vector describing the difference between the amounts of nutrients in the food the user is purchasing and the recommended consumption for the nutrients. The values in the vector represent the percentage gap (or excess) of each nutrient.

In order to speed up computation, a nutrition profile vector, N_p, describing the nutrient content of each recommendable food product was pre-computed. This vector is also in *n*-space, as the nutritional gap vector above.

3.3.2 Computing Recommendations

Food recommendations that nudge the user towards a nutritional goal (in this case, the DRI) are generated by identifying food items (along with the number of servings) that best fill the gap in nutrition described by the nutrition gap vector. To do this, we compute the cosine similarity between N_g, the user's nutrition gap vector and N_p, the nutrition profile vector for each food product in the recommendable foods database.

The 12 nutrients used fall into two categories - *nutrients to limit* (e.g. fat and sodium), and *nutrients to meet* (e.g. calcium, iron, and vitamin C). Elements of the "gap" vector may actually be negative, especially due to an excess of the former category of nutrients. This first exploratory implementation uses a weighted form of the nutrition gap vector N_g, to control the relative importance of the nutrients, as shown below. In particular, nutrients to limit are assigned a weight of 0, which effectively excludes them from the computation.

$$N_{g\text{-weighted}}[i] = N_g[i] * \begin{cases} 0, i \in \{nutrients\ to\ limit\} \\ 2, i = nutrient\ with\ largest\ gap \\ 1, i \in \{all\ other\ nutrients\} \end{cases}$$

The cosine similarity is thus given by

$$\text{Similarity } (food_i) = \frac{Ng-weighted \cdot Np(foodi)}{\|Ng-weighted\|\|Np(foodi)\|}$$

The five highest scoring foods are presented as recommendations. We also present to the user the number of servings for a given time interval that would be needed to fill the gap *for the nutrient with the largest gap*. The number of servings needed is computed as the nutrition gap value for this nutrient divided by the corresponding value in the nutrition profile vector for that food, and displayed as part of the recommendation.

4. RESULTS AND DISCUSSION

The initial Foodle prototype allowed us to demonstrate our research objective of providing automated, personalized and goal driven food recommendations to users. The recommendations are driven by the gap in nutrition between the present state (defined by a sliding temporal window) and the target state corresponding to DRIs. It is expected that these recommendations will be reflected in a user's grocery purchases when adopted, and the nutrition gap will change, as will the recommendations, constantly narrowing the gap. Thus a feedback loop is built into the system. We observed that for users with sufficient grocery data, the recommendations delivered by Foodle are sensible and address

significant nutrient shortages in their diets. For user "Eugene" who showed a 50% deficiency in dietary fiber, the top three recommendations were spaghetti squash (3 servings/week), green beans (3 servings/week) and vegetable bean medley (1 serving/week), all rich sources of dietary fiber.

In this process, we also identified several conceptual and practical challenges:

Under-estimating nutritional intakes: The percentage of food purchases that we were able to match to nutrition information varied between 30 and 60%, depending on the particular user's purchase habits. This resulted primarily from the difficulty in matching product names of purchased foods with their nutrition data, and was particularly significant in the case of house brands, which represented a significant fraction of purchases. Semantic analysis of product descriptions to identify brand names, adjectives (e.g. "fresh", "organic") and ingredients would help increase the accuracy of matching. Other options include screen scraping and crowdsourcing to capture nutrient information from product pages found online. More fundamentally however, nutritional intakes are under-estimated because users may shop at farmer's markets, eat at restaurants or forget to use their loyalty cards. There may be ways to mitigate these effects, such as providing the user a capability to enter that information, either manually, or automatically e.g. by processing restaurant receipts and matching them with nutrition information provided by restaurant chains. Despite these gaps, grocery purchase data has been shown to be a reliable indicator of nutrition, and useful recommendations for nudging may be obtained by making some assumptions about the fraction of consumption going unrecorded (perhaps input by the user) and scaling the available nutrient data.

Individual vs. family purchases: We observed significant excess of all nutrients for some users, and these turned out to be users purchasing groceries for their families. Capturing information about members of the family would allow more accurate estimation of the *aggregate* target nutrition for family units. The recommendations provided would be "personalized" at the level of these family units.

Recommendation algorithms: The algorithm used in the prototype for recommending foods works only for gaps caused by insufficient nutrients. To compute recommendations for excess nutrients, for example, salt or sugar, currently purchased products that contribute high amounts of these nutrients could be identified, and recommended for reduced purchase. However, it is not guaranteed that this process would converge since a food product contains several nutrients. We have developed an alternative formulation of the problem as an unbounded Knapsack problem. Intuitively, the objective is to fill the knapsack with different servings of available food items such that the total "nutritional value" of the knapsack is maximized, subject to the constraint that the net calories from the selected items does not exceed the recommended caloric intake for the individual user, given their profile. But in doing so, our solution attempts to make the least possible change to the current diet.

5. SUMMARY & FUTURE DIRECTIONS

In this paper, we described our research in using grocery receipt purchase data to generate recommendations of foods to nudge the user to a healthier diet. The initial prototype has several advantages over available tools and aids to promote healthier diets, and shows considerable promise for generating highly personalized recommendations that directly address nutritional deficiencies in users' diets. The persuasiveness and adoption of these recommendations needs to be validated using longitudinal user studies.

Beyond addressing the shortcomings of the initial prototype as described, there are several interesting directions for future research: (i) User profiles: The user can specify dietary restrictions or preferences; there may also be ways to learn some of these through data analysis (ii) With a larger population of users, other recommender system architectures can be explored, such as user-user collaborative filtering. This may have the benefit of recommending food items the user is more likely to adopt. (iii) New ways of delivering recommendations can be explored, e.g. in the context of the grocery shopping experience through mobile app, shopping list integration, or even as part of a "smart" grocery cart. (iv) Social navigation may be useful to help users select and adopt food recommendations (v) The scope of recommendations could be extended to school boards, nursing homes and SNAP (food stamp) programs to improve public health outcomes.

6. REFERENCES

[1] A Brief History of USDA Food Guides: 2011.

[2] Armstrong, K. 2010. *Stumped at the Supermarket*. Technical Report. Public Health Law Center, William Mitchell College of Law, St. Paul, Minnesota.

[3] Carol Byrd-Bredbenner, C.A.B. 2010. Assessing the home food environment nutrient supply using mobile barcode (Universal Product Code) scanning technology. *Nutrition & Food Science*. 40, 3 (2010), 305–313.

[4] DRI Tables and Application Reports: *http://fnic.nal.usda.gov/dietary-guidance/dietary-reference-intakes/dri-tables-and-application-reports*.

[5] El-Dosuky, M.A., Rashad, M.Z., Hamza, T.T. and EL-Bassiouny, A.H. 2012. Food Recommendation using Ontology and Heuristics. *Advanced Machine Learning Technologies and Applications* (Cairo, 2012), 423–429.

[6] Food Surveys : FNDDS 5.0: *http://www.ars.usda.gov/Services/docs.htm?docid=22370*.

[7] Obesity Epidemic "Astronomical": *http://www.webmd.com/diet/obesity-epidemic-astronomical*.

[8] Phanich, M., Pholkul, P. and Phimoltares, S. 2010. Food Recommendation System Using Clustering Analysis for Diabetic Patients. *International Conference on Information Science and Applications* (Seoul, Korea, 2010), 1–8.

[9] Tatli, I. 2009. Food Recommendation System Project Report 1- Introduction and Problem Definition (2009).

[10] Tin, S.T., Mhurchu, C.N. and Bullen, C. 2007. Supermarket Sales Data : Feasibility and Applicability in Population Food and Nutrition Monitoring. 65, 1 (2007), 20–30.

[11] Atkinson, Fiona S., R., Kaye Foser-Powell, Kaye, R. and Brand- Miller, Jennie C., P. 2008. International Tables of Glycemic Index and Glycemic Load Values : 2008. Diabetes Care. 31, 12 (2008), 2281–2283.

[12] Guenther, P.M., Kirkpatrick, S.I., Reedy, J., Krebs-smith, S.M., Buckman, D.W., Dodd, K.W., Casavale, K.O. and Carroll, R.J. 2014. The Healthy Eating Index-2010 Is a Valid and Reliable Measure of Diet Quality According to the 2010 Dietary Guidelines for Americans 1 – 3. (2014), 399–407.