Nudging Towards Health in a Conversational Food Recommender System Using Multi-Modal Interactions and Nutrition Labels

Giovanni Castiglia[1], Ayoub El Majjodi[2], Federica Calò[1], Yashar Deldjoo[1], Fedelucio Narducci[1], Alain Starke[2,3] and Christoph Trattner[2]

[1]Polytechnic University of Bari, Bari, Italy
[2]Department of information science and media studies, University of Bergen, Bergen, Norway
[3]Marketing and Consumer Behaviour Group, Wageningen University & Research, Wageningen, The Netherlands

Abstract

Humans engage with other humans and their surroundings through various modalities, most notably speech, sight, and touch. In a conversation, all these inputs provide an overview of how another person is feeling. When translating these modalities to a digital context, most of them are unfortunately lost. The majority of existing conversational recommender systems (CRSs) rely solely on natural language or basic click-based interactions.

This work is one of the first studies to examine the influence of multi-modal interactions in a conversational food recommender system. In particular, we examined the effect of three distinct interaction modalities: pure textual, multi-modal (text plus visuals), and multi-modal supplemented with nutritional labeling. We conducted a user study (N=195) to evaluate the three interaction modalities in terms of how effectively they supported users in selecting healthier foods. Structural equation modelling revealed that users engaged more extensively with the multi-modal system that was annotated with labels, compared to the system with a single modality, and in turn evaluated it as more effective.

Keywords

Personalization, Health, Food recommendation, Digital Nudges, Nutrition labels

1. Introduction and Context

Conversational recommender systems (CRSs) represent a hotly debated area of study in the field of information seeking [1, 2]. They combine the power of recommendation algorithms with conversational strategies. Using multi-turn conversations, CRSs are able to collect users' nuanced and dynamic preferences in more depth, which can enhance recommendation outcomes and user experience. CRSs are utilized in a variety of domains, including medical diagnosis [3], e-commerce [4], and entertainment [5, 6]. Only a few studies have investigated their merit for food recommendation [7], and in particular for encouraging users to make *healthier* food decisions.

Over 60% of all deaths are caused by non-communicable diseases, which are preventable by tackling risk factors, such as attaining a healthy food intake [8]. While our food decisions are driven by our overall preferences, the food selection process is extremely contextual and influenced by a variety of factors, such as the user's mood and dietary constraints. Moreover, many of the decisions are made spontaneously and consumers' judgments are influenced by factors unrelated to the food content, such as their perception of the food's visual characteristics [9]. For instance, the packaging of items with nutritional labels can serve to highlight the nutritious nature of the food (cf. [10]). Moreover, people generally prefer food that has a more visually appealing presentation, such as food that is presented in an attractive way [11]. People are willing to pay extra for food whose ingredients are tastefully/attractively organized, and restaurants strive to generate Instagram-friendly photographs by enhancing the color composition of their plates.

To surface effective and healthy food recommendations it is crucial to understand these underlying decision factors. Regrettably, the large majority of existing conversational recommender systems [12, 13] only consider a single type of interaction, such as natural language or click-based interaction, thereby neglecting a wealth of information in the actual imaging of meals [14]. The goal of the present work at hand is to employ a new conversational model for food recommendation that permits more natural, multi-modal user-system interaction.

4th Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) Workshop @ RecSys 2022, September 18–23 2023, Seattle, WA, USA.

g.castiglia@studenti.poliba.it (G. Castiglia); ayoub.majjodiu@uib.no (A. E. Majjodi); f.calo8@studenti.poliba.it (F. Calò); yashar.deldjoo@poliba.it (Y. Deldjoo); fedelucio.narducci@poliba.it (F. Narducci); alain.starke@uib.no (A. Starke); christoph.trattner@uib.no (C. Trattner)
https://www.christophtrattner.info/ (C. Trattner)
0000-0002-7478-5811 (A. E. Majjodi); 0000-0002-6767-358X (Y. Deldjoo); 0000-0002-9255-3256 (F. Narducci); 0000-0002-9873-8016 (A. Starke); 0000-0002-1193-0508 (C. Trattner)

To attain this goal, this paper introduces a *multi-modal conversational food recommender system* (MMCFRS). It implements different user-system interaction modes, along with nutrition labelling in order to assist the user in making dietary decisions. Our objective is to examine the effects of three distinct interaction modes: pure textual, multi-modal (text plus visuals), and multi-modal supplemented with *nutritional labeling*. While multimodal conversational information seeking (MMCIS) is gaining attention by the research in the RecSys/IR/HCI communities [15, 1, 16], only a few practical studies have been published that focus on topics other than food and health, such as conservational systems on tourism [17] and fashion [18, 19]. In the field of food recommendation, Elsweiler et al. [20] provide a good frame of reference for recent advances in the field of food recommender systems in general. Specifically for conversational systems, Barko-Sherif et al. [21] investigate the possibility for conversational preference elicitation in a food recommender environment, using a Wizard of Oz study design (see also [22]). Using a between-groups approach, they compare spoken and text-input chat interfaces and reported that such interfaces are useful for users to describe their needs and preferences. In other studies, Samagaio et al. [23] present a RASA-based chatbot that can recognize and categorize user intentions in the conversation aimed to elicit food preferences for recommendation purposes. Another study of Samagaio et al. [24] applies more knowledge-based elements based on word embedding to optimize conversational ingredient retrieval. These studies, however, focus less on aspects pertaining to health, health labelling, or elicitation modalities. In a non-conversational recommender context, El Majjodi et al. [25] recently indicated that nutritional labels can reduce user's choice difficulty in non-conversational context. The primary distinction between our work and previous studies is the lack of multiple modalities (typically only text is used), as well as that only a few studies (e.g., [25]) have used nutrition labelling.

To summarize, the goal of this study is to compare the impact of three user-system interaction and explanation modalities (textual, multi-modal, and multi-modal with nutritional labels) on both behavioral aspects (what type of recipe is chosen? How healthy is that recipe?) and evaluation aspects (how does the user evaluate the system or their chosen recipe?). Using a mediation analysis (structural equation modelling), we answer the following research question:

- *RQ:* To what extent do different interaction modalities affect a user's recipe choices and evaluation in a conversational food recommendation scenario?

To address this question, we consider different dimensions of analysis. This includes system interaction length, presentation time, healthiness of recipes chosen and a user's level of choice satisfaction and experienced system effectiveness.

2. System Design

In this section we describe the features of our conversational food recommender system, which supports users in making healthier choices.[1]
We designed a system-driven conversation in which the system requires user feedback (response/input) to continue. The main steps of the conversational flow are shown in Figure 1. Users can interact with the system using both buttons and textual messages[2]. The main steps of the interaction are reported below:

- *Food category acquisition*: The user was presented with a choice of *four* different food categories that were considered in this work: Pasta, Salad, Dessert, and Snack.
- *User constraints acquisition*: The user was then prompted to indicate any potential dietary constraints. Initially, the system used an interface with a single checkbox for each of the most prevalent *intolerances* and *allergies*: Lactose, Meat, Alcohol, Seafood, Reflux, Cholesterol, Diabetes. Afterwards, the system asked the user to disclose a list of ingredients she could not consume.
- *Preference elicitation*: According to the constraints specified by the user, the user was prompted to submit preferences for five of the dishes proposed by the system. Each dish was accompanied with two buttons: "Like" and "Skip". The skip option was provided to encourage users to inspect an addition dish, which was retrieved from the randomly sorted menu. The retrieval was based on a random active learning strategy. This way, users were encouraged to like five dishes they were interested in, after which the user profile was built by the system.
- *Processing*: The system constructed the *user profile* by analyzing the user's five preferences from the previous stage. The cosine similarity was computed between the user profile and each of the available foods in the catalog, to provide a list of dishes from which recommendations would be selected. The algorithm also provided a list of dishes ranked according to their healthiness (based on their FSA score; see Section 3).

[1] Code and recipe data used for implementing the chatbot are available at https://github.com/giocast/MMCFRS
[2] A video demo of the three versions of our system is available at https://tinyurl.com/mtzxr2sw

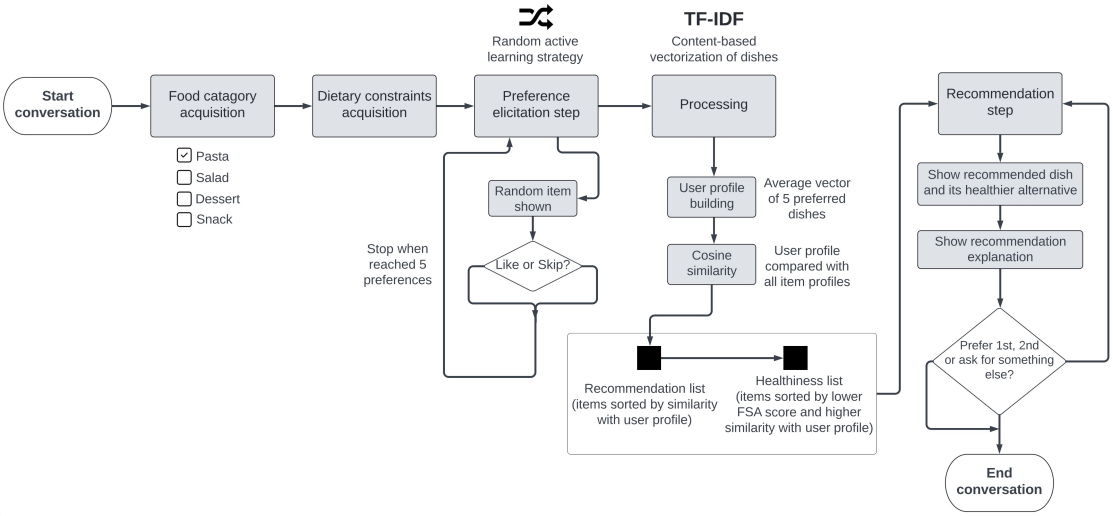

Figure 1: Our conversational recommender system flow.

For each food category we built a matrix containing the TF-IDF representation (dish vs. ingredient) of dishes in the catalog. The higher the TF-IDF score, the greater the ingredient's significance to this dish (as opposed to other dishes).

- *Recommendation and explanation*: The system provided two personalized recommendations, based on the user's preferences. The system constrained the retrieval to ensure that the two options differed in terms of healthiness, so that one option was healthier than the other. Thus, the algorithm provided a description of the suggested dishes. Specifically, it explained why the second dish was healthier than the first and why the advice was made. The user would then be prompted to select one or request a new recommendation. The two recommended dishes were chosen using the following strategy: The first dish would be the most similar to the user profile, while the second dish (the healthier alternative) was selected from a list of most similar dishes ranked on their FSA scores, selecting the healthiest one (i.e. with the lowest FSA score).

Three different interaction modes were implemented by modifying the values associated with the two manipulated variables: interaction *I* and explanation *E*, according to Table 1.

In the *Pure text* version (T + T), the system communicates with the user solely through text, displaying simply the dish titles and offering textual explanations of the food recommendations. In the *Multi-modal* version (MM + T), the system engages the user in a multi-modal manner by displaying the name and image of each dish throughout the dialogue. However, the supplied explanation remains textual. For the first dish, the explanation can be like "I recommend these dish because I know that you have diet constraints due to: meat, zucchini. The first dish I proposed contains ingredients that you might like: carrot, lemon, tuna, olive oil". For the second recommendation, the explanation further provides information about macro nutrients quantities of the two recommended dishes and can be in the form of "The second dish I proposed has less calories (54 Kcal) than the first one (123 Kcal) and has less fats than the first one. The third version *MM-Label* (MM + MM) likewise employs a multi-modal interaction approach, but it also makes use of nutritional explanations in the form of a front-of-package nutrition label with FSA's Multiple Traffic Lights (MTL) [25]. MTL nutrition labels depicted the intake adequacy of a dish in terms of energy and nutritional content, along five dimensions: energy (kcal), fat, saturates, sugars, and salt. This adequacy, per serving and per 100g, was depicted using the colors green, yellow and red, where green indicated a dish to adhere to the nutritional intake guideline, while red indicated that the content was unacceptable. These labels were generated

Table 1
Differences between three implementations of the system.

Interaction Mode	I	E
Pure text (T)	T	T
Multi-modal (MM)	MM	T
Multi-modal with labels (MM-Label)	MM	MM

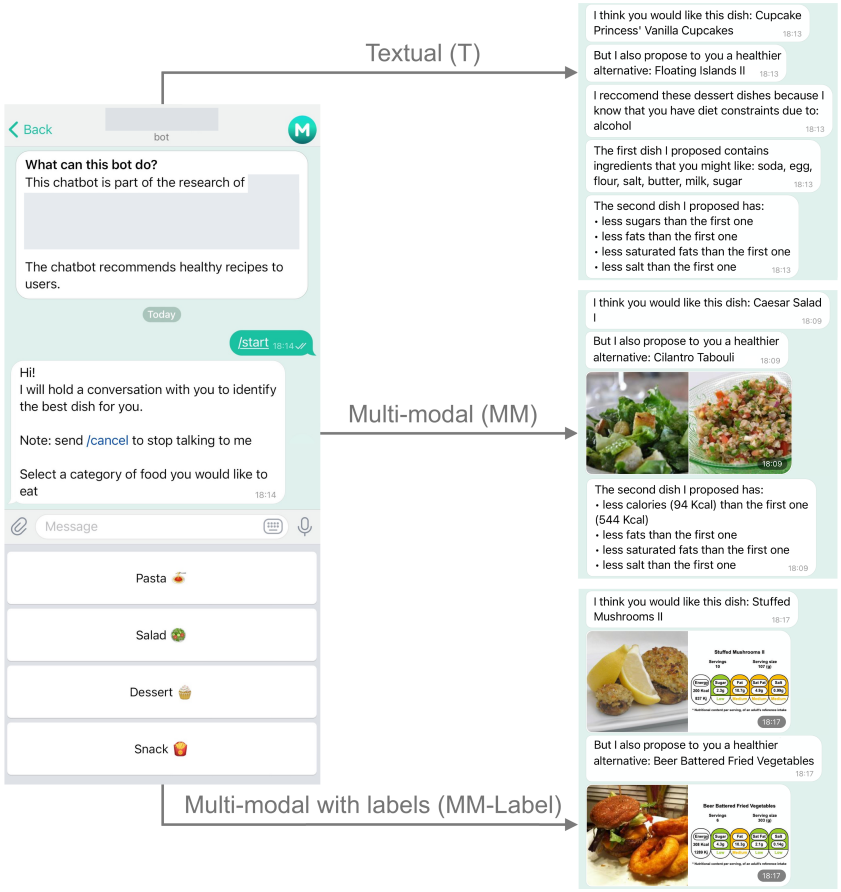

Figure 2: The three implementations of the system. Some details displayed on the interface, such as the chatbot's and authors' names are anonymized and will be added after peer review.

for each dish by following the directives of Food Standard Agency and UK department of health [26].

Figure 2 depicts a snapshot of the chatbot prototype, visualizing the different interaction phases.

In the *Textual (T)* version, the user received recommendations identified by only the names of the dishes (e.g., Cupcake Princess' Vanilla Cupcakes, Floating Island II). The recommendations were followed by textual explanations, based on the ingredients in the dish that the user likes. A comparative analysis of the nutritional facts (e.g., 'less sugars') would also be provided. In the *Multi-modal (MM)* version, the system additionally provided images of the recommended dishes. The explanation was similar to the one presented in the *T* version. Finally, the *Multi-modal with labels (MM-Label)* version provided nutritional labels that were annotated to the depicted images (e.g., Sugar 2.3g, Fat 10.7g, etc.) presented with red, yellow, and/or green colors according to the FSA score. As stated previously, following the presentation of the recommendations, we provide the user with an explanation that helps her comprehend the health benefits of the second alternative above the first, which is the dish that best matches her preferences. This is accomplished either by text (T and MM variants) or a multiple traffic light nutritional label (MM-Label).

The user can accept one of the two dishes proposed or can ask for another recommendation.

3. Experimental Evaluation

To evaluate the extent to which different versions of the chatbot affected users' evaluations and decisions, we recruited 195 participants from Amazon MTurk to use our system. Participants had to have a hit rate of 95% at least and were compensated with 2 dollars. On average, user required around 15 minutes to complete the study.[3] Users

[3]The research conformed to the ethical standards of the Norwegian Centre for Research Data (NSD). The collected data is available in

Table 2
Questionnaire items used in the confirmatory factor analysis. Alpha denotes Cronbach's Alpha, AVE denotes the Average Variance Explained, indicating construct validity if AVE > 0.5. Items in gray and without loading were omitted from analysis. Choice Satisfaction did not form a sensible aspect, because of a lack of construct validity.

Aspect	Item	Loading
Choice Satisfaction	I think, I would enjoy eating the dish I have chosen in the end	
	I would recommend the dish I've chosen in the end to others	
	My chosen dish could become my favorite	
System Effectiveness Alpha = 0.740 AVE = 0.534	It was easy to make my final choice on the dish	0.737
	I interacted a lot with the system before getting the dish of my choice	
	The explanation influenced my final choice of dish	
	I think, that I would use this system frequently	
	I found the system easy to use and understand	0.724
	I felt very confident using the system	0.661
	I would imagine that most people would learn to use this system very quickly	0.722

performed the processes outlined in Section 2, interacting with our chatbot for preference elicitation, evaluating recipe recommendations, selecting one recipe, and evaluating the experience. A user's experience was evaluated through choice satisfaction and system effectiveness, using questionnaire items that were evaluated on 5-point Likert scales.

Chosen recipes were evaluated according to their healthiness. This was evaluated using the FSA score [27]. Each recipe was scored between 4 and 12, where 4 indicated that all four nutrients (sugar, fat, saturated fat, salt) adhered to nutritional guidelines per 100g [9, 28], while 12 would indicate that a recipe was unhealthy because of all nutritional contents being too high.

The responses to the evaluation questionnaire item were submitted to a confirmatory factor analysis (CFA; see Table 2). Unfortunately, we could not infer a reliable construct for choice satisfaction, as the variance explained by the questionnaire items was too low, while Cronbach's Alpha was only acceptable (0.60). Other items were dropped from the system effectiveness aspect because of low factor loadings.

We organized the different factors (e.g., conversation time, condition factors) and aspects (i.e., system effectiveness) into a path model using Structural Equation Modelling. Figure 3 depicts the resulting model, which had decent fit statistics: $\chi^2(17) = 28.064$, $p < 0.05$, $CFI = 0.969$, $TLI = 0.954$, $RMSEA = 0.058$, $90\% - CI$: $[0.009, 0.095]$. The relevant AVEs of the aspects was sufficiently high to form a path model [29].

Our analysis revealed that the MM-Label condition with nutrition labels (MM-label) stood out in terms of how long users interacted with our chatbot. Figure 3 illustrates this, while the use of multi-modal approaches alone had no effect on the interaction or evaluation factors considered. For MM-Label, our mediation analysis suggested that in the MM-Label condition, the conversation duration was significantly longer ($p < 0.05$) than in the text-based condition . This indicated that the usage of nutrition labels affected conversation time, on top of the other modalities.

The duration of the conservation affected, in turn, the evaluation of the user. Inferred from our confirmatory factor analysis (cf. Table 2), users who interacted with the chatbot for longer periods of time indicated greater levels of system effectiveness ($p < 0.01$). This indicated that an extended engagement did not frustrate users. Instead, it indicated that they were enthusiastic about using the system. Figure 3 also shows that the healthiness of chosen recipes was not significantly related to any of the other aspects or factors. Note that the MM-Label condition led the healthiest recipe choices, but the differences with the other conditions were not significant.

4. Conclusion and Future Work

We have presented a novel chatbot-like recommender system that introduces multi-modality in interaction with user, presentation of results and explanation of the recommendations with nutrition labels in a conversational scenario. We have designed and analyzed the impact of three distinct version of our chatbot: pure textual, multi-modal (use of text and images), and multi-modal supplemented with nutritional labels.

Our experimental evaluation reveals that our chatbot is the most effective when accompanied by explanatory labels. This is indicated by the length of conversation, as well as by the user's evaluation of the system effectiveness.

Limitations to this study could be viewed from different viewpoints. In terms of analysis, we have been unable to infer the choice satisfaction evaluation aspect. Other research have demonstrated that decision satisfaction is a good predictor of post-interaction engagement with selected item, such as for household energy con-

the project's GitHub repository.

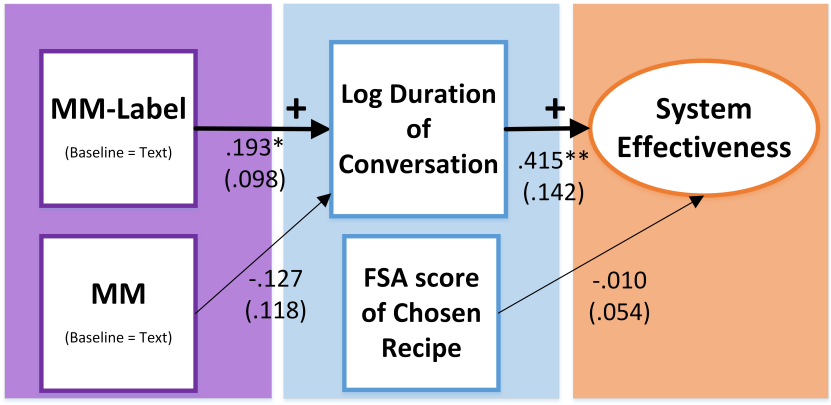

Figure 3: Structural Equation Model (SEM). Numbers on the arrows represent the β-coefficients, standard errors are denoted between brackets. Effects between the subjective constructs are standardized and can be considered as correlations, other effects show regression coefficients. Aspects are grouped by color: Objective system aspects are purple, behavioral indicators are blue (note: the FSA score represents recipe unhealthiness) and experience aspects are orange. The thinner arrows are non-significant relations, in addition: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

servation [30]. Moreover, rather than relying solely on system-driven interaction, it might be intriguing and natural to investigate *user-driven* scenarios in which users might query the system with an image and textual query. The food categories considered in this work (pasta, salad, dessert, snack) could additionally be expanded to include more meal categories and their combinations, such as to create a complete meat (first dish, second dish and vegetables). On top of that, the distinctions between various label modalities are an additional intriguing topic we wish to investigate more in-depth [31].

References

[1] H. Zamani, J. R. Trippas, J. Dalton, F. Radlinski, Conversational information seeking, arXiv preprint arXiv:2201.08808 (2022).

[2] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys 54 (2022) 1–36. doi:10.1145/3453154.

[3] P. Cordero, M. Enciso, D. López, A. Mora, A conversational recommender system for diagnosis using fuzzy rules, Expert Systems with Applications 154 (2020) 113449. doi:10.1016/j.eswa.2020.113449.

[4] D. Griol, J. Milina, From voicexml to multimodal mobile apps: development of practical conversational interfaces, ADCAIJ Adv. Distrib. Comput. Artif. Intell. J. 5 (2016) 43.

[5] F. Narducci, P. Basile, M. de Gemmis, P. Lops, G. Semeraro, An investigation on the user interaction modes of conversational recommender systems for the music domain, User Model. User Adapt. Interact. 30 (2020) 251–284. URL: https://doi.org/10.1007/s11257-019-09250-7. doi:10.1007/s11257-019-09250-7.

[6] A. Iovine, F. Narducci, G. Semeraro, Conversational recommender systems and natural language: : A study through the converse framework, Decis. Support Syst. 131 (2020) 113250. URL: https://doi.org/10.1016/j.dss.2020.113250. doi:10.1016/j.dss.2020.113250.

[7] C. Trattner, D. Elsweiler, Food recommendations, in: Collaborative recommendations: Algorithms, practical challenges and applications, World Scientific, 2019, pp. 653–685.

[8] R. Y. Toledo, A. A. Alzahrani, L. Martinez, A food recommender system considering nutritional information and user preferences, IEEE Access 7 (2019) 96695–96711.

[9] A. D. Starke, M. C. Willemsen, C. Trattner, Nudging healthy choices in food search through visual attractiveness, Frontiers in Artificial Intelligence 4 (2021) 621743.

[10] E. J. Van Loo, C. Grebitus, J. Roosen, Explaining attention and choice for origin labeled cheese by means of consumer ethnocentrism, Food Quality and Preference 78 (2019) 103716.

[11] Y. Peng, J. B. Jemott III, Feast for the eyes: Effects of food perceptions and computer vision features on food photo popularity., International Journal of Communication (19328036) 12 (2018).

[12] C. Zhou, Y. Jin, K. Zhang, J. Yuan, S. Li, X. Wang, Musicrobot: Towards conversational context-aware music recommender system, in: International Conference on Database Systems for Ad-

vanced Applications, Springer, 2018, pp. 817–820.
[13] J. Schaffer, T. Hollerer, J. O'Donovan, Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems, in: The Twenty-Eighth International Flairs Conference, 2015, pp. 507–512.
[14] Y. Deldjoo, M. Schedl, P. Cremonesi, G. Pasi, Recommender systems leveraging multimedia content, ACM Computing Surveys (CSUR) 53 (2020) 1–38.
[15] Y. Deldjoo, J. R. Trippas, H. Zamani, Towards multimodal conversational information seeking, in: Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval, 2021, pp. 1577–1587.
[16] R. G. Sousa, P. M. Ferreira, P. M. Costa, P. Azevedo, J. P. Costeira, C. Santiago, J. Magalhaes, D. Semedo, R. Ferreira, A. I. Rudnicky, et al., ifetch: Multimodal conversational agents for the online fashion marketplace, in: Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI, 2021, pp. 25–26.
[17] L. Liao, L. H. Long, Z. Zhang, M. Huang, T.-S. Chua, Mmconv: an environment for multimodal conversational search across multiple domains, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 675–684.
[18] S. Moon, S. Kottur, P. A. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho, et al., Situated and interactive multimodal conversations, arXiv preprint arXiv:2006.01460 (2020).
[19] Y. Yuan, W. Lam, Conversational fashion image retrieval via multiturn natural language feedback, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 839–848.
[20] D. Elsweiler, H. Hauptmann, C. Trattner, Food recommender systems, in: Recommender Systems Handbook, Springer, 2022, pp. 871–925.
[21] S. Barko-Sherif, D. Elsweiler, M. Harvey, Conversational agents for recipe recommendation, in: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, 2020, pp. 73–82.
[22] A. Steinfeld, O. C. Jenkins, B. Scassellati, The oz of wizard: simulating the human for interaction research, in: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, 2009, pp. 101–108.
[23] Á. Mendes Samagaio, H. Lopes Cardoso, D. Ribeiro, A chatbot for recipe recommendation and preference modeling, in: EPIA Conference on Artificial Intelligence, Springer, 2021, pp. 389–402.
[24] Á. M. Samagaio, H. Lopes Cardoso, D. Ribeiro, Enriching word embeddings with food knowledge for ingredient retrieval, in: 3rd Conference on Language, Data and Knowledge (LDK 2021), Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.
[25] A. El Majjodi, A. D. Starke, C. Trattner, Nudging towards health? examining the merits of nutrition labels and personalization in a recipe recommender system, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 48–56.
[26] Department of Health and Social Care UK, Front of Pack nutrition labelling guidance, 2016. URL: https://www.gov.uk/government/publications/front-of-pack-nutrition-labelling-guidance.
[27] D. of Health UK, F. S. Agency, Guide to creating a front of pack (fop) nutrition label for pre-packed products sold through retail outlets (2016). URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566251/FoP_Nutrition_labelling_UK_guidance.pdf.
[28] C. Trattner, D. Elsweiler, Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems, in: Proceedings of the 26th international conference on world wide web, ACM, New York, NY, USA, 2017, pp. 489–498.
[29] B. P. Knijnenburg, M. C. Willemsen, Evaluating recommender systems with user experiments, in: Recommender systems handbook, Springer, 2015, pp. 309–352.
[30] A. Starke, M. Willemsen, C. Snijders, Effective user interface designs to increase energy-efficient behavior in a rasch-based energy recommender system, in: Proceedings of the eleventh ACM conference on recommender systems, 2017, pp. 65–73.
[31] Y. Deldjoo, M. Schedl, B. Hidasi, Y. Wei, X. He, Multimedia recommender systems: Algorithms and challenges, in: Recommender systems handbook, Springer, 2022, pp. 973–1014.