Castiglia Giovanni, Calò Federica

Final Writeup AIML/BDA Project work

Abstract

We designed and developed a multi-modal conversational food recommender system for driving users toward healthier choices. The majority of existing conversational recommender systems (CRSs) rely solely on natural language or basic click-based interactions. Here we examined the effect of two interaction modalities: pure textual and multi-modal (text plus images). We conducted a within-subject user study (N=30) to evaluate the two interaction modalities in terms of how effectively they supported users in selecting healthier foods and in terms of satisfaction, effectiveness and efficiency of their choices. We also evaluated other system-measured variables using T-test analysis.

Introduction

Multi-modal Conversational Information Seeking systems are gaining popularity and consist of introducing multiple signals during conversational tasks. This could include different combinations of input and outputs, such as **text and images**. This powerful combination can influence user behavior and lead her towards different directions. Imagine the possibility of driving users using really good quality images.

In the present work, we designed and developed a multi-modal content-based conversational food recommender system, aimed at recommending healthy dishes to users. We used the *python-telegram-bot* library for developing our system as a Telegram interactive chatbot (**@mlfoofbot**). We extracted 2000 items from Allrecipes.com dataset (under EULA agreement) and generated a simple dataset used by our system.

The main goal of this work is understanding the impact of two different interaction modalities on user healthy choices, satisfaction, effectiveness and efficiency, and on system-measured metrics. The two interaction modalities are the Textual (T) one which consists in a pure text-based interaction of the user with the system, and the Multi-modal (MM) one, which combines text and images to interact with the user. In the first case, the system shows, recommends and explains items using pure text. In the second case, images help enrich the presentation and recommendation. The explanation remains textual.

Related Work

Conversational Recommender Systems (CRSs) combine the power of recommendation algorithms with conversational strategies. Using multi-turn conversations, they are able to collect users preferences and provide personalized recommendations. CRSs are used in a various domains, such as medical diagnosis (Cordero et al.), e-commerce (Griol and Milina), and entertainment (Narducci et al.) (Iovine et al.). Only a few studies have investigated the use of CRS in healthy food recommendation (Trattner and Elsweiler). Food selection process is contextual and influenced by a variety of factors, such as the user mood and dietary constraints, but also by factors unrelated to the food content, such as the perception of the food's visual characteristics by individual users (Starke et al.). Moreover, people generally prefer food that has a more visually appealing presentation, such as food that is presented in an attractive way (Peng and Jemott).

While multi-modal conversational information seeking (MMCIS) is gaining attention by the research in the RecSys/IR/HCI communities (Deldjoo et al.) (Zamani et al.) (Sousa et al.), only a few practical studies have been published that focus on topics other than food and health, such as conservational systems on tourism (Liao et al.) and fashion (Moon et al.) (Yuan and Lam). In the field of food recommendation, (Elsweiler et al.) discussed recent advances in the field of food recommender systems in general.

Dataset and features

To reach food data, we started from the Allrecipes.com dataset, which was obtained in July 2015 by implementing a standard web crawler. It contains 60983 recipes published between the years 2000 and 2015 on the popular Allrecipes.com website. In particular, 58263 out of these recipes contain basic nutritional details to assess the healthiness of recipes.

We accessed the dataset for educational scopes under EULA (End User License Agreement) agreement and we extracted 2000 recipes.

We decided to consider four food categories:

- pasta (500 dishes)
- salad (500 dishes)
- dessert (500 dishes)
- snack (500 dishes)

The acquired information helped us creating a simpler dataset (composed of 4 smaller datasets), containing dishes characterized by the attributes reported here:

Field name	Description
id	URL of dish recipe on Allrecipes.com
category	Dish category in [Pasta, Salad, Dessert, Snack]
image_url	Dish image URL on Allrecipes.com
name	Dish recipe name
energy_Kcal	Calories of dish (Kcal)
energy_Kj	Calories of dish (Kj)
ingredients	List of ingredients of the dish
servings	Servings of the recipe
proteins	Proteins in the dish (g)
carbohydrates	Carbohydrates in the dish (g)
fibers	Fibers in the dish (g)
sugar	Sugar in the dish (g)
fats	Fats in the dish (g)
saturates	Saturates in the dish (g)
sodium	Sodium in the dish (g)
salt	Salt in the dish (g)
tot_grams_weight	Total weight of the dish (g)
serving_size	Weight of the serving (g)
proteins_100_grams	Proteins in 100 g of dish (g)
carbohydrates_100_grams	Carbohydrates in g of the dish (g)
fibers_100_grams	Fibers in 100 g of dish (g)
sugar_100_grams	Sugar in 100 g of dish (g)
fats_100_grams	Fats in 100 g of dish (g)

saturates_100_grams	Saturates in 100 g of dish (g)
sodium_100_grams	Sodium in 100 g of dish (g)
salt_100_grams	Salt in 100 g of dish (g)
nutri_score	Computed nutriscore
fsa_score	Computed fsa_score

In order to filter the menu according to user constraints, we build a dataset containing for each well-known intolerance/disease a list of ingredients to avoid. This will help us remove dishes which have dangerous ingredients for a certain user. We started from the diseases and manually scan dishes from the menu and collected ingredients in the dataset. We also looked at online medical resources to investigate diseases and intolerances.
The simple dataset we created is composed by the attributes reported here:

Field name	Description
name	Name of the disease/intolerance
type	Disease or intolerance flag
ingredients_to_avoid	List of ingredients to avoid

An example row is reported here:
"**Alcohol**","**Intolerance**","sherry,vermouth,vodka,cognac,white wine,wine,red wine,beer,liquor,liqueur,whiskey,rum,peach schnapps,apricot brandy,bourbon,triple sec".

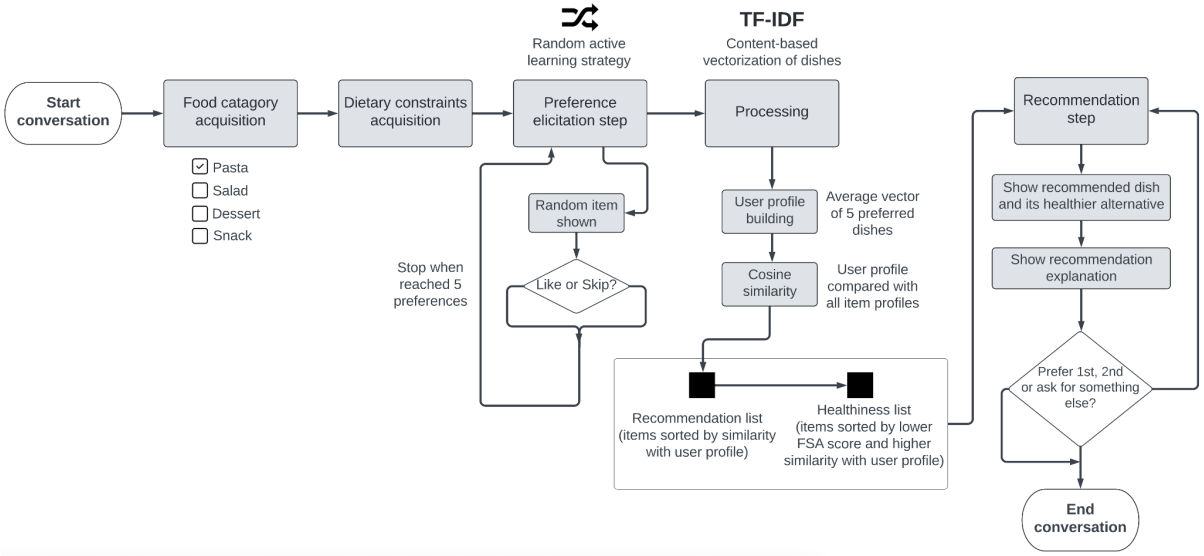
Methods

The main steps of the conversation flow are described below:

- Food category acquisition: choice of food category (Pasta, Salad, Dessert, and Snack)
- User constraints acquisition: input of potential dietary constraints (intolerances/diseases or specific ingredients to avoid)
- Preference elicitation: submission of preferences for five of the randomly proposed dishes (“Like” and “Skip” buttons used during the interaction)
- Processing: item representation by TF-IDF vectorization. User profile building as the mean of the vectors of preferred items. Use of cosine similarity for recommendation list ranking. Extraction of an item from recommendation list

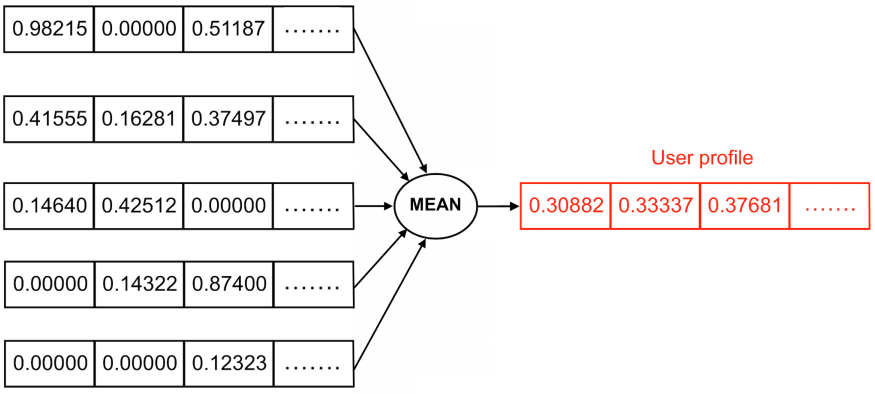
(which fits user preferences) and generation of a healthier alternative (rerank based on 60% of healthiness and 40% of similarity with the current recommended dish).

- Recommendation and explanation: recommendation of the two dish items and textual explanation which describes why the second dish is healthier than the first one. The user can choose one of them or ask for another comparison (iterative process).



In order to represent items and their features, we decided to use the vector representation using TF-IDF (not normalized). Each dish has been represented by a vector of ingredients, represented by their feature score which represents the importance of that ingredient in that dish according to other dishes in the collection (and their ingredients). We used for this task the *TfIdfVectorizer* module of the *scikit-learn* library. We preprocessed the TF-IDF vectors for each category of dish, in order to avoid that computation during the user-system conversation. The system uses the matrix of vectors in order to compute similarity among dishes.

In order to represent user preferences, we build each user profile by computing the mean of the vectors of dishes liked by each user. In this way, the score associated with each ingredient is the result of the average of scores of the vector representation of the dishes liked by the user (5 in optimal cases). A representation of this process is reported below. The resulting vector has the same number of features of dish items of the menu for the category selected by the user. In this way it is easy to compare vectors of dishes with the user profile.



TF-IDF vectors of five preferred dishes

In order to compare the user profile with dish items to produce a recommendation list, we chose the Cosine similarity metric. We performed the cosine similarity score among each pair of dishes of each category. To do so, we used the *cosine_similarity* module of *scikit-learn* library. For each food category, we append the user profile vector to the TF-IDF matrix containing item representations and we perform cosine similarity. We obtain a new matrix of cosine scores representing the similarity evaluation between two corresponding items (by looking at their ingredients scores). We extracted this last row of this matrix (represents the similarity scores of each dish of the menu with the user profile) and removed its last element, with value 1, referring to the similarity score performed between user profile and itself. Then, we matched each dish item with its similarity score with the user profile. In order to generate the recommendation list, we sorted the menu according to decreasing cosine similarity with the user profile. Then, we generated a new healthy list, by sorting the recommendation list according to each dish's FSA score. This metric is commonly used in the culinary recommendation scenario and helps evaluate the healthiness of a dish. We recommend the best item by looking at its rankings in the two list according to this score: *0.4 * rank_rec_list + 0.6 * rank_health_list*.

Experiments and Results

We set up a small user study (N=30) with family-and-friends participants. We let them take part in this within-subject study which consisted in interacting with our two versions of the bot (different interactions modality) in turn: in our case we decided to present firstly the textual modality bot (T) and after that the multi-modal modality bot (MM). The users who completed the task of interaction with the two bot versions had to compile a short questionnaire with 5-point likert scales. We performed a simple analysis of variance (T-test) for investigating the impact of interaction modality (T, MM) on various acquired variables.
We designed a short Google Form questionnaire in which we asked user to specify the preferred modality and evaluate on 5 point scales the following metrics:

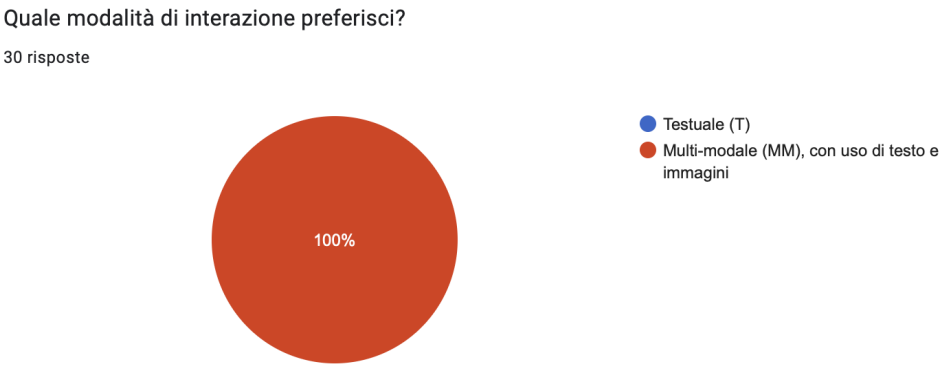
- satisfaction (what interaction modality is **more satisfying**?)

- effectiveness (what interaction modality is **easier to use**?)
- efficiency (what interaction modality is **faster to use**?)

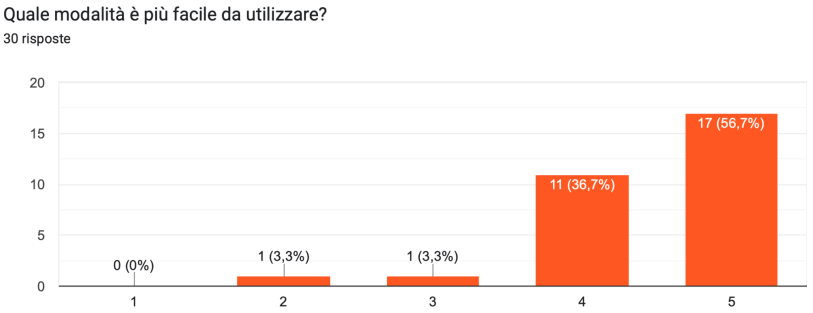
We also asked users to express the reasons for their choice using a multiple checkboxes question.
Our form is available here: https://forms.gle/okQYPcGSLqMNPKky9.

We retrieve the following insights:

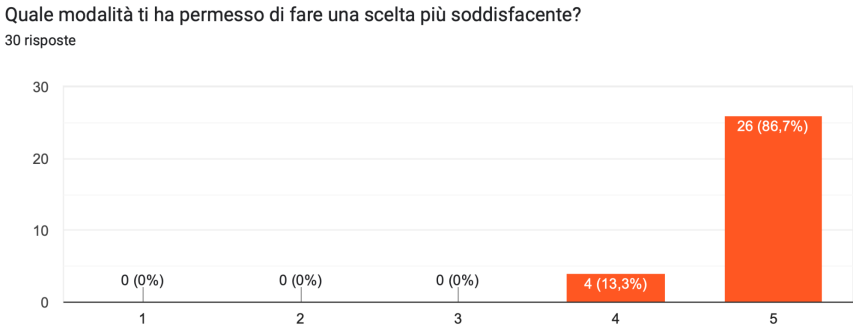
1. All the users (100%) preferred the multi-modal modality.



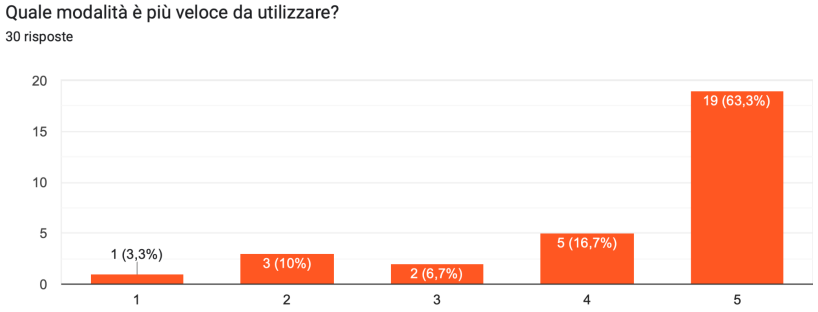
2. The main reasons for their choices is that MM representation allows them to identify clearly the dishes wrt the T representation (96.7%) and that image quality is good (20%).
3. The majority of users strongly thinks MM modality is easier to use (56.7%) but also in a partially decisive way (36.7%).



4. The majority of users are completely satisfied with MM modality (86.7%) and also partially satisfied with MM modality (13.3%).



5. The majority of users think that the MM modality is faster to use (63.3%)



We performed a T-test to understand if two considered populations are statistically different from each other. Our aim is to understand if the change of modality (T, MM) directly impacts certain measured variables. T-test (like ANOVA with more than two groups) looks at the difference in means and the spread of distributions across **two groups**. It tests the so-called **null hypothesis**: it defines if two groups have the same population mean. In our case, if the null hypothesis is rejected, there is a good signal from the test, so the change of modality has a direct impact on the tested variable. T-test produces a test statistic value (T-statistic) which is then converted to a p-value. Generally a p-value lower than 0.05 allows us to reject the null hypothesis.

Example: the null hypothesis tells that the variable X = conversation duration is the same for all two groups (one per modality, T and MM)

We tested the following variables:

- conversation duration
- preference elicitation step duration
- recommendation step duration
- number of interactions (Note: it increases when user continues the recommendation step, by asking for new dish comparisons)
- has user performed a healthier choice (boolean flag)
- FSA score of selected dish
- number of skips during preference elicitation step

We grouped a dataframe (obtained from our measurements dataset) by the field scenario (Textual or Multi-modal) and we split it into two different dataframes. This helped us easily perform means and creating groups of values for each studied variable.

We obtained these results from the mean calculation of some fields:

Variable	Mean of values for T group	Mean of values for MM group
Conversation duration	148.46 s	132.03 s
Preference elicitation step duration	51.06 s	56.86 s
Recommendation step duration	37 s	28.13 s
Number of interactions	6.03	6.3
Number of skips	5.06	6.13

We obtained the following results from the T-test.

Variable	T-statistic	p-value	Null Hyp. Rejected?
Conversation duration	0.777	0.440	No
Preference elicitation step duration	-0.749	0.456	No
Recommendation step duration	1.880	0.065	**No***
Number of interactions	-1.705	0.093	**No***
Has user performed a healthier choice	-0.255	0.799	No
FSA score	-1.223	0.225	No
Number of skips	-0.872	0.386	No

* We obtained a good p-value but not sufficient for declaring rejected the null hypothesis. This could be caused by the relatively small number of participants. The p-value is greater than 0.05 so we cannot reject the null hypothesis of the test. We don't have sufficient evidence to say that the mean duration time between the two modalities is different.

In a certain way, hypothetically we could assume that the recommendation step time is influenced by the modality proposed to the user (T,MM). This could tell us that the user spent less time when the recommendation is accompanied by images. They have the power of driving users faster to the desired item.

Moreover, the majority of users have performed a healthier choice using both interaction modalities. We can also state that the MM modality has slightly better influenced the user to perform healthier choices. Results, which could be not so significant due to the small user set, are displayed below.

Has the user performed a healthier choice?	True	False
T modality	16	14
MM modality	17	13

Conclusion

In conclusion, we developed a multi-modal conversational approach for healthy food recommendation. We were able to test the functionalities of our system and investigate the impact of the two versions (T and MM interaction mods.) on user satisfaction, effectiveness and efficiency and on system measurements (timings, etc.). Unfortunately we could not infer that modalities have a direct impact on the variables due to T-test incomes. This is probably caused by the limited number of the user study participants. We applied statistical analysis and techniques learned from the AIML/BDA course in this project work and we equally distributed the work tasks.

Works cited

Cordero, P., et al. “A conversational recommender system for diagnosis using Fuzzy Rules.” *Expert Systems with Applications*, vol. 154, 2020, p. 113449. 10.1016/j.eswa.2020.113449.

Deldjoo, Yashar, et al. “Towards multi-modal conversational information seeking.” *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval*, 2021, 1577--1587.

Elsweiler, David, et al. “Food recommender systems.” *Recommender Systems Handbook*, 2022, 871--925.

Griol, D., and J. Milina. “From VoiceXML to multimodal mobile Apps: development of practical conversational interfaces.” *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, vol. 5, 2016, p. 43.

Iovine, A., et al. “Conversational Recommender Systems and natural language: : {A} study through the ConveRSE framework.” *Decis. Support Syst.*, vol. 131, 2020, p. 113250. 10.1016/j.dss.2020.113250.

Liao, Lizi, et al. “MMConv: an environment for multimodal conversational search across multiple domains.” *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, 675--684.

Moon, Seungwhan, et al. “Situated and interactive multimodal conversations.” *arXiv preprint arXiv:2006.01460*, 2020.

Narducci, F., et al. “An investigation on the user interaction modes of conversational recommender systems for the music domain.” *User Model. User Adapt. Interact.*, vol. 30, 2020, 251--284. 10.1007/s11257-019-09250-7.

Peng, Yilang, and John Jemott. “Feast for the Eyes: Effects of Food Perceptions and Computer Vision Features on Food Photo Popularity.” *International Journal of Communication (19328036)*, vol. 12, 2018.

Sousa, Ricardo Gamelas, et al. “iFetch: Multimodal Conversational Agents for the Online Fashion Marketplace.” *Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI*, 2021, 25--26.

Starke, Alain, et al. “Nudging healthy choices in food search through visual attractiveness.” *Frontiers in Artificial Intelligence*, vol. 4, 2021.

Trattner, C., and E. Elsweiler. “Food Recommendations.” *Collaborative recommendations: Algorithms, practical challenges and applications*, World Scientific, 2019, 653--685.

Yuan, Yifei, and Wai Lam. “Conversational fashion image retrieval via multiturn natural language feedback.” *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, 839--848.

Zamani, Hamed, et al. “Conversational information seeking.” *arXiv preprint arXiv:2201.08808*, 2022.