A systematic review on food recommender systems

Citation for published version (APA):

DOI:
10.1016/j.eswa.2023.122166

Document status and date:
Published: 15/03/2024

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
A systematic review on food recommender systems

Jon Nicolas Bondevik a, Kwabena Ebo Bennin a,c, Önder Babur a,b, Carsten Ersch c

a Information Technology Group, Wageningen University and Research, Wageningen, 6706KN, The Netherlands
b Digital Research & Development, Friesland Campina, Wageningen, 6708WH, The Netherlands
c Department of Mathematics and Computer Science, Eindhoven University of Technology, Groene Loper, Eindhoven, 5612AZ, The Netherlands

ARTICLE INFO

Keywords:
- Systematic literature review
- SLR
- Recommender system
- Food recommender system
- Food recommendation

ABSTRACT

The Internet has revolutionised the way information is retrieved, and the increase in the number of users has resulted in a surge in the volume and heterogeneity of available data. Recommender systems have become popular tools to help users retrieve relevant information quickly. Food Recommender Systems (FRS), in particular, have proven useful in overcoming the overload of information present in the food domain. However, the recommendation of food is a complex domain with specific characteristics causing many challenges. Additionally, very few systematic literature reviews have been conducted in the domain on FRS. This paper presents a systematic literature review that summarises the current state-of-the-art in FRS. Our systematic review examines the different methods and algorithms used for recommendation, the data and how it is processed, and evaluation methods. It also presents the advantages and disadvantages of FRS. To achieve this, a total of 67 high-quality studies were selected from a pool of 2,738 studies using strict quality criteria. The review reveals that the domain of food recommendation is very diverse, and most FRS are built using content-based filtering and ML approaches to provide non-personalised recommendations. The review provides valuable information to the research field, helping researchers in the domain to select a strategy to develop FRS. This review can help improve the efficiency of development, thus closing the gap between the development of FRS and other recommender systems.

1. Introduction

With the emergence of the Internet, a lot of conveniences have been brought to all aspects of everyday life (Xie & Lou, 2022). The process of information retrieval has been influenced markedly, with the current norm diverging from using offline sources towards retrieving information online (Fischer, Stronati, & Lanari, 2017). As technology becomes increasingly available, more individuals gain access to the internet. This is causing a surge in the volume of available information, often heterogeneous in nature (Bai et al., 2019). Both the massive amount and heterogeneity of available data is far beyond the comprehension of a single user (Feng, Meng, & Zhang, 2021; Lops, De Gemmis, & Semeraro, 2011), making it difficult to retrieve correct and relevant information in this day and age (Trang Tran, Atas, Felfernig, & Stettninger, 2018).

The concept of recommender systems has become a popular solution to combat the problem of information retrieval. A recommender system aims to retrieve items a user is likely to interact with (Koren, Bell, & Volinsky, 2009). This is achieved by predicting individual preferences based on interaction with the system (Xie & Lou, 2022). Prediction of preferences enables the system to filter out non-relevant information and present the most relevant content to the user (Jannach & Jugovac, 2019). This can reduce the time spent to query information and greatly improve the efficiency of information acquisition (Xie & Lou, 2022). Recommender systems can be found in a vast amount of domains ranging from the recommendation of scientific articles and news, to entertainment such as movies and books (Ricci, Rokach, & Shapira, 2015; Trang Tran et al., 2018), e-commerce and most notably, food.

The sharing of cooking recipes has been common for thousands of years and this has naturally become prominent on the internet (Teng, Lin, & Adamic, 2012). This has led to a fast growth of food computing (Li, Zaki, & Chen, 2023). Information overload is very much present in this domain, as the popularity of recipe websites has increased over the past years (El Majjodi, Starke, & Trattner, 2022). The overflow of published recipes is making the decision process challenging for the user (El Majjodi et al., 2022). To overcome this overload of information, Food Recommender Systems (FRS) have proven to be useful (Trattner & Elsweiler, 2017).

Compared to other fields, the recommendation of food has many specific characteristics that need to be considered (Min, Jiang and
Jain, 2019). First of all, food and diet are complex domains bringing many challenges for recommendation (Trang Tran et al., 2018). Food preference is highly personal and varies considerably between individuals. Deciding what to eat is a complex process, influenced by many biological, personal, and socio-economic factors (Leng et al., 2017). These are factors such as taste preference, cultural background, and even genetic influence (Min, Jiang and Jain, 2019). Furthermore, thousands of ingredients have to be collected in order to produce the available recipes. These can be combined together in several ways, exponentially increasing the complexity of food recommendation (Freyne & Berkovsky, 2010). To further complicate matters, the studies by Elseweler, Trattner, and Harvey (2017) and Yang et al. (2015) reveal the importance of visual attributes for food recommendation (Meng, Feng, He, Gao, & Chua, 2020). To consider this, the systems must be able to use the information contained within more complex data formats such as images or videos.

As a result of the increased complexity of food recommendation systems, the research is not as widespread and common. This is causing the development of FRS to lag behind recommendation systems in other areas (Min, Jiang and Jain, 2019). The amount of research in this domain has only increased within the past decade (Trattner & Elsewiler, 2019). In addition, systematic literature reviews (SLR) in this research domain are scarce (Min, Jiang and Jain, 2019). Therefore, we believe a systematic approach is necessary to provide a thorough investigation into the state-of-the-art in this research field. By providing a thorough overview and analysis of the methods used in FRS, our review can help researchers select the correct approach for their problem. In addition, this review can be used to assist in the implementation of general and specific recommendation systems. Thus, this SLR can help improve the efficiency of development, closing the current gap in development between FRS and other domains of recommendation.

The overall contribution of this paper is an SLR with a comprehensive overview of the current state-of-the-art in FRS in all domains. This is done by summarising the many important aspects to consider during the development of these systems. The techniques used for recommendation, the data used and the preprocessing applied, the evaluation method and the metrics that are calculated, in addition to the advantages and disadvantages of food recommendation are considered in this review. Furthermore, the extent of reproducible research in this domain is assessed based on the availability of codebases and datasets. In order to extract valuable data on these aspects, a total of 67 high-quality studies were selected from a pool of 2738 studies using strict quality criteria. These papers describe the implementation of 71 FRS. Five research questions were formulated to summarise the state-of-the-art of FRS, as shown in Table 1.

The organisation of the remainder of this paper is as follows. In Section 2, a theoretical background for food recommendation is provided. In addition, related work are summarised and compared to our study. Section 3 presents the research objectives and methodology used for primary study collection. In Section 4, the results are analysed and presented to be further discussed in Section 5. Finally, a conclusion is presented together with a proposal for future research in Section 6.

Table 1

<table>
<thead>
<tr>
<th>Research question (RQ)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1</td>
<td>What are the techniques used in Food Recommendation Systems?</td>
</tr>
<tr>
<td>RQ2</td>
<td>What kind of data and preprocessing techniques are used in Food Recommendation Systems?</td>
</tr>
<tr>
<td>RQ3</td>
<td>How are the Food Recommendation Systems evaluated?</td>
</tr>
<tr>
<td>RQ4</td>
<td>To what extent is recent research on Food Recommendation Systems reproducible?</td>
</tr>
<tr>
<td>RQ5</td>
<td>What are the advantages and disadvantages of Food Recommendation Systems?</td>
</tr>
</tbody>
</table>

2. Background

2.1. Food Recommender Systems

FRS can be defined as systems with the aim of recommending food items that match certain eating preferences or characteristics (Meng et al., 2020; Trang Tran et al., 2018). These preferences are based on the interaction between users and the system (Meng et al., 2020; Xie & Lou, 2022). Based on the learned preferences, food items can be recommended by retrieval of existing items or generation of novel items. This is done by linking the preferences to various attributes of food items, such as ingredients, cooking methods and nutrition composition (Meng et al., 2020). Generally, the FRS takes input data and compares this with a database of items. The recommendation is evaluated based on an objective ground truth value or a subjective review. The general architecture is shown in Fig. 1.

At the highest level, FRS can be broadly divided into two main categories: implicit and explicit recommendation. The first category is the most common. For implicit recommendation, preferences obtained from users are utilised. These attributes are processed in such a way that they are comparable to the food items. They consist of ratings for recipes, historical data and contextual information. A similarity measure is then obtained, and similar food items are retrieved and presented to the user. This implicit recommendation caters to specific user preferences, making it inherently personalized.

The second category of FRS relies solely on attributes from food items for recommending food items. With this approach, the system is given these attributes explicitly in a query. It is not necessary to learn user preferences, as the necessary attributes for recommendation are given. Based on user preferences, the system can either retrieve similar food items or generate novel ones. These systems are not dependent on personalization, although implicit user preferences can be considered. If that is the case, different users would receive different recommendations based on the same query.

2.2. Recommendation methods

The recommendation method of an FRS describes the rationale and features used for producing the recommendations or generating recipes. The methods can be classified based on the features used for generating the recommendation. By following established taxonomy by Bai et al. (2019) and Ghanadrad et al. (2022), the methods are divided into four classes. These classes are content-based filtering, collaborative filtering, graph-based methods and hybrid methods.

2.2.1. Content-based filtering

With content-based filtering, similarities are defined based on the content of two objects. These two objects can be a user profile and an item, or two items. In the first case, items are compared to the user preferences stored in the user profile (Amami, Pasi, Stella, & Faiz, 2016; Ghanadrad et al., 2022). The user preferences can be based on queries (explicit) or historical interaction (implicit) (Pazzani & Billsus, 2007). Similar features are extracted from both the user profile and the food items. This is based on the rationale that a user is likely interested in
recommendations that are similar to known items. In the latter case, similar features are extracted from the two items to be compared. This is based on the rationale that items with similar attributes can be used interchangeably. The final recommendation is based on the similarity between these features (Lops et al., 2011). The features that are used for assessing similarity differ between the FRS and can range from ingredients to images of cooked dishes. This is the most traditional and elementary method of recommendation for general recommender systems (Bai et al., 2019).

2.2.4. Hybrid methods

Methods based on collaborative filtering exploit similarities between two user profiles in order to generate the recommendations (Bai et al., 2019; Zhang, Liu, Guo, Bai, & Gan, 2021). These methods are based on the principle that users who have similar preferences and interests tend to be interested in the same items (Ghannadrad et al., 2022). To achieve this, FRS based on collaborative filtering employ various algorithms to measure similarities between users. If several users are found to be similar, one of the users can be recommended novel items based on the user profiles of other, similar users (Bai et al., 2019).

2.2.3. Graph-based methods

Graph-based methods mainly revolve around the formation of graphs to represent the item and user relationships (Bai et al., 2019; Ghannadrad et al., 2022). With these methods, users and items are represented as vertices. The relationships between users and items are represented by connecting the vertices with edges (Bai et al., 2019). This type of recommendation is able to exploit data from a vast amount of different sources (Bai et al., 2019). In the case of food recommendation, the relationship between ingredients can be inferred based on co-occurrence in recipes. In addition, structural information on the relationships between users and recipes can be considered. Furthermore, some FRS consider both the co-occurrence of ingredients and the relationships between users and recipes to recommend food items. Finally, it is also possible to map relationships between users based on certain features. This opens the possibility of using collaborative filtering to retrieve recommendations. Graph recommendation consists of two phases: the construction of the graph, and the generation of recommendations based on the graph (Bai et al., 2019).

2.2.4. Hybrid methods

As suggested by the name, this class of recommendation methods can combine two or more recommendation methods to leverage their individual strengths. Hybrid methods also allow the different methods to alleviate the shortcomings of each other (Bai et al., 2019). This can improve the accuracy and performance of the recommendations (Bai et al., 2019; Ghannadrad et al., 2022). However, the increased performance relies on how the methods are combined together (Bai et al., 2019). This class of methods is able to base the recommendation on multiple sources of information (Bai et al., 2019). As a result, this method is able to provide comprehensive and accurate recommendations to users.

2.3. Recommendation algorithms

While the method is based on the overall rationale and features used for recommendation, the algorithm describes how the specific recommendations are calculated and retrieved. Three main classes of algorithms are defined as machine learning, statistics and querying.

2.3.1. Machine learning

With this class of algorithms, the recommendations are retrieved based on machine learning. In order to do this, a machine learning algorithm is implemented in order to predict certain features of the user or the food items. These are features such as a user’s score for a recipe or a recipe’s missing ingredient.

2.3.2. Statistics

With statistics, various statistical measures are employed in order to calculate similarities. The similarities can be measured between a user and a food item, two users or two food items, depending on the recommendation method. The calculated measure is then used to select recommendations, which are then presented to the user.

2.3.3. Query

With an algorithm based on queries, the recommendations are retrieved directly from a database consisting of food items. The queries are designed based on predefined rules to filter and sort the data in the correct manner. In addition, one can employ specific database tools in a query in order to predict various relationships between users, food items or both.

2.4. Evaluation

Evaluation is the process of measuring the quality of the recommendations. There are two important aspects to consider when evaluating the FRS. First, it is necessary to decide which evaluation method to use. Furthermore, it is important to choose the correct evaluation metrics. While the method describes how the evaluation is performed, the metrics describe the specific aspects of a system that are measured. Following the taxonomy by Ghannadrad et al. (2022), the evaluation methods are divided into Online-, Offline-, and Hybrid-methods. For online evaluation, a user’s experience of the system and recommendations are measured. The experience is based on interactions between users and the system (Vrijenhoek et al., 2021), and is captured with
questionnaires or interviews. This method of evaluation is the most desired since it provides accurate results of the performance in real-world use cases (Silveira, Zhang, Lin, Liu, & Ma, 2019). For offline evaluation, a specific dataset is used to assess the accuracy of the recommendations (Ghannadrad et al., 2022). A portion of the dataset is used for training the system, while a different portion is used for evaluation (Silveira et al., 2019). In most cases, the recommendations are compared with ground truth values. This method of evaluation is less costly and can be automated, but requires specific evaluation metrics.

Three main classes of evaluation metrics can be defined based on the current research. These classes are Accuracy-based, Ranking-based and Error-based. The metrics that do not fit the defined classes are extracted as Miscellaneous.

Accuracy-based metrics capture how close the recommendations are to the ground truth values. As a result, it is often used for classification problems (Gallo, Landro, La Grassa, & Turconi, 2022). Accuracy, Precision, Recall and F1 are commonly used measures. Precision captures the proportion of recommendations that are relevant, while recall is defined as the number of relevant recommendations that are retrieved (Ghannadrad et al., 2022; Ng & Jin, 2017; Powers, 2016; Rostami, Oussalah, & Farrahi, 2022). The F1 measure is the harmonic mean between precision and recall (Ghannadrad et al., 2022; Powers, 2016; Wang, Li, Pavlu, & Aslam, 2018). Together, these metrics provide a picture of the quality of the recommendations.

Ranking-based metrics measure how many times a relevant recipe is found within a certain range of the top recommendations or predictions (Ghannadrad et al., 2022). Examples of ranking-based metrics are Normalised Discounted Cumulative Gain, Mean Reciprocal Rank, and Average Precision. Common for all metrics is that they provide an intuitive sense of how quickly the best recipe can be located (Chen, Ngo and Chua, 2017; Chen, Pang and Ngo, 2017).

Error-based metrics evaluate performance by measuring the difference between predicted and ground truth values (Ghannadrad et al., 2022). These metrics are useful for regression-like problems (Gallo et al., 2022), and can be useful for FRS where the end goal is to predict numerical features. These are features like the pairing scores (Park, Kim, Park, Shin, & Kang, 2019) and amount of ingredients (Li et al., 2021). However, this class can also be used on numerical measures like the percentage of relevant recommendations and the recommendation scores.

2.5. Related work

Numerous attempts have been made to examine the corpus of studies on FRS. In this section, we summarise related surveys and compare them to our SLR in order to situate our SLR within the context of FRS research. In their survey, Min, Jiang, Liu, Rui and Jain (2019) formalise the term food computing. This term is defined as the perception, recognition, retrieval, recommendation and monitoring of food. By considering about 300 papers, they identify the methods used and the problems solved with the recommender systems. The applications, tasks and data sources used for food computing are categorised. In addition, they summarise challenges and suggest possible future research directions. The same authors propose a unified framework for food recommendation in another survey (Min, Jiang and Jain, 2019). Approaches for personalization, unique food characteristics, incorporation of contextual information and other factors affecting the recommendations are identified. They provide a summary of existing methods, current challenges and future research directions. The databases used for retrieval and the number of selected studies are not mentioned in this survey.

Most of the related works consider systems proposed in a specific domain, and the health domain is frequently considered. In their SLR, Abhari et al. (2019) define and consider nutrition-based recommender systems. These are FRS that are based on health in addition to a user’s preferences. By selecting 25 studies from a pool of 781 they identify the different types of systems, AI techniques and modules used for recommendation. In addition, they describe the platforms where such systems are used. They find that hybrid recommendation with a rule-based technique is the most common, with mobile as the most common platform. De Croon et al. (2021) select 73 studies from a pool of 2397 in order to identify the current state of the art, the main trends and current gaps of FRS in the same domain. They characterise several subdomains, recommendation techniques and evaluation methods. Furthermore, they consider how the recommendations are presented to the user. In the end, several guidelines are proposed in an effort to standardise the reporting of research in this domain. In order to identify techniques and algorithms used to suggest diet plans, Jain and Singhal (2022) collect 15 of 353 studies. Four categories of FRS are defined based on the information considered during the recommendation. These categories are similar user behaviour, user preferences for item features, hybrid filtering and knowledge-based filtering.

The recommendation of healthy diets is also considered by Kallel, Kanoun, and Dhoub (2022). They use 13 papers to identify the methods used, user types and various constraints considered by the systems. Two categories of recommendation methods are defined as exact and approximation. User types are categorised as elderly, teens, athletes, sick patients and other. Furthermore, they find that nutrient and cost are the most common constraints that are considered. Kirk, Catal, and Tekinerdogan (2021) consider 60 studies where machine learning is used for precision nutrition. They define regression, classification, recommendation and clustering as machine learning tasks and identify 30 machine learning algorithms. Several categories of evaluation methods and metrics are proposed. Systems exploiting the internet of medical things are considered by Soma and Dyapur (2022). By reviewing various machine learning algorithms and their performance, they find that deep learning yields better results than what they define as classic machine learning algorithms. The number of retrieved studies and the databases that were searched in order to produce their findings are not mentioned. This is also the case in a survey by Trang Tran et al. (2018). The researchers provide an overview of techniques to assess the state of the art of recommendation in the same domain. A high-level classification is proposed based on whether the system caters to individuals or groups. They define four types of systems based on the factors considered during the recommendation. These subtypes are user preferences, nutritional needs or a balance of both, in addition to systems catering to groups. In another SLR, Yera, Alzahrani, Martinez, and Rodriguez (2023) evaluate systems designed for diabetic patients. 34 studies are analysed to identify the state of the art, the recommendation methods and the evaluation methods. A taxonomy is proposed based on the terms used in the abstracts. They define four methods for recommendation, which are semantic-based, optimisation-based, rules-based and classification-based, and interaction-based. The individual strengths and weaknesses of each approach are identified.

By considering the domain of big data, Chavan, Thoms, and Isaacs (2021) investigate its use in FRS. They assess the emergence of literature in this domain, and analyse many aspects of the 38 selected publications. These are aspects such as the number of citations and authors, in addition to the theoretical underpinnings and themes of systems. They define nine different themes, challenges and possible research questions that can be raised for food recommendation based on big data. However, the different methods used for recommendation are not considered. In an SLR by Wang et al. (2021), the use of graphs for recommendation is evaluated. They coin the term Graph learning based recommender systems and identify its three types. These categories are based on the information considered in the system, which can come from interaction data, sequential interaction data or side information. Three recommendation methods are defined as random walk, graph embedding and graph neural networks. Additionally, they classify the type of data that is used based on its characteristics and source. However,

J.N. Bondevik et al.

Export Systems With Applications 238 (2024) 122166
they do not mention the number of primary studies considered in their survey.

The collection of studies is not described in some cases, leaving the number of studies and databases used unknown to the reader. Where the collection is described, many of the related surveys base their findings on a small number of studies. This introduces some uncertainty to the findings by making it difficult to infer the possibility of bias during primary study collection. Furthermore, little evidence means that the findings are not as reliable. In addition, most surveys only consider one domain. While this might provide a higher level of detail, it is difficult to get an overview of food recommendation as a whole. In contrast to the related work, the methods used for collection, selection and analysis are described rigorously in this SLR. This makes it possible for other researchers to evaluate the possibility of bias in our research. Furthermore, it provides a template for how SLRs can be conducted in the research of FRS. This SLR takes a more broad approach in order to classify the current state of the art of FRS in all domains. This opens up for a greater study collection as more studies are available. As a result, an overview based on a larger number of existing attempts is produced. While the results might lose some granularity, the findings of this SLR are more reliable.

3. Research objectives and method

We followed a systematic approach to collect, classify and analyse the current body of knowledge on FRS. There is little published research with attempts to systematically review research of FRS. Therefore, we mainly based the research method on established strategies for performing SLRs (Ali & Petersen, 2014; Keele et al., 2007; Wohlin, 2014; Wright, Brand, Dunn, & Spindler, 2007), previous experiences from conducting SLRs (Brereton, Kitchenham, Budgen, Turner, & Khalil, 2007; Staples & Niazi, 2007) and previous SLRs conducted in the field of Software Engineering (Ali, Babar, Chen, & Stol, 2010; Giray, Bennin, Köksal, Babur, & Tekinerdogan, 2023; Gurbuz & Tekinerdogan, 2018; Tummers, Kassahun, & Tekinerdogan, 2019; Tüzün, Tekinerdogan, Kalender, & Bilgen, 2015; Villegas, Sánchez, Díaz-Cely, & Tamura, 2018). By doing this, we ensured the robustness of our systematic method.

Initially, we defined a review protocol. This is a necessary step in order to reduce the possibility of bias in the selection and analysis of studies (Keele et al., 2007). The summarised protocol, shown in Table 2, was based on a previously established format by Motta, de Oliveira, and Travassos (2018). The review process consisted of four main parts, which are described in more detail throughout this section:

1. Definition of goal and research questions.
2. Selection of databases and collection of primary studies.
3. Definition of extraction points and the extraction of these.
4. Analysis, synthesis and reporting of the data.

3.1. Goal and scope

The goal of this SLR is to explore and analyse the current state-of-the-art on FRS. We followed the established review protocol summarised in Table 2 in order to accomplish this.

The scope of this study encompasses five principal aspects of FRS. The first aspect considers the diverse techniques employed in FRS. The techniques consist of the different possible combinations of recommendation methods and algorithms, as well as the utilisation of personal attributes. Another important aspect of FRS is the data employed for food recommendations. This aspect comprises multiple features such as the source, format, size, features, preprocessing and representation of the data. The third aspect of this study is related to the evaluation of FRS, encompassing not only the evaluation method but also the metrics that are measured and calculated. The fourth aspect is associated with the repeatability of research in the domain of food recommendation. The aspect is based on the availability of the code used for the experiments, as this is necessary to reproduce the experiments and results of a study. The last aspect in the scope is the advantages and disadvantages of FRS. In order to ensure that current works are reviewed in this SLR, we defined a threshold for the publication dates. As such, we only considered research published after 2017.

3.2. Research questions

The research questions drive the entire systematic review (Staples & Niazi, 2007), and are described by Brereton et al. (2007) as the most critical element of any systematic review. This is due to the fact that the questions set the basis for the entire research methodology (Keele et al., 2007). An early definition is recommended in order to reduce bias throughout the research process (Wright et al., 2007). Therefore, the research questions were defined after establishing the goal and the scope of this study.

In order to make the method as reliable as possible, it is necessary to formulate research questions that can capture the entire scope of this study. Furthermore, Staples and Niazi (2007) recommends choosing clear and narrow research questions in order to limit the scope of the literature review. The questions are based on criteria defined by Keele et al. (2007). In addition, previous SLRs conducted in the field of Software Engineering are used for inspiration. These are studies by Ali et al. (2010), Giray et al. (2023), Gurbuz and Tekinerdogan (2018), Tummers et al. (2019), Tüzün et al. (2015) and Villegas et al. (2018).

As in the research protocol presented by Motta et al. (2018), the population, intervention and outcomes were defined. The population of this research is defined as FRS and the intervention as the different recommendation techniques. The outcomes are defined as the advantages and disadvantages of the FRS. By following guidelines and taking inspiration from other SLRs, we have ensured that the research questions are clear and specific. This provides a clearly defined scope for this research, in which valuable knowledge on FRS can be extracted. Five research questions were formulated, as shown in Table 1. The relevant aspects these research questions explore are then linked to the general architecture of FRS in Fig. 1.

<table>
<thead>
<tr>
<th>Research questions</th>
<th>RQ1 What are the techniques used in Food Recommendation Systems?</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ2 What kind of data and preprocessing techniques are used in Food Recommendation Systems?</td>
<td></td>
</tr>
<tr>
<td>RQ3 How are the Food Recommendation Systems evaluated?</td>
<td></td>
</tr>
<tr>
<td>RQ4 To what extent is recent research on Food Recommendation Systems reproducible?</td>
<td></td>
</tr>
<tr>
<td>RQ5 What are the advantages and disadvantages of Food Recommendation Systems?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search string</th>
<th>Population: Food recommendation systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>search string</td>
<td>Intervention: Recommendation method</td>
</tr>
<tr>
<td>food AND (system OR model OR network) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*) NOT safety NOT calculator</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search strategy</th>
<th>Database search: ACM, IEEE, ISI, SCOPUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward snowballing using Google Scholar</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>IC1 Full text is available.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC2 Paper is written in English.</td>
<td></td>
</tr>
<tr>
<td>IC3 Paper describes a food recommendation system.</td>
<td></td>
</tr>
<tr>
<td>IC4 Paper is peer-reviewed.</td>
<td></td>
</tr>
<tr>
<td>IC5 Paper is a primary study.</td>
<td></td>
</tr>
<tr>
<td>IC6 Paper is not a duplicate publication already considered.</td>
<td></td>
</tr>
<tr>
<td>IC7 Paper is not a demo paper.</td>
<td></td>
</tr>
</tbody>
</table>

| Study type | Primary studies. |
3.3. Primary study selection

An overview of the primary study selection process is shown in Fig. 2. The process consists of six steps which are described in the sections below.

3.3.1. Database selection

The primary study selection was carried out on literature retrieved from ACM Digital Library, IEEE Xplore, ISI Web of Science and SCOPUS. These databases were selected in order to get a good coverage of the existing literature. In addition, they were based on the previously mentioned SLRs conducted in Software Engineering (Ali et al., 2010; Giray et al., 2023; Gurbuz & Tekinerdogan, 2018; Tummers et al., 2019; Tüzün et al., 2015; Villegas et al., 2018).

3.3.2. Search string generation

An initial exploration of available research revealed important keywords and technical jargon used for FRS. A search string prototype was generated based on these findings. The ISI Web Of Science database was used to evaluate the performance of the search string, as it is the most concise database of the selected databases. As such, the effect of including or excluding keywords from the search string was well demonstrated here. Furthermore, a satisfactory yield in this database ensures a satisfactory yield from the other databases as well. Initially, many false positives were retrieved, leading to low precision. An iterative approach was therefore taken to enhance the precision and recall obtained with the search string, as shown in the Appendix in Table A.1.

During this process, boolean operators were used as a way to assess the performance of specific keywords. This was done by adding the AND operator followed by the keyword to assess. If no relevant papers were retrieved with the addition of the keyword, it was added to the search string together with the NOT operator. This improves the precision of the results, as many false positives are removed. With the addition of safety to the search string, 53 papers were retrieved. Most papers were related to evaluating the safety of certain foods and ingredients. In addition, research into the safety of bioactive compounds, and how they can be used to combat different diseases was retrieved. Therefore, no research related to FRS was to be found in the results. This was also the case for calculat*, where 26 papers were retrieved. The results were about the prediction of nutritional values, digestibility and other aspects of nutrients not related to human food. These two keywords were added after the NOT operator, producing a search string with the general syntax of food AND (network OR system OR model) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*) NOT safety NOT calculat*. This search string provided a satisfactory output with good precision and recall.

Table 3

<table>
<thead>
<tr>
<th>Database</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM Digital Library</td>
<td>100</td>
</tr>
<tr>
<td>IEEE Xplore</td>
<td>104</td>
</tr>
<tr>
<td>ISI Web of Science</td>
<td>807</td>
</tr>
<tr>
<td>SCOPUS</td>
<td>1727</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2738</td>
</tr>
</tbody>
</table>

3.3.3. Database retrieval

We performed a search in each of the four selected databases using the search string. The actual search string that was used differed slightly based on the standards of the different databases, but the same keywords and boolean logic were used. The search was matched against the title, abstract and keywords of the papers. As the aim of the research usually is described in the abstract, this ensured that only research with an aim related to the search string was identified. We conducted the search in October 2022, and a total of 2738 papers were retrieved. The outcomes of the different databases are shown in Table 3.

3.3.4. Inclusion criteria

Strict inclusion criteria were specified in order to retrieve a subset of high-quality papers from the outcomes of the different databases. Using only high-quality research to answer the research questions assures a certain level of validity to the results. Seven inclusion criteria were defined based on the previously mentioned strategies, experiences and SLRs (Ali et al., 2010; Ali & Petersen, 2014; Brereton et al., 2007; Giray et al., 2023; Gurbuz & Tekinerdogan, 2018; Keele et al., 2007; Staples & Niazi, 2007; Tummers et al., 2019; Tüzün et al., 2015; Villegas et al., 2018; Wohlin, 2014; Wright et al., 2007).

The inclusion criteria defined in Table 4 were applied to all studies retrieved by the search string. Initially, only the abstract and title of the papers were considered. However, it is important to note that the standard of abstracts of research in information technology and software engineering is poor (Brereton et al., 2007). Therefore, it is necessary to look beyond the title and abstract on some occasions. If this was the case, a three-step strategy proposed by Ali and Petersen (2014) was used. These three steps describe three sections to review if no definite decision can be made from the abstract. The strategy is described in Table 5. After applying inclusion criteria to the retrieved research, we were left with 83 of the primary studies.

3.3.5. Quality assessment

It is critical to assess the quality of primary studies (Keele et al., 2007). A quality assessment will help support the study-selection process (Brereton et al., 2007), and improve the reliability of the SLR (Ali et al., 2010; Ali & Petersen, 2014; Brereton et al., 2007; Giray et al., 2023; Gurbuz & Tekinerdogan, 2018; Staples & Niazi, 2007; Tummers et al., 2019; Tüzün et al., 2015; Villegas et al., 2018; Wohlin, 2014; Wright et al., 2007).
Applying the inclusion criteria, 22 of the collected references remained.

The publication date of the references in the primary studies was increased precision, only forward snowballing was used for this review. Forward and backward snowballing is shown to exhibit similar effectiveness (Giray et al., 2023). However, forward snowballing has been shown to retrieve fewer non-relevant papers than its backward counterpart (Badampudi, Wohlin, & Petersen, 2015). Because of this increased precision, only forward snowballing was used for this review.

As recommended by Wohlin (2014), snowballing was conducted with Google Scholar in order to further improve the recall of papers. Forward and backward snowballing is shown to exhibit similar effectiveness (Giray et al., 2023). However, forward snowballing has been shown to retrieve fewer non-relevant papers than its backward companion (Badampudi, Wohlin, & Petersen, 2015). Because of this increased precision, only forward snowballing was used for this review. The publication date of the references in the primary studies was checked against the threshold, and 27 papers were retrieved. After applying the inclusion criteria, 22 of the collected references remained.

The quality was assessed, and 15 papers scored above the threshold. In the end, 67 high-quality primary studies were selected based on the primary study selection. An overview of the selected primary studies from each source is shown in Table 7. The primary studies excluded during the quality assessment can be seen in the Appendix.

### 3.4. Data extraction

After the collection and review of primary studies, the collection of metadata was initiated. Extraction points are defined based on the research questions defined in 3.2. The resulting extraction form, shown in Table 8, consists of 23 extraction points. With the first seven extraction points, metadata describing the primary studies is extracted. The remaining 16 extraction points collect information on FRS that is necessary to answer the research questions. Four input types are collected with these extraction points. The personalization of FRS is recorded as binary data, as personalization is either considered or not. The advantages and disadvantages are recorded as free text for further analysis. The dataset size is stored as integers. The remaining extraction points record categorical data. These categories were obtained during the extraction and refined iteratively. For the extraction point considered for RQ5, the location of available repositories is collected. It was not possible to summarise this data, as the location for each repository is unique. Therefore, the locations were used to produce binary data describing whether the research is open source or not.

### 3.5. Data analysis

After extracting data from the primary studies, the results were analysed. For the categorical and binary data, the frequencies and percentages of each category are reported. For the numerical data, the distribution is reported. The collected data on the advantages and
disadvantages of each FRS requires qualitative analysis. The reported advantages and disadvantages are summarised and categorised. Each category is reported in the context of the primary studies where it is stated.

4. Results

The synthesised data is reported in this section. General information about the primary studies is provided first. This is done in order to give the reader an overview of the papers included in this study. Afterwards, the more specific results are presented in order to answer the research questions defined in Section 3.2. The results are presented in the order and context of the research questions.

4.1. General statistics

The total amount of primary studies initially retrieved for this literature review is 2738, shown in Table 3. Applying the inclusion criteria shown in Table 4 reduces the amount to 105 primary studies. By applying the quality criteria from Table 6 and calculating the scores, the primary studies obtain an average score of 13.72 with a standard deviation (SD) of 4.64. The distribution of quality scores is shown in Fig. 3, with the quality threshold illustrated by a black vertical line. A total of 67 papers score above the threshold. 38 (56.7%) of the primary studies are conference proceedings, while the remaining 29 (43.3%) are articles.

The amount of primary studies published per year is shown in Fig. 4. The lowest number of publications is found in 2017, with six (9.0%) papers. 16 (23.9%) FRS studies are published in 2021, making this year the most active of the last five years. There is a slight decrease in research efforts in 2019, 2020 and 2022. However, there is an overall increase in research efforts from 2017 to 2022.

The number of papers retrieved from the different sources is shown in Fig. 5. It is shown that ISI Web Of Science contains the largest body of research into FRS, with 17 (25.4%) of the total 67 papers. Research from the past five years is retrieved, and a generally increasing trend of publications can be inferred. There is a dip in publications in 2019 and 2020. A total of 15 (22.4%) papers are collected by conducting forward snowballing. The largest amount of the retrieved primary studies are from 2019, and no papers from 2022 are retrieved. SCOPUS contains 14 (20.9%) of the papers. All years are represented, and an increasing trend of publications can be inferred. ACM yields 11 (16.4%) papers. No studies from 2019 are retrieved, and the distribution indicates a decreasing trend of publications. Finally, 10 (14.9%) papers are retrieved from IEEE. This database contains papers from all years except 2022.

A total of 227 authors have contributed to the 67 publications included in this study. The average amount of published papers per author is 1.19, with a SD of 0.55. The number of publications for authors with more than two publications is shown in Table 9. Four (1.8%) authors are involved in four primary studies. Five (2.2%) authors are part of three studies, and 22 (9.7%) authors contribute to two studies. The remaining majority of 196 (86.3%) authors contribute to one study.

4.2. RQ1 what are the techniques used in Food Recommendation Systems?

In order to learn about the techniques used for recommending food, this section explores three important aspects of the FRS. These aspects
The yearly distribution of the recommendation methods used in FRS is shown in Fig. 6. The figure shows that FRS with content-based filtering is used in 37 (52.1%) FRS. All years are represented in the papers. Graph-based methods are used in 15 (21.1%) of the FRS. This category of methods is proposed in all years except 2017 and 2020. 12 (16.9%) of the FRS employ collaborative filtering. No papers published in 2017 did propose these methods. The remaining seven (9.9%) FRS are based on hybrid methods. No papers published in 2018 and 2022 are based on these methods.

4.2.2. Personalization

Personalization describes whether personal user attributes are used to generate recommendations. Where non-personalized FRS base the recommendations on queries or food items, personalized recommendation is based on preferences learned from historical interaction (Zhang et al., 2021) and personal attributes. These preferences are then used to generate the recommendations (Bai et al., 2019). The use of personalization in the retrieved FRS is shown in Table 10. It is shown that 26 (50.7%) FRS are not personalized. 35 (49.3%) of systems are personalized.

The personalization is separately assessed for each group of recommendation methods. Among content-based FRS, 28 (73.68%) do not consider personal attributes, while the remaining nine (26.32%) are personalized. With collaborative-based filtering, three (25%) of the FRS are not personalized. The remaining nine (75.0%) systems are personalized. Among FRS employing graph-based methods, four (26.67%) FRS are not personalized. Personalization is applied in the remaining 11 (73.33%) systems. One (14.3%) system based on a hybrid method is not personalized, while the remaining six (85.7%) systems consider personal attributes.

4.2.3. Algorithm

The distribution of algorithms for each class of methods is shown in Fig. 8. The most common algorithm in all categories of methods is machine learning, used in 49 (64.79%) FRS in total. Statistical measures are used to generate recommendations in 15 (21.13%) FRS. Both algorithms can be found in FRS from all methods. The remaining two (2.82%) FRS do not describe the algorithm used. The systems proposed by Adaji, Sharmane, Debrowney, Oyibo, and Vassileva (2018) and Clunis (2019) employ graph-based methods. All classes of algorithms can be found in the FRS employing content-based filtering and graph-based methods. For FRS based on collaborative filtering or hybrid methods, the recommendations depend on machine learning or statistics.

4.2.4. Combinations

Since the aspects described above are necessary to consider when generating the recommendations, all FRS are built on certain combinations of these. To get an overview of the overall techniques used in FRS, these combinations are assessed. A total of 18 different combinations are found, and 12 combinations are used more than once. These 12 combinations are presented in Table 11. Six of them are described in this section. 20 (28.17%) FRS are based on a content-based filtering method, where non-personalized recommendations are predicted with machine learning algorithms. The combination of a graph-based method with personalized recommendations based on machine learning is used in eight (11.3%) FRS. Seven (9.86%) FRS generates personalized recommendations with machine learning, using a content-based filtering method. A combination of collaborative filtering, personalization and a machine learning algorithm is used in six (8.45%) FRS. Six (8.45%) FRS also consist of a content-based filtering method where personalization is not considered and recommendations are predicted.
Table 11
The combinations of methods, personalization and algorithms used more than once.

<table>
<thead>
<tr>
<th>Method</th>
<th>Personalization</th>
<th>Algorithm</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-based</td>
<td>Not personalized</td>
<td>Machine learning</td>
<td>20</td>
</tr>
<tr>
<td>Graph-based method</td>
<td>Personalized</td>
<td>Machine learning</td>
<td>8</td>
</tr>
<tr>
<td>Content-based</td>
<td>Personalized</td>
<td>Machine learning</td>
<td>7</td>
</tr>
<tr>
<td>Collaborative filtering</td>
<td>Personalized</td>
<td>Machine learning</td>
<td>6</td>
</tr>
<tr>
<td>Content-based</td>
<td>Not personalized</td>
<td>Statistics</td>
<td>6</td>
</tr>
<tr>
<td>Hybrid method</td>
<td>Personalized</td>
<td>Machine learning</td>
<td>5</td>
</tr>
<tr>
<td>Collaborative filtering</td>
<td>Personalized</td>
<td>Statistics</td>
<td>3</td>
</tr>
<tr>
<td>Graph-based method</td>
<td>Not personalized</td>
<td>Query</td>
<td>2</td>
</tr>
<tr>
<td>Content-based</td>
<td>Personalized</td>
<td>Statistics</td>
<td>2</td>
</tr>
<tr>
<td>Collaborative filtering</td>
<td>Not personalized</td>
<td>Query</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 12
The sources used in more than one system.

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
<th>Percentage of FRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allrecipes</td>
<td>21</td>
<td>29.58%</td>
</tr>
<tr>
<td>Food</td>
<td>11</td>
<td>15.49%</td>
</tr>
<tr>
<td>Yummly</td>
<td>8</td>
<td>11.27%</td>
</tr>
<tr>
<td>Recipe1M</td>
<td>8</td>
<td>11.27%</td>
</tr>
<tr>
<td>USDA</td>
<td>5</td>
<td>7.04%</td>
</tr>
<tr>
<td>Kaggle</td>
<td>5</td>
<td>7.04%</td>
</tr>
<tr>
<td>Geniuskitchen</td>
<td>4</td>
<td>5.63%</td>
</tr>
<tr>
<td>Not mentioned</td>
<td>4</td>
<td>5.63%</td>
</tr>
<tr>
<td>Meishiji</td>
<td>4</td>
<td>5.63%</td>
</tr>
<tr>
<td>Go Cooking</td>
<td>3</td>
<td>4.23%</td>
</tr>
<tr>
<td>FoodKG</td>
<td>3</td>
<td>4.23%</td>
</tr>
<tr>
<td>Novel dataset</td>
<td>2</td>
<td>2.82%</td>
</tr>
<tr>
<td>Cooks</td>
<td>2</td>
<td>2.82%</td>
</tr>
<tr>
<td>Epicurious</td>
<td>2</td>
<td>2.82%</td>
</tr>
<tr>
<td>Foodnetwork</td>
<td>2</td>
<td>2.82%</td>
</tr>
</tbody>
</table>

4.3. RQ2 what kind of data and preprocessing techniques are used in FRS?

To generate recommendations, FRS are dependent on data. This data is treated differently based on the method of recommendation. For FRS based on statistical measures, the data is used to calculate similarities. With machine learning prediction, data is used for training the recommender system. With a query-based recommendation, the data is used to build a database. In this case, recommendations are retrieved directly from a database with queries. In order to fully understand the data used for food recommendation, several aspects are considered. These aspects are the source, format, size, features, preprocessing, and finally representation of the data.

4.3.1. Source

A total of 46 unique sources are found in the papers retrieved for this research. Of these, 33 are only used once. The distribution of the sources used more than once is shown in Table 12. Allrecipes is most common, and used in 21 (29.58%) of the FRS. The second most common dataset is Food, used in 11 (15.49%) of the FRS in this study. Yummly and Recipe1M are both used in eight (11.27%) FRS, while USDA and Kaggle are used in five (7.04%). Geniuskitchen and Meishiji are used in four (5.63%) FRS. This is also the amount where the source is not mentioned. Go Cooking and FoodKG are used in three (4.23%) FRS each. Cooks, Epicurious and Foodnetwork are all used twice (2.82%). Finally, two FRS are also based on novel datasets.

The average number of sources used per model is 1.65, with a SD of 1.69. The relative number of sources is calculated for the FRS separately per method. A heatmap, shown in Fig. 9 is generated to provide an overview of the number of sources used for the different methods. A total of 51 (71.8%) FRS use one source. Two sources are used in 12 (16.9%) systems. The remaining eight (11.27%) FRS use more than two sources, with one (1.41%) content-based system using 12 sources.

4.3.2. Size

The average number of items among the datasets is 585 830.20, with an SD of 2.32 million. The largest dataset found consists of 16 million items, while the smallest contains 105. The distribution of dataset sizes is shown in Fig. 10. The size is not mentioned for 13 (18.31%) systems. The size distribution is assessed per class of methods in Table 13. When considering all datasets used for content-based filtering, the average size and SD are 135 039.00 and 216 069.56, respectively. For collaborative systems, the average size and SD are 227 122.20 and 308 306.56. FRS where a graph-based method is used obtain a lower average of 67 008.10 items and an SD of 100 447.20. Finally, the average dataset size for hybrid FRS is 182 719.00 and the SD is 115 461.39.

4.3.3. Format

The format of the data can be divided into four classes based on the selected papers. These classes are text, images, graphs and video. The distribution of data formats was assessed separately for each class of methods and plotted in Fig. 11. Text is used in 59 (83.10%) FRS, while images are used in 23 (32.40%). Both graphs and videos are used in one (1.41%) FRS with a graph-based method each. In 11 (15.5%) FRS, the data format is not mentioned. The distribution of data formats is similar for the different methods, but FRS with graph-based recommendation is the only category where graphs and videos are used.

4.3.4. Attributes

The datasets contain many different attributes that range from ingredients and cooking method to images of dinner recipes. The different features are shown in Fig. 12. A total of 25 unique features are found. Ingredients are used in 58 (81.7%) FRS while cooking instructions and title are used 37 (52.1%) and 30 (42.3%) times, respectively. Images are used in 25 (35.21%) cases, while nutrition is used 20 (28.17%) times. The ratings of recipes are used in 19 (26.76%) FRS. The features contained in the data are not mentioned for 10 (14.1%) systems. Most of the remaining features are only used in specific cases that are highly relevant to the proposed system.

The distribution of the number of attributes considered in the FRS is shown in Table 13. The plot shows that 14 (19.72%) FRS use four attributes, 13 (18.31%) FRS use three attributes and 11 (15.50%) FRS use five attributes.

The distribution of the number of attributes is also considered for the different recommendation methods and plotted in Fig. 14. The distribution is normalised within each recommendation method. The heatmap shows that all FRS based on collaborative filtering mention the attributes that are considered. However, this is not the case for any of the other methods. No FRS employing graph-based methods consider more than seven attributes, while the other methods consider up to nine attributes. In systems with content-based filtering methods, it is most common to consider three attributes. This is the case for eight (21.62%) FRS. Seven (18.92%) systems lack a description of the attributes. For methods where collaborative filtering is used, four and five attributes are most common with three (25%) FRS each. However, two (16.67%) also consider seven attributes. FRS with a graph-based method usually use three and four attributes, with three (20%) and four (26.67%) of the systems, respectively. Of the FRS based on a hybrid method, the smallest amount of attributes considered is 3. All amounts of attributes above this are used in one (14.29%) system each.
4.3.5. Preprocessing

In order to extract the appropriate features, preprocessing is a necessary step. Six different preprocessing techniques are found in the primary studies. These techniques are text cleaning, exclusion, duplicate removal, image resizing, and per-pixel mean subtraction. Text cleaning consists of filtering out unnecessary words and phrases from the data. This is done by removing non-relevant words and merging words with similar semantics. Translation is also considered a part of text cleaning. Exclusion is a technique used to remove recipes based on certain criteria in order to reduce noise from the data. The removal of duplicates and resizing of images is also done during preprocessing. Finally, per-pixel mean subtraction is concerned with normalising all pixels in an image by subtracting the mean value for the specific pixel locations.

The distribution of the different preprocessing techniques is shown in Fig. 15. It is shown that 29 (40.8%) FRS lack a description of how the data is preprocessed. Following this, 24 (33.8%) FRS apply text cleaning. The same amount of systems applies exclusion. Duplicates are removed in four (5.63%) FRS. Three (4.23%) FRS apply image resizing, while one (1.41%) FRS apply per-pixel mean subtraction.

4.3.6. Representation

To make the data available for FRS, it is helpful to make suitable representations. Eight different classes of representation are found in the collected body of research. The distribution of the different classes is shown in Fig. 16. The Figure shows that embedding is used in 27 (38.03%) FRS. In 18 systems, the representation of data is not described. Matrix and TF-IDF scores are used to represent data in 13
Fig. 12. The distribution of attributes the data contains.

Fig. 13. The overall distribution of the number of attributes used in the FRS.

Fig. 14. The number of attributes found in the FRS with the different methods. When the attributes are not mentioned, the model is assigned NA. The distribution is normalised within each method.

Fig. 15. The different preprocessing strategies found in the FRS.

(18.31%) and 11 (15.49%) systems, respectively. Topic modelling and tokenisation are both used to represent data in six (8.45%) systems. Vectorisation is used four (5.63%) times and three (4.23%) systems use raw data. The remaining system (1.41%) represents the data as vertices.

4.4. RQ3 how are the Food Recommendation Systems evaluated?

4.4.1. Evaluation method

The distribution of evaluation methods found in this study is shown in Fig. 17. Offline evaluation is used in 47 (66.20%) FRS. Online evaluation is performed in nine (12.68%) FRS, while hybrid evaluation accounts for eight (11.27%) FRS. Five (7.04%) systems do not mention the evaluation method used, and two (2.82%) recommendation systems are not evaluated.

4.4.2. Evaluation metrics

As online evaluation is based on specific questionnaires and subjective experiences, no evaluation metrics are used to measure the performance. As such, the metrics are only extracted from papers describing offline- and hybrid evaluation. The obtained distribution of evaluation metrics is shown in Fig. 18. Accuracy-based metrics are
used in 39 (54.92%) FRS, while ranking-based metrics are used in 25 (35.21%). Error-based metrics group of metrics is used in eight (11.27%) FRS. Finally, three FRS (4.23%) measure aspects of the FRS that do not fit into the three defined classes and are classified as miscellaneous. In eight (11.27%) FRS, the evaluation metrics are not mentioned.

4.5. RQ4 to what extent is recent research on Food Recommendation Systems reproducible?

With available datasets and codebases, it is evident that other researchers can access and reproduce the process and results of a study with a higher success rate. As the description of datasets is already extracted for Section 4.3.1, we will focus on the availability of the codebase in this section.

It is found that 55 (77.46%) FRS have not published their codebase, and are therefore not open source. The remaining 16 (22.54%) systems have made their codebase available. To examine if there is a trend towards making the codebase available, the fraction of open source systems relative to the year is plotted in Fig. 19. It is shown that no research published before 2019 has available code. In 2019, 30.77% of the primary studies made the code available. 45.45% of the primary studies published the code in 2020. In 2021, 16.67% of the research had available code. Finally, the percentage of primary studies with open code was 36.36% in 2022.

4.6. RQ5 what are the advantages and limitations of current techniques used in FRS?

From our analysis, 63 (94.03%) papers state the advantages of food recommendation and their method. Limitations are only stated in 26 (38.81%) primary studies. However, it is possible to infer limitations based on the proposed method, conclusion or future research directions for 35 (52.24%) primary studies. In total, disadvantages are either stated or inferred from 47 (70.15%) primary studies. Both the stated advantages and the disadvantages are grouped into three main categories. An overview of the advantages is given in Table 14, while the limitations are shown in 15. Furthermore, an explanation of each category, and how it is used in the primary studies is given below.

4.6.1. Advantages

Heterogeneous attributes. FRS makes it possible to consider many more attributes than a user is able to consider individually. The attributes are captured in different formats that can be exploited by FRS when generating recommendations. These multi-modality provides many heterogeneous attributes that can be used to describe food items. These are attributes like visual appearance, nutritional value, flavour, and geographical regions. Among the primary studies stating advantages, 17 (25.37%) describe the use of rich attributes from multi-modal data.
learn spice preference, while Ribeiro et al. (2017) use the activity level of other recipes. Li et al. (2018) use the spiciness of food in order to recommend restaurant dishes to users. The main aim of this study is to achieve this can be found. First of all, it is possible to achieve this by providing explanations for the recommendations. Furthermore, some systems are able to answer questions or have conversational abilities. Eight (11.94%) of the collected primary studies propose systems where a human-like experience is considered.

Three (4.48%) primary studies aim to provide a more human experience by designing FRS that generates explanations for the recommendations. With their multi-list approach, Starke et al. (2021) are able to provide explanations for each list of recipes, e.g. “Similar, but with fewer calories”. Guidotti and Viotti (2020) takes advantage of a memory-based approach in order to provide concrete explanations for the recommendations. Historical data for the users are stored in the system, and explanations are generated based on the earlier purchases, e.g. “Because you purchased X, the following recommendation is made”. In a study, Lei et al. (2021) proposes to use deep learning models in order to generate explanations from images and videos of the recommended food.

Visual attributes are used by Chen, Ngo et al. (2018), Chen, Pang et al. (2017), Gao et al. (2019), Khan et al. (2019), Lei et al. (2021), Meng et al. (2020), Min, Bao et al. (2017), Min, Jiang et al. (2017) and Salvador et al. (2019) in order to capture how users visually judge food. This is used to make personalized recommendations that cater to the visual preferences of the users. In their research, Rostami et al. (2022) propose using time in order to create a time-aware similarity measure. This measure takes the changes in food preferences over time into account, making their FRS able to handle users with dynamic eating habits. Maia and Ferreira (2015) consider contextual information, such as the current location of a user. By doing this, they are able to recommend recipes based on available offers by a close food place. The research by Jin et al. (2020) considers restaurant-related dish quality in order to recommend restaurant dishes to users. One primary study by Gallo et al. (2022) considers the use of water resources when recommending recipes. The main aim of this study is to make recipes less dependent on water as a resource. Ingredients from resource-friendly dishes are tagged and used as substitutions in other recipes. Li et al. (2018) use the spiciness of food in order to learn spice preference, while Ribeiro et al. (2017) use the activity level and anthropometric measures of users to suggest meals based on their dietary needs. Adaji et al. (2018) define personality types to cluster users. These clusters then receive similar recommendations. Finally, in a study by Khan et al. (2021), the significance is calculated for a total of 288 attributes extracted from recipes.

Nutrition-aware. By using FRS, numerous aspects of nutrition can be considered when recommending recipes. These are aspects that can describe both users and food items. For food items, attributes like calories are the most prevalent. Furthermore, the items can be compared to nutritional guidelines. For users, attributes like health- and cultural constraints, dietary requirements and disease are used. 12 (17.91%) primary studies report using nutrition when recommending food items.

By associating recipes compliant with the health- and cultural constraints of a user, Bianchini et al. (2017) is able to improve the nutritional habits of the users using the system. Similarly, Chavan et al. (2021) recommends recipes based on calorie requirements. This is a function of the age, activity level, height and weight of a user. Nirmal et al. (2018) consider nutrients and flavour as the main variables of food with a focus towards identifying alterations for ingredients in order to optimise recipes to become more healthy. In a study by Li et al. (2023), separate representations are learned for user preferences and food healthiness. Increasingly healthier recipes are recommended to nudge users towards healthier habits. This concept of nudging is also described by Pecune, Callebert, and Marsella (2020). By using nutrition guidelines, Ng and Jin (2017) obtain a list of nutritional foods. This is used to recommend recipes to toddlers. This is also the case in a study by Chavan et al. (2021), where guidelines from the American Diabetes Association are used. Song et al. (2023) consider calories when recommending recipes. They base this on the hypothesis that the amount of calories in a recipe influences the preference of an individual. Finally, Starke et al. (2021) propose a system where five lists of recommendations are generated. The lists consider various nutritional aspects and contain either recommended recipes with fewer calories, fewer carbohydrates, less fat and more fibre.

Recommendations can also cater to users suffering from various chronic diseases in order to improve their health. By communicating with hospital servers, Ivașcu et al. (2018) is able to obtain health information on users. This makes it possible to extract knowledge on recommended and restricted ingredients based on the disease to generate recipe recommendations. With a focus on blood pressure, Clunis (2019) divide users into several groups. Based on food nutrients, and their known interactions with prescribed drugs and blood pressure, recipes are recommended. Finally, by identifying different levels of obesity among users, Mckensy-Sambola et al. (2021) presents a system recommending diets to help users lose weight.

### Table 14

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Primary studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneous attributes</td>
<td>Adaji et al. (2018), Chen, Ngo et al. (2018),</td>
</tr>
<tr>
<td></td>
<td>Chen, Pang et al. (2017), Gallo et al. (2022),</td>
</tr>
<tr>
<td></td>
<td>Jin, Zhang et al. (2020), Rokne, Smyth et al. (2019),</td>
</tr>
<tr>
<td></td>
<td>Khan, Smyth et al. (2019), Lei, Hug et al. (2018),</td>
</tr>
<tr>
<td></td>
<td>Suraud et al. (2019), Zhang et al. (2021), Li et al.</td>
</tr>
<tr>
<td></td>
<td>Maia and Ferreira (2018), Meng et al. (2020),</td>
</tr>
<tr>
<td></td>
<td>Min, Jiang et al. (2017), Sang and Mei (2017),</td>
</tr>
<tr>
<td></td>
<td>Ng and Jin (2017), Ribeiro et al. (2017),</td>
</tr>
<tr>
<td></td>
<td>Rostami et al. (2022) and Salvador, Drouzdal et al.</td>
</tr>
<tr>
<td></td>
<td>Giró-i Nieto and Romero (2019)</td>
</tr>
</tbody>
</table>

### Table 15

<table>
<thead>
<tr>
<th>Limitation</th>
<th>Primary studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited data</td>
<td>Bisaj et al. (2018), Panda, Dabas, and Kaur (2018),</td>
</tr>
<tr>
<td></td>
<td>Chen et al. (2021), Chumins et al. (2019),</td>
</tr>
<tr>
<td></td>
<td>Garrido-Merchán et al. (2018), Hausmann et al. (2019)</td>
</tr>
<tr>
<td></td>
<td>Li et al. (2018), Liu, Chen et al. (2019), Lai, Wu,</td>
</tr>
<tr>
<td></td>
<td>Wei et al. (2018), Maheshwari et al. (2019),</td>
</tr>
<tr>
<td></td>
<td>Maia and Ferreira (2018), Min et al. (2017),</td>
</tr>
<tr>
<td></td>
<td>Shchukina et al. (2018), Zhang, Zhou et al. (2019)</td>
</tr>
<tr>
<td></td>
<td>Chen et al. (2017), Pang and Ngo (2018), Gao et al.</td>
</tr>
<tr>
<td></td>
<td>(2019), Guidotti et al. (2020), Ivașcu et al. (2018),</td>
</tr>
<tr>
<td></td>
<td>Loesch et al. (2022), Nadee et al. (2021), Starke et</td>
</tr>
<tr>
<td></td>
<td>al. (2021) and Zhang et al. (2019), Duan et al. (2019)</td>
</tr>
<tr>
<td></td>
<td>Chen et al. (2017), Ngo et al. (2018), Gao et al. (2019),</td>
</tr>
<tr>
<td></td>
<td>Guidotti et al. (2020), Ivașcu et al. (2018),</td>
</tr>
<tr>
<td></td>
<td>Loesch et al. (2022), Nadee et al. (2021), Starke et</td>
</tr>
<tr>
<td></td>
<td>al. (2021) and Zhang et al. (2019), Duan et al. (2019)</td>
</tr>
<tr>
<td></td>
<td>Adaji et al. (2017), Altosaar et al. (2021), Ivașcu</td>
</tr>
<tr>
<td></td>
<td>et al. (2018), Maheshwari et al. (2019), Ribeiro et</td>
</tr>
<tr>
<td></td>
<td>al. (2017) and Twomey et al. (2020)</td>
</tr>
</tbody>
</table>

### Table 16

<table>
<thead>
<tr>
<th>Overview of the limitations present in the primary studies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limitation</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Heterogeneous attributes</td>
</tr>
<tr>
<td>Nutrition-aware</td>
</tr>
<tr>
<td>Human-like</td>
</tr>
</tbody>
</table>
Three (4.48%) primary studies provide a more human-like experience by designing FRS able to answer questions raised by users. By using a food knowledge graph, Chen et al. (2021) is able to conduct knowledge-based question-answering. The main topic of a question is detected, and a subgraph of candidate recipes is extracted. A similarity score is calculated between the query and the potential recipes, and the potential recipes are ranked. Similarly, Haussmann et al. (2019) builds a knowledge graph that can be queried with three question types. One system proposed by Khilji et al. (2021) classifies and labels input questions. This label is used to retrieve recipes, which are recommended to the user in the end.

Two (2.99%) studies explore the use of chatbots to provide a more human-like experience with the recommender system. In the system proposed by Mendes Samagao et al. (2021), food preferences are extracted by intent classification and sentiment analysis. Both ingredients and dishes matching a user’s preferences are then retrieved from the database. For managing dialogue, a set of possible paths the chatbot can take is collected through simulated conversations. The extraction of preferences and the dialogue management are provided by the Rasa framework (Bocklisch, Faulkner, Pawlowski, & Nichol, 2017). Similarly, Pecune et al. (2022) extracts intent from user input with a module for natural language understanding. The intent is given as input to the dialogue management, which keeps track of a user’s preferences in a user frame. Another module generates a response, while the user frame is given to a recommendation engine that returns recipes.

4.6.2. Limitations and open problems

Limited data. Several limitations can be found when it comes to data used for creating FRS. Three categories of data limitations are defined from the collected body of research. First of all, some proposed FRS are based on small datasets. In order to combat this, some researchers use simulated datasets. Finally, region imbalance in the collected data is a recurring theme. This disadvantage is present in 13 (19.40%) of the selected primary studies. Six (8.96%) primary studies use datasets that contain little data. This disadvantage is found in primary studies by Adaji et al. (2018), Bajaj et al. (2018), Maia and Ferreira (2018), Min, Jiang et al. (2017) and Zhang, Zhou et al. (2019). Furthermore, Liu et al. (2018) only consider a small test set for assessing the performance of their FRS.

In order to combat this, several papers describe the use of simulated datasets when creating FRS. In order to recommend recipes, Chen et al. (2021) describes the use of simulated food logs. This is defended by the fact that they do not have access to food logs that matches recipes from their database. For a similar question-answering approach, Haussmann et al. (2019) use a simulated dataset consisting of questions and answers. This data is based on a set of manually designed question templates. In order to test their system, Clunis (2019) create simulated user profiles. Finally, Garrido-Merchán and Albarca-Molina (2018) use expert knowledge to generate synthetic data. This data is used to simulate the cooking process by a fitted machine-learning model. This is further optimised using Bayesian Optimisation to generate recipes that are given as recommendations.

Region imbalance is when a large part of the data comes from one certain region. A system proposed by Shchuka et al. (2020) is aimed at recommending Russian cuisine. This is done by collecting recipes from Russian websites, causing the regional imbalance. In order to design a system for individuals that are prescribed medications for hypertension, Clunis (2019) include the interaction between medicine and food in their system. However, they only focus on a subset of medications prescribed in the USA. Furthermore, Li et al. (2018) recommend Chinese foods and only use evaluators from China. Similarly, Maheshwari and Chourey (2019) only focus on Indian cuisine when creating their system.

Few attributes. Even though an advantage of FRS is the possibility to consider many more attributes than a single user, the number of attributes actually used is often small. Furthermore, the importance of attributes is also not considered in one case. Finally, many papers state a lack of specific attributes that could improve their system as a recommendation for future work. Limitations regarding attributes are found in 9 (13.43%) of the primary studies.

In the research by Ivaçu et al. (2018), an ontology is designed to recommend recipes based on health information. In future work, it is suggested to extend the data to contain more diseases in addition to ingredient quantities. The quantity of ingredients is also not considered by Nadee and Unankard (2021). The recommendations provided by Starke et al. (2021) are only based on the title of the recipes. Similarly, Guidotti and Viotto (2020) only consider the co-occurrence of ingredients for their recommendations. Zhang, Zhao et al. (2019) state that the combination of ingredients, nutrition and the special needs of a user should be incorporated into their system in the future. In research by Chen, Pang et al. (2018), it is mentioned that cutting and cooking attributes and food categories can improve the performance of their system. Gao et al. (2019) only considers one image per recipe. Furthermore, they state that the use of various health attributes can produce healthier recommendations. Somewhat vaguely, Loesch et al. (2022) mention that they want to include more attributes to compare food products to each other. Furthermore, they want to extend their system by including nutrients. In their research, Chen, Ngo et al. (2017) attempt to recommend recipes based on images. They state that there is a possibility to enhance the recommendations by assigning different weights to the attributes predicted from these images.

Limited evaluation. The evaluation is another theme of limitations found in the primary studies. This theme concerns the lack of any evaluation of the recommendations. Furthermore, some papers use few metrics during evaluation. Six (8.96%) of the primary studies are limited by this. The lack of any evaluation is found in the research by Adaji et al. (2018), Ivaçu et al. (2018) and Maheshwari and Chourey (2019). Ribeiro et al. (2017) only evaluate some aspects of their system due to its stage of development. In research by Altosaar et al. (2021) and Twomey et al. (2020), only one evaluation metric is used to assess the performance of their systems.

5. Discussion

5.1. Primary study selection

During the study selection process, the number of collected papers was drastically reduced from 2738 to 67. The exclusion of 97.55% of the studies may indicate that the search string does not fit the purposes of this study and yields papers with low precision. However, the iterative and heuristic approach taken to develop the search string (Table A.1) decreases the likelihood of this. As stated in Section 4.1, the majority of retrieved research is published in conference proceedings. The quality of conference papers is often considered inferior to papers published in journals (Freyne, Coyle, Smyth, & Cunningham, 2010). As such, it is more likely that the findings indicate a low standard of reporting present in this domain. This further supports the findings by Min, Jiang and Jain (2019), stating that the recommendation of food is lagging behind other domains of recommendation.

The yearly publication of papers indicates that there is a decrease in publications from 2021 to 2022, as shown in Fig. 4. However, it is more likely that this is due to the collection date of this SLR. As the collection occurred in October 2022, the whole year has not been considered in this research. With this in mind, it can be argued that the amount of research did not decrease. This supports the findings by Trattner and Elsweller (2019) stating that the amount of research in this domain is steadily increasing.
5.2. Recommendation techniques

There is a high variety of techniques used to recommend food. The most common technique consists of a content-based filtering method with a machine learning algorithm. This technique does not generate personalized recommendations. This is likely due to the fact that content-based filtering can be used to define similarities between items that are not necessarily users, while the other methods either exploit the similarity between users or user–item relationships to create recommendations. Personalization is indeed more present with these methods, as shown in Fig. 7. It is evident that content-based FRS is the only class of methods where personalization is the least common. This is likely the leading cause for the results shown in Table 10, where a slight preference towards not considering personal attributes is indicated. The second most common technique is where a graph-based method is used to produce personalized recommendations based on a machine learning algorithm. However, there is a large difference between this and the most common technique. This may be due to the fact that graph-based methods have been developed rapidly in recent years (Wang et al., 2023). As such, it is useful to evaluate whether the low count in Fig. 4 was due to the algorithm being newly established. To review this, the relative usage of this algorithm is plotted over the years in Fig. 20. From the plot, there is an increase in the relative amount of systems employing this method. However, no FRS was proposed in 2017 and 2020. This may indicate that graph-based methods are gaining increased interest and will become gradually more common throughout the years.

As mentioned in Section 2.1, an FRS based on a hybrid method can overcome shortcomings (Chavan et al., 2021; Rostami et al., 2022) and leverage individual strengths (Sun & Huang, 2019) of recommendation methods based on content-based and collaborative filtering. As such, this category is found to improve the performance of the other methods (Chavan et al., 2021; Khan et al., 2021; Rostami et al., 2022; Twomey et al., 2020). Therefore, it seems surprising that this category of methods is used the least among the four. To investigate the trend of hybrid methods, the relative amount of systems using this method per year is calculated and shown in Fig. 21. The figure shows a general decrease in the relative amount of systems adopting this method. Furthermore, no hybrid systems are proposed in 2022. It might be the case that this method is not popular because the disadvantage of increased complexity exceeds the advantages of combining methods.

5.3. Data

The most common dataset is Allrecipes, followed by Food and Yummly. Furthermore, two novel datasets are proposed in the collected research. Both datasets are extensions of existing ones. Gim, Park, Spranger, Maruyama, and Kang (2021) base their novel dataset on the Reciptor (Li & Zaki, 2020) dataset. This is extended to include a multi-dimensional information vector containing information such as cooking difficulty and cuisine type. In the research Park et al. (2019), Recipe1M (Marín et al., 2021) is utilised and extended with ingredient-pairing scores ranging from −1 to 1. These scores are calculated based on point-wise mutual information and are used to train the deep-learning models. Most of the datasets that are used once are based on other research efforts. Amongst all methods, the average number of sources is 1.65, with a SD of 1.69. The number of sources, shown in Fig. 9, shows that using one source is most common for all methods. One content-based system uses the maximum amount of 12 sources, proposed by Ng and Jin (2017). No hybrid FRS use more than two sources, while FRS based on the other methods exhibit a higher variance in source use. Other than this, no relation is indicated between the method and the number of sources used.

The average amount of items based on all datasets is $5.86 \times 10^5$, with a SD of 2.32 million. As shown in Fig. 10, there is a large difference between the smallest and largest datasets. This is reflected by the SD shown in 13. The largest and smallest datasets are both used in content-based FRS. The largest dataset is used in an approach by Altosaar et al. (2021) and consists of 16 million items. In this study, the data consists of meals obtained from a diet-tracking app. The smallest dataset, consisting of 105 items, is used by McKensy-Sambola et al. (2021) to recommend healthy recipes to users based on their health status. Here, the data is a subset of randomly selected recipes from Allrecipes. The highest variability of sizes is found in FRS based on collaborative filtering. In addition, this group of FRS also obtain the largest average dataset size. The lowest variability of sizes is found in FRS with graph-based methods. This is also where the average size of the datasets is lowest.

The most common format is text, followed by images. Graphs and videos are the least common. Data formatted as graphs is used in one graph-based system by Gao et al. (2022) in order to skip the process of graph formulation. Instead, they directly exploit the graphs contained in their dataset. This is the only graph-based system where the graph is not constructed by the researchers. Videos are used in the FRS proposed by Lei et al. (2021). In their study, feature extraction is performed on the videos in order to generate textual descriptions used to train the recommender system. The ingredients are by far the most common attribute to include in all methods. Many attributes specific to the aim of the FRS are used. For example, Jin et al. (2020) use restaurants in order to focus on the dining out scenario. Guidotti and Viotto (2020)
exploit transactions to learn a user’s eating habits and recommend ingredients based on shopping behaviour.

For most papers, the preprocessing is not mentioned. Among the papers where it is mentioned, the most common strategies are text cleaning and exclusion. Nadee and Unankard (2021) manually merge ingredients with similar meanings as a preprocessing step. Pan, Xu, and Li (2020) also merge manually, but also employ NLP techniques like tokenisation and stemming. Based on the tokens, it is possible to remove any word that is not a noun or a verb from the cooking instructions. For exclusion, a threshold for removing items from the dataset. Gim et al. (2021) use the number of ingredients as a threshold, and excludes recipes with less than four ingredients. Guidotti and Viotto (2020) sets the threshold to six ingredients. Furthermore, they exclude users with less than 10 products in their profile. The threshold is set based on the amount of data that is available to the researcher, and there are no strict rules for this. Duplicate removal is not common, indicating that the number of duplicates present in the most common datasets is low. For recommendation and prediction based on images, image resizing is common as it reduces training times of computer-vision tasks (Saponara & Elhanashi, 2022). Meng et al. (2020), Salvador et al. (2019) and Shchuka et al. (2020) all downsize the images to 224 × 224 pixels before considering them in their FRS. Per-Pixel Mean Subtraction is only done in one paper by Shchuka et al. (2020). With this technique, the mean value of each pixel location is calculated over all the images. This value is then subtracted in order to centre the pixel location to around 0. This inhibits gradient explosion, increasing the efficiency of the training (Shchuka et al., 2020).

Three to five attributes are used by most FRS, and it is possible to deduce some connection to the recommendation method in from 14. While most methods have different frequencies for a different number of attributes, the frequencies of the number of attributes used in hybrid FRS are constant. As described by Bai et al. (2019), a feature of the hybrid methods is that it is possible to base recommendations on multiple sources of information (Section 2.2). Additionally, only seven FRS are based on a hybrid method. The fact that a range from three to nine attributes is captured from the small set of FRS supports that there is a high variance in the number of attributes used. As such, it is not surprising that there is not any preference towards having a certain number of attributes, which is the case for the other methods. These findings support (Bai et al., 2019) by showing that no hybrid methods use less than three attributes and that nine attributes are just as common.

To make the data available for FRS, it is helpful to make suitable representations. It is evident that embedding is the most common representation of data. Similar to vectorisation, embedding replaces words with numerical vectors that are machine-readable. However, embedding can be learned unsupervised and is usually used as a layer in a deep learning model (Almeida & Xexéo, 2019). One advantage of embedding is that it is applicable to multi-modal data. Chen, Ngo et al. (2017) extract features from images which are embedded in order to predict ingredients and cooking method. However, embedding is also used to understand natural language. Mendes Samagalo et al. (2021) use pre-trained embeddings to retrieve ingredients based on queries. Matrix is also a common representation in the primary studies. In these papers, the matrix usually denotes the presence of ingredients in a recipe. Nirmal et al. (2018) models the dataset as a binary matrix, where the columns are different cuisines. The ingredients are rows, and a binary vector denotes in which cuisines the ingredients can be found. Khan et al. (2021) follows a similar approach, using columns for recipes instead of cuisines.

5.4. Evaluation

As online evaluation considers subjective experiences, it is costly and time-consuming. Offline evaluation is faster and less expensive (Ghannadrad et al., 2022). As such, it is not surprising that it is most common to evaluate FRS offline. When looking at the distribution of recommendation methods for the different evaluation methods in Fig. 17, there does not seem to be any relation. However, no hybrid FRS are evaluated online. Furthermore, content-based FRS is the only category where systems are not evaluated. Three FRS use metrics that do not fit into any of the previously defined categories. Bajaj et al. (2018) and Bianchini et al. (2017) measure the response time of the system, while Li, Zaki and Chen (2021) captured the diversity and healthiness of the retrieved recipes based on self-made metrics.

5.5. Availability

A scientific process that is reproducible is widely considered essential for scientific research (Ivie & Thain, 2018). Despite this, more than 70% of 1576 researchers in a survey conducted by Baker (2016) have failed to reproduce the experiments of other researchers. There is an increase in concerns regarding this issue in several research domains (Goodman, Fanelli, & Ioannidis, 2016; Ivie & Thain, 2018) such as drug development (Begley, 2013), psychology (Collaboration, 2015) and software engineering (Lewowski & Madeyski, 2022). As a result of these concerns, an assessment of the research that is reproducible is done for this domain of food recommendation. Without available datasets and codebases, it becomes more difficult to reproduce the process and results of a study. By making the research less reproducible, the validity of results is decreased. As such, the availability of the code-base is used to answer this research question. As such, the availability of the dataset and the codebase plays a crucial part in making research reproducible.

Most datasets are mentioned, as shown in 12. In research by Garrido-Merchán and Albarca-Molina (2018), Li et al. (2018), Mendes Samagalo et al. (2021) and Shchuka et al. (2020), the dataset used is not described. As such, the availability of datasets is adequate. On the other hand, only a minority of researchers publish the code used in their research. However, as shown in 19, there is a general increasing trend in the availability of codebases after 2018. While the majority of FRS are still not available, the results seem to be indicative of increasing availability in this domain.

5.6. Advantages

As mentioned in Section 1, food and diet are complex domains. Therefore, there are many attributes that can be used to describe food items. These attributes come in many different formats. Therefore, it is advantageous to have a system that is able to consider this multi-modal data. Furthermore, the addition of many heterogeneous attributes allows the system to capture more information on food and preferences. As such, the use of multi-modal heterogeneous attributes can improve recommendations and increase the level of personalization in the system. In addition, the performance can potentially be improved if the attributes are weighted differently (Chen, Ngo et al., 2017).

FRS can support nutrition-based recommendations based on both preferences and dietary requirements (Chavan et al., 2021). Several aspects of nutrition can be exploited to cater to patients with chronic diseases. Furthermore, it can be used to infer a user’s health status and recommend food items with similar healthiness. Nutrition can also be exploited to provide healthier recommendations and nudge users towards healthier recipes with the aim of improving a user’s eating habits. The latter becomes increasingly important as a growing proportion of the global population is becoming overweight or obese (Tirosh et al., 2019). In a publication by the World Health Organisation from 2016, it was estimated that 39% of adults were overweight and 13% were obese (Organization et al., 2002). There are many factors that contribute to this (Mehrzad, 2020). Unhealthy eating habits is one of these factors (Tirosh et al., 2019). As such, it is a great advantage to consider the nutritional attributes of food items and users when
providing recommendations. Therefore, FRS can be implemented as a tool to combat this.

Finally, providing conversational abilities to FRS can improve the users’ perception of the system (Pecune et al., 2022). Furthermore, the lack of reasonable explanation often finds people more reluctant to accept the recommendations (Lei et al., 2021). As a result, a more human-like interaction is advantageous as it can increase the persuasiveness of the recommendations.

5.7. Limitations

When creating FRS, small datasets can result in problems with the overall generality of a system. In order to combat this, some primary studies describe the simulation of datasets. However, this adds another bias to the research. As researchers simulate datasets to fit the capabilities of their system, this might lead to some issues when real users are applied. Additionally, some researchers use datasets specific to one region only. This results in a region imbalance, again limiting the overall generality of the system.

Surprisingly, the actual number of attributes used for recommendation is often quite small (13). This may be due in part to many early-stage systems being presented in conferences (see Section 5.1). This is further supported by the fact that most primary studies with a lack of attributes state that including more attributes is a future research direction. Some papers state that the weighting of attributes is not considered. This causes attributes to be considered equally important when this might not be the case. Furthermore, similar attributes may be weighted unequally, even though they hold the same amount of information. This can degrade the performance of a system.

5.8. Threats and limitations of our study

The following parameters limit the scope of this study:

- Date: This SLR covers primary studies published from January 2017 to October 2022.
- Type of literature: Primary studies are collected from peer-reviewed journals, and conference- and workshop proceedings. Only grey literature from arxiv.org is considered.
- Perspective: Primary studies are selected using the inclusion criteria and quality threshold. Primary studies that do not describe the implementation of a food recommender system are excluded as well.

For this SLR, the limitations are mostly related to the search string, selection bias, data extraction and data analysis. The selection of keywords to use in the search string, in addition to the limitations of the search engines, can possibly lead to an incomplete pool of candidate papers. In order to combat this, preliminary searches were conducted to learn technical jargon and important keywords used in this domain. Furthermore, the iterative process of generating the search string shown in A.1 further combats this threat. Furthermore, forward snowballing was conducted using Google Scholar. While there is a risk of missing candidate papers as backward snowballing was not conducted (Badampudi et al., 2015), we believe we were able to obtain an adequate selection of primary studies.

The definition of inclusion criteria is of course subject to the bias of the researchers, and can be a potential threat to the validity of our results. However, the inclusion criteria are defined with this in mind. Furthermore, the inclusion criteria are deeply rooted in established strategies for performing SLRs (Ali & Petersen, 2014; Keele et al., 2007; Wohlin, 2014; Wright et al., 2007), previous experiences from conducting SLRs (Breteron et al., 2007; Staples & Niazi, 2007) and previous SLRs conducted in the field of Software Engineering (Ali et al., 2010; Giray et al., 2023; Gurbuz & Tekinerdogan, 2018; Tummers et al., 2019; Tüzün et al., 2015; Villegas et al., 2018). The three-step strategy in Table 5 was also used in the case of uncertainty.

The validity of data extraction directly affects the results of this study. However, the categories were formed and refined iteratively during the extraction. Whenever an author was unsure about the correct category, the case was recorded and resolved by discussions among the authors. In order to analyse data from RQ1 to RQ4, descriptive statistics was used. We believe the threats regarding this data is relatively small. For RQ5, the extracted advantages were stated. However, the limitations were only stated in a small number of papers. Therefore, limitations were inferred from the research using the conclusion or stated future research directions. The threats to validity is not as small here, but we still think this data covers important aspects to consider when implementing FRS.

6. Future work

Research into FRS has been an active field within the last five years, with the number of publications indicating an increasing trend. However, the quality of the primary studies is generally bad with most research being presented in conference proceedings. As such, many primary studies exclude the description of important aspects of their system. Most FRS are not open source, but the datasets used are rarely excluded from the research paper. The exclusion of important aspects and unavailabe code damages the overall availability of the proposed systems. This indicates that the extent of reproducible research is low in this domain. Based on this, we propose several directions to pave the way for further research:

- Publication of high-quality articles by including all aspects of FRS.
- Increase the availability of code used for the implementation and experimentation of a system.
- Decrease the use of simulated or region-specific datasets.
- Include more attributes of different formats.
- Evaluate all aspects of the system.

This research can be beneficial for future research to understand the current state-of-the-art of FRS. Furthermore, established researchers can use this review to make sure all aspects are considered when implementing FRS. Furthermore, it can be used to retain a good quality of reporting in this domain. In the future, it will be valuable to increase the granularity of the aspects considered in this research. It is possible to have a more in-depth look at the algorithms used, and it can be valuable to not categorise this aspect. However, the field of food recommendation is very diverse and no standardised methodology has been proposed yet.

7. Conclusion

Food Recommender Systems use various recommendation methods, algorithms, datasets, preprocessing techniques and data representations in order to recommend food items to users. Furthermore, different evaluation methods and metrics are used to measure the quality of recommendations and the system as a whole. These systems are beneficial to handle information overload by filtering out non-relevant items. The importance of these systems is expected to increase over the years, as more and more recipe-sharing sites and recipes become available. The use of such systems is advantageous, but several limiting factors can be found for the different implementations. In this systematic literature review, several aspects of food recommendation are exhaustively explored. This provides a thorough description of the state-of-the-art of FRS. In addition, it provides an overview of the techniques, data, evaluation, availability, advantages and disadvantages of FRS.

There is a high variety of methods for recommending food. The widely used method for the last five years is content-based filtering. As discussed, this can lead to the fact that most FRS are non-personalized. However, the use of graph-based methods, which are often personalized, seems to be on the rise. The most common class of algorithms
### Table A.1
The iterative development of the search string.

<table>
<thead>
<tr>
<th>Search string</th>
<th>Yield</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>food AND (recipe OR ingredient) AND (recommend* OR predict*)</td>
<td>2250</td>
<td>Many false positives causing a low precision. Add &quot;system&quot;.</td>
</tr>
<tr>
<td>food AND system AND (recipe OR ingredient) AND (recommend* OR predict*)</td>
<td>341</td>
<td>Many false negatives giving low recall. Add &quot;model&quot;.</td>
</tr>
<tr>
<td>food AND (system OR model) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*)</td>
<td>784</td>
<td>Improved recall. Add &quot;network&quot;.</td>
</tr>
<tr>
<td>food AND (system OR model OR network) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*)</td>
<td>938</td>
<td>Slightly improved recall. Precision is low. Assess keywords to subtract.</td>
</tr>
<tr>
<td>food AND (system OR model OR network) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*) AND safety</td>
<td>111</td>
<td>No relevant paper includes &quot;Safety&quot;. Add &quot;NOT safety&quot;.</td>
</tr>
<tr>
<td>food AND (system OR model OR network) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*) NOT safety AND image</td>
<td>31</td>
<td>Relevant paper includes &quot;Image&quot;. Assess &quot;calculat*&quot;.</td>
</tr>
<tr>
<td>food AND (system OR model OR network) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*) NOT safety AND calculat*</td>
<td>60</td>
<td>No relevant paper includes &quot;calculat*&quot;. Add &quot;NOT calculat*&quot;.</td>
</tr>
<tr>
<td>food AND (system OR model OR network) AND (recipe OR ingredient) AND (recommend* OR predict* OR adapt*) NOT safety NOT calculat*</td>
<td>807</td>
<td>Precision and recall of the yield is adequate.</td>
</tr>
</tbody>
</table>

### Table A.2
Quality criteria and scoring descriptions.

<table>
<thead>
<tr>
<th>Quality criteria</th>
<th>Yes</th>
<th>Partially</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 Is the aim of the research clearly stated?</td>
<td>The aim is described consistently in a straightforward manner.</td>
<td>The aim is described inconsistently.</td>
<td>The aim is not clearly stated.</td>
</tr>
<tr>
<td>Q2 Is the scope of the study defined?</td>
<td>The scope is clearly defined.</td>
<td>The scope is vague but can be inferred.</td>
<td>The scope is not defined, and cannot be inferred.</td>
</tr>
<tr>
<td>Q3 Does the study provide additional value to academia or the industry community?</td>
<td>New contributions to either academia or industry are explicitly stated.</td>
<td>Contributions are not specifically stated but can be inferred.</td>
<td>Contributions cannot be inferred or are not new.</td>
</tr>
<tr>
<td>Q4 Is the context of the study clear?</td>
<td>Several relevant research efforts are described thoroughly and critically.</td>
<td>A smaller amount of relevant papers are described.</td>
<td>Relevant research is not described at all.</td>
</tr>
<tr>
<td>Q5 Is the data-collection process clear?</td>
<td>The data source, collection and preprocessing are described.</td>
<td>Only one aspect of data collection is mentioned.</td>
<td>Source and collection of data are not described.</td>
</tr>
<tr>
<td>Q6 Is the data described?</td>
<td>The data description includes the size, type and features of the data.</td>
<td>Data is described but is lacking some aspects.</td>
<td>Data is mentioned, but not described.</td>
</tr>
<tr>
<td>Q7 Is the system for food recommendation well documented?</td>
<td>The system is well described, and all important aspects are included.</td>
<td>The system is decently described, but some logical steps are missing.</td>
<td>Model, algorithm or method is not reproducible because of missing description.</td>
</tr>
<tr>
<td>Q8 Is the evaluation process well documented?</td>
<td>The evaluation is well described.</td>
<td>The evaluation is decently described without particular detail.</td>
<td>Evaluation is not reproducible because of missing steps and details.</td>
</tr>
<tr>
<td>Q9 Are the findings of the study clearly stated and supported by the reported results?</td>
<td>Findings are presented, discussed and unambiguously supported by the results.</td>
<td>Findings are discussed and reported results can be reasonably interpreted as supporting the findings.</td>
<td>Findings are either not presented or are contradictory to reported results, or no results are reported.</td>
</tr>
<tr>
<td>Q10 Is the conclusion related to the aim of the research?</td>
<td>Strongly related, all aspects of the aim are concluded.</td>
<td>Some relation, but all aspects are not answered.</td>
<td>No relation to the aim or no conclusion is made.</td>
</tr>
<tr>
<td>Q11 Are limitations of the study discussed?</td>
<td>&gt;3 limiting factors (or possible improvements) are described with solutions.</td>
<td>&lt;3 limitations (or possible improvements) are described.</td>
<td>No limitations or possible improvements are described.</td>
</tr>
</tbody>
</table>
to use for generating recommendations is machine learning. The popular combination is non-personalized content-based machine learning recommendation. However, the second most common combination for food recommendation is a personalized graph-based machine learning approach. This seems to be the most current method at the moment. Most systems use one source for their data, which is commonly Allrecipes. The average size of the datasets is 585,830.20. With an SD of 2.32 million, the size varies a lot between the proposed systems. The average sizes and SD vary between the methods as well. The most common format is text, and the most common attribute is ingredients. Most systems use somewhere between three and five attributes, which are usually represented with embeddings. Offline evaluation with accuracy-based metrics is the most common.

CRediT authorship contribution statement

Jon Nicolas Bondevik: Conceptualization, Writing – original draft, Visualization. Kwabena Ebo Bennin: Investigation, Writing – review & editing. Önder Babur: Investigation, Writing – review & editing. Carsten Ersch: Writing – review & editing.

Declaration of competing interest

Potential conflict of interest exists:

No conflict of interest exists.

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Data availability

Data will be made available on request.

Intellectual property

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Appendix

See Tables A.1 and A.2.

References


Shchuka, I., MiRakhov, S., Patrushev, V., Tikhonova, M., & Fenogenova, A. (2020). Dish-id: A neural-based method for ingredient extraction and further recipe suggestion. In 2020 international conference engineering and telecommunication (En&T) (pp. 1-5). IEEE.