Expert Systems With Applications 154 (2020) 113449

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A conversational recommender system for diagnosis using fuzzy rules

P. Cordero [a], M. Enciso [b], D. López [a,*], A. Mora [a]

[a] Dept. of Applied Mathematics, Universidad de Málaga, Andalucía Tech, Málaga, Spain
[b] Dept. of Computer Science, Universidad de Málaga, Andalucía Tech, Málaga, Spain

ARTICLE INFO

Article history:
Received 6 November 2019
Revised 3 March 2020
Accepted 10 April 2020
Available online 14 April 2020

Keywords:
Recommendation
Diagnosis
Fuzzy logic
Critiquing
Formal concept analysis

ABSTRACT

Graded implications in the framework of Fuzzy Formal Concept Analysis are used as the knowledge guiding the recommendations. An automated engine based on fuzzy Simplification Logic is proposed to make the suggestions to the users. Conversational recommender systems have proven to be a good approach in telemedicine, building a dialogue between the user and the recommender based on user preferences provided at each step of the conversation. Here, we propose a conversational recommender system for medical diagnosis using fuzzy logic. Specifically, fuzzy implications in the framework of Formal Concept Analysis are used to store the knowledge about symptoms and diseases and Fuzzy Simplification Logic is selected as an appropriate engine to guide the conversation to a final diagnosis. The recommender system has been used to provide differential diagnosis between schizophrenia and schizoaffective and bipolar disorders. In addition, we have enriched the conversational strategy with two strategies (namely critiquing and elicitation mechanism) for a better understanding of the knowledge-driven conversation, allowing user's feedback in each step of the conversation and improving the performance of the method.

1. Introduction

Recommender systems constitute one of the emerging issues in different areas. In some well known surveys (Bobadilla, Ortega, Hernando, & Gutiérrez, 2013; Lu, Wu, Mao, Wang, & Zhang, 2015) a categorization was presented, remarking that, in most cases, a hybrid approach is used. Two of these categories are usually merged: the collaborative filtering approach –which introduces a cluster-like approach to associate the user with some user community so that the recommendation can be guided by the community's previous recommended items– and the knowledge-based approach –where the previously declared preferences of the user is used to build new recommendations.

A simple and efficiently manageable way for knowledge representation are, in its general sense, the rule-based systems, which requires two issues: the construction of the set of rules and the design of an automated reasoning method to infer new knowledge from these rules. Formal Concept Analysis (FCA), introduced by Ganter and Wille (1999), constitutes a solid mathematical framework to manage information. It provides several methods to extract rules –known as implications– from datasets and introduces a logic to reason and infer new knowledge. FCA provides the two elements needed to be a suitable framework for recommender systems: the construction of clusters –used in the collaborative recommender systems– and the knowledge reasoning capabilities –used in the knowledge-based ones. The first role is played by the so-called concept lattice, a dual cluster of items and attributes, and the second one by the implicational logic.

As Renjith, Sreekumar, and Jathavedan (2020) mentioned, recommender systems are evolving to use intelligent engines strongly based on rules as a way for knowledge representation. Some other works also emphasize the use of different kinds of rules in recommendations: fuzzy rules are used by Borrás, Moreno, and Valls (2014), Vesin, Ivanović, Klašnja-Milićević, and Budimac (2012) use rules expressed in terms of first-order logic for course personalization, while others propose the use of association rules (Cakir & Aras, 2012; Jooa, Bangb, & Parka, 2016; Khanian Najafabadi, Naz'ri Mahrin, Chuprat, & Sarkan, 2017).

Thus, the motivation of this work is the following question: *Can FCA contribute to the research on recommender systems?*. More specifically, can FCA provide some light to the well-known problems in this area (sparsity, cold-start, scalability, overspecialized recommendation, etc)? Our proposal is to build a recommender system following the conversational paradigm (Christakopoulou, Radlinski, & Hofmann, 2016). It works by building a conversation with the user, who interacts with the system by iteratively selecting features. Then, the system provides, in each step, a narrowing of the set of items to be recommended.

* Corresponding author.
E-mail addresses: pcordero@uma.es (P. Cordero), enciso@uma.es (M. Enciso), dominlopez@uma.es (D. López), amora@ctima.uma.es (A. Mora).

https://doi.org/10.1016/j.eswa.2020.113449
0957-4174/

As stated in (Ricci, Rokach, Shapira, & Kantor, 2010), "users may not be fully aware of their preferences until they have interacted to a certain extent with the system and roughly understand the range of alternatives". This is specially true when the user has not all the information beforehand, thus the conversational paradigm appears as a promising alternative to collaborative-filtering and knowledge-based recommenders. This conversational paradigm avoids two of the classical problems –cold-start and data sparsity– whereas scalability aggravates. The so-called *curse of dimensionality* appears in those problems with a high number of features, causing user overwhelming.

In our opinion, FCA is a suitable framework to tackle this problem. We can build the conversation guided by implications and reason with the logic methods. Since evaluation of features is commonly imprecise, vague or graded, we consider Fuzzy Formal Concept Analysis. The knowledge is described by using graded implications, that can be automatically discovered (Belohlavek, 2002) from the fuzzy datasets, providing the background knowledge in a complete and smart way. We will use the so-called *Fuzzy Attribute Simplification Logic*, FASL (Belohlavek, Cordero, Enciso, Mora, & Vychodil, 2016) and its automatic reasoning method for implications in data with grades.

Our proposal includes the design of a suitable knowledge representation, considering both the features of the items and the choices to be recommended as propositions in the rules and the use of the attribute closure operator for fuzzy logic to guide the conversation until a recommendation is reached. In addition, in this work we also discuss and evaluate two strategies (namely critiquing and elicitation mechanism) for a better understanding of the knowledge-driven conversation.

Finally, to show the benefits and the practical relevance of our proposal, we have built a conversational recommender system for medical diagnosis and we have designed some experiments where our system has been confronted with other recommender systems and with other techniques (machine learning methods, such as random forests and eXtreme Gradient Boosting). Moreover, some criteria have been defined to give an objective measure of its promising expectations (session length and accuracy).

As research method, we have traversed the following way:

- Step#1 The first issue was to review the literature and to identify two key points: classical paradigms in recommender systems, their main problems and strategies (Section 2).
- Step#2 We have chosen the elements and methods of the fuzzy formal concept analysis framework suitable for the identified problems. More specifically, we have adapted the structure of the dataset (formal context), we have designed a method to manage the features (user input) and the recommendation (system output) in a integrated way and we have used the fuzzy attribute closure as the core of this method (Section 3).
- Step#3 We have collected some available libraries and algorithms in the recommendation area. We have also studied the structure of the datasets managed by these approaches and selected a dataset collecting some real data. In addition we have also gathered some strategies proposed in the literature to enrich the recommender systems and we have tested whether our method can be improved with them or not (Section 4).
- Step#4 We have explored the results and illustrate how they confirm our initial hypothesis. We have also identified the elements that support such confirmation (Section 5).

Consequently, the paper is organized as follows: the following section is focused on the literature review. Section 3 presents our proposal: the fundamental concepts of fuzzy formal analysis are presented, providing a description of our method. The Results section shows the experiment, introducing the application of the proposed framework to the differential diagnosis of schizophrenia and comparing its execution with some other previous recommender systems and other methods available in the literature. In addition, we present a discussion section to highlight the findings provided in the experiments. The paper ends with the conclusions, impact and further research (Section 6).

2. Literature review

In this section, we review some works in the general area of recommender systems. Some surveys have established a solid categorization of these systems (Bobadilla et al., 2013; Lu et al., 2015): collaborative filtering, content-based, knowledge-based, fuzzy set based, social network-based, trust-based, context awareness-based, and group recommendation approaches.

From all of them, collaborative recommendations (Nilashi, Ibrahim, & Ithnin, 2014; Ricci et al., 2010) have been used extensively and with very good results. In Ikemoto, Asawavetvutt, Kuwabara, and Huang (2019), similarity among users and/or items and clustering techniques are the central points of their collaborative filtering approach to recommend the most relevant items. Zhang, Xie, Li, and Lui (2019) propose a novel algorithm based on feedback with users, indicating whether they are interested in the key-terms. This information allows the optimization of the selection strategy of key-terms through user feedback. Other filtering techniques based on statistical measures are used in Phan, Huynh, and Huynh (2017).

Specifically, in this work, we focus on Conversational Recommender systems (Christakopoulou et al., 2016), which work by interacting with the user and building a conversation that ends in the recommendation. Thus, these systems are considered as an alternative to the most popular approaches: collaborative filtering and content-based recommenders. These two approaches to automatic recommendation present two well-known problems (Guo, 2012): the cold-start problem (difficulty to react when new users or new items appears) and sparsity (low number of ratings for a low number of items). Conversational recommender systems avoid both of them since they are not strongly based on the user preferences. Thus, they can be considered an emerging alternative approach with particular characteristics and new challenges to be addressed.

In these methods, it is a key point to reduce the number of interaction cycles with the users using intelligent techniques. But accuracy also matters. Recently, some authors (Jannach, Shalom, & Konstan, 2019) study how to develop recommender systems with more impact in the users. The offline experimentation and accuracy measures are not the only way to measure the impact, in fact, they ensure that "in conversational recommendation, even more foundational research is needed to understand how humans interact". Moreover, conversational recommenders should face questions based on its ability "to uncover user preference and narrow down recommendation candidates effectively" (Priyogi, 2019).

Recently, the conversational paradigm has been enriched by the so-called critiquing recommender systems (Chen & Pu, 2012a). These systems propose to enrich the users elicitation by giving them the opportunity to provide a dynamic feedback in each step of the conversation, refining their preferences when more options are presented. The current state of maturity of the critique-based recommenders is leading to the development of the first systems with industrial application (Christakopoulou, Beutel, Li, Jain, & Chi, 2018).

Some modern and popular techniques are beginning to be used. Thus, neural networks (Christakopoulou et al., 2018) or deep learning (Wu, Luo, Sanner, & Soh, 2019) offer high-quality personalized items suggestions. Unfortunately, as it is well known, these

techniques work as black boxes without any explanation feedback of the results of the recommendations. They do not allow to build transparent recommender systems. In (Musto, Narducci, Lops, de Gemmis, & Semeraro, 2019), the authors claim that "the recent advances in recommender systems research are facing a sharp dichotomy between the need for effective and precise recommendation techniques and the development of transparent algorithms" and in their approach propose the usage of Linked Open Data for explanation aims. In this line, Tran et al. (2019) affirm that explanations help users "have an insight into recommendation processes, choose better solutions, and increase the acceptance of recommended items".

Regarding the application point of view, the main issues of recommender systems (Lu et al., 2015) have focused on recommendations of movies, music, television programs, books, documents, websites, conferences, touristic scenic spots and learning materials, and involve the areas of e-commerce, e-learning, e-library, e-government and e-business services. Since recommender systems have succeeded in many different areas, it is of great interest to transfer their benefits to healthcare. In this case, one application paradigm is the identification of diseases given the patient's symptoms or tests results, in contrast to offering suitable products or services according to the given profile.

Several studies apply collaborative techniques of recommender systems to medicine, for example Davis, Chawla, Christakis, and Barabási (2010), Folino and Pizzuti (2010), among others.

To practically demonstrate the benefits of our proposal, we have built a recommender system for medical diagnosis. As Wiesner and Pfeifer (2014) propose, health recommender systems can be classified in two categories: systems for health professionals as end-users and systems for patients as end-users. For health professionals, recommender systems focus on the idea of building a clinical guideline for a specific case. Our work belongs to the latter category. Calero Valdez and Ziefle (2019) made an orthogonal classification of the health recommender systems according to the medical issue they deal with: diagnosis, therapy or health behaviour. We focus on the diagnosis recommendation. Thong and Son (2015) show a collaborative filtering recommender and they approach the medical diagnosis by using clustering to identify users with similar profiles and fuzzy and intuitionistic logic for reasoning purposes. As in the cited work, fuzzy logic is a key point to properly capture user's information: *"it helps professionals by providing fuzzy picture clustering and recommendation for possible illnesses, thus improving diagnostic accuracy. "* Another unrelated diagnosis approach is carried out by Lafta, Zhang, Tao, Li, and Tseng (2015). In their work, they use time series analysis to predict short-term risk for heart disease.

For a comprehensive review of healthcare recommendation, see the works of Wiesner and Pfeifer (2014), Hors-Fraile et al. (2018) and Afolabi and Toivanen (2018), presenting a very deep study of the area, including its current challenges. Finally, some interesting approaches in the healthcare conversational area, where the dialogue is verbatim built, are the telegram chatbot for healthcare of Narducci, de Gemmis, Lops, and Semeraro (2018) and the conversational agents surveys of Laranjo et al. (2018) and Montenegro, da Costa, and da Rosa Righi (2019).

Now, we summarized some FCA-based approaches to build recommender systems, which can be considered closely related to our selected framework. FCA allows to efficiently represent user preferences and interests in the dataset, also known as *formal context*. This representation seems to be particularly oriented to the collaborative approach, since FCA is employed to propose recommender systems based on user clustering (Alqadah, Reddy, Hu, & Alqadah, 2015; Aufaure & Le Grand, 2013; Chemmalar Selvi, Lakshmi Priya, & Joseph, 2019). Users in the same community should have similar interest since these communities are based on their common interests. Another approach for collaborative filtering which is based on boolean matrix factorization inside FCA is proposed by Nenova, Ignatov, and Konstantinov (2013). The authors use the rating matrix to learn how to compute recommendations for users. They use the information automatically inferred from the dataset and organized in a dual lattice (of users and items) named *concept lattice*. FCA provides an equivalent representation of the concept lattice by means of the so-called implications. As far as we know, implications have not been used to build a recommender in the framework of FCA. A closer work was presented by Ignatov and Kuznetsov (2008), where the authors provide recommendations for Internet advertisement based on FCA. However, they use association rules whereas we use fuzzy implications.

Several works support the application of fuzzy FCA for recommendation. Thus, Castellanos, De Luca, Garcia-Serrano, and Cigaran Recuero (2015) mentioned "the suitability of FCA for context-aware recommendation, outperforming other state-of-the-art proposals". Mezni and Abdeljaoued (2018) propose an explicit description of the objects of the cloud system environment (users, services, ratings), which makes the recommendations more suitable for the targeted user using the fuzzy formal concepts in the built concept lattice. Medina, Pakhomova, and Ramírez-Poussa (2017) introduce a mechanism based on fuzzy FCA developing social network analysis.

This proposal considers the work of Benito-Picazo, Enciso, Rossi, and Guevara (2018) as a previous starting point. In this work we use the Simplification closure operator for implications on fuzzy formal contexts to find the recommendation, an extension of the framework used in that work. We remark that the core of the method takes a linear time since the Simplification closure (Mora, Cordero, Enciso, Fortes, & Aguilera, 2012) outputs the new set of attributes and, with the same cost, a new set of implications corresponding to the complementary knowledge. This new set can be efficiently used in the next step of the conversation without the extra data mining step of re-inferring these implications from scratch. These solid characteristics have been preserved in the extended approach we present in this paper.

Our approach follows the motivation of Anelli and Noia (2019), emphasizing that their "aim is to go beyond the traditional accuracy goal and to start a new generation of algorithms and approaches which exploit the knowledge encoded in ontological and logic-based knowledge bases". Moreover, Priyogi (2019) propose that conversational recommenders should have the following strategies: (1) a set of answer suggestions can assist users in eliciting their preference; (2) feedback received, the next crucial phase is to utilize it for improving recommendation quality; (3); efficiency means to minimize interaction length. The introduction of feedback in conversational recommendation provides some benefits and it has been well studied. Here we consider the inspiring work of Reilly, McCarthy, McGinty, and Smyth (2005) about the so-called *Incremental Critiquing*. They propose a paradigm which includes a recommend-review-revise strategy. In this paradigm, the cycle of the recommendation has two stages: in the first one, the system builds a new milestone of the conversation and then, in the second one, it offers the user an opportunity to provide feedback on the information produced. Our intention is to evaluate if the benefits provided by the critiquing stage are really supported by the experiments. In addition to the reduction in the conversational sessions lengths, Narducci et al. (2018) also emphasize that critiquing strategies improve the recommendation accuracy as well. They argue that the feedback allows to build an effective user-recommender interaction. We also test the validity of this hypothesis in this paper.

3. Method and implementation

As it has been mentioned in the introduction, we present a recommender system following the conversational paradigm. Our method is data-driven and it is strongly based on the expertise stored in the dataset. First, we don't consider rules as a repository of the human expertise but as a collection of the semantics of the system, directly mined from the data. We use rules for knowledge representation but with some significant differences from the classical ruled-based expert systems. In those systems rules follow the causality paradigm; i.e., their interpretation is "if premise occurs, then choose its conclusion", where usually premises and conclusions belong to two different sets of propositions. In logic programming, rules are statements following a given normal form defined to be efficiently executed by some specific automated reasoning method. Rules in this area play the role of a link in the deduction chain to provide an output whenever a proposition is introduced as input. In FCA, implications are just a declarative relation among two subset of attributes, variables or features. They are highly flexible from the syntax point of view and they are not tied to a specific reasoning method, as other frameworks do. Such a flexible orientation opens the door to multiple applications, but it also requires to fix two issues: the allocation of the information in a dataset properly establishing the objects (rows) and the attributes (columns) and the definition of a reasoning method to infer new knowledge.

This work is based on the fuzzy variant of Formal Concept Analysis (FCA) (Belohlavek, 2002), which formalize the dataset as a fuzzy/graded relation between objects and attributes. One of the two ways of representing the knowledge are rules, named implications, that can be automatically obtained from the dataset. In the fuzzy version, attribute implications are formulas $A \Rightarrow B$ where A and B are fuzzy sets over an attribute set M and, informally, an implication such as $\{a, {}^{0.5}/b\} \Rightarrow \{{}^{0.9}/c\}$ means that every object that has attribute a to degree 1 (i.e. fully possesses a), and attribute b to degree 0.5, has attribute c to degree at least 0.9.

A basis of implications associated to a dataset is a set of implications (minimum according to some criteria) that allows to derive in some way all the implications that are satisfied in the dataset. We must therefore distinguish the computation of a basis from the techniques to deduce new implications (automatic reasoning) from the basis. For the first one, the recommender system presented in this paper uses, as a source of knowledge, a basis of fuzzy implications extracted from the dataset by using the NextClosure for Graded Attributes algorithm (Belohlavek, 2002).

This basis is used as background knowledge to guide the conversation towards some user recommendation. More precisely, the knowledge retrieved from the dataset is shaped like a set (basis) of graded implications over which we will reason by using the *Fuzzy Attribute Simplification Logic* (FASL) (Belohlavek et al., 2016). The automated method based on this logic allows us to reach a recommendation.

Now, we briefly present FASL. Truthfulness structures in FASL are tuples $\langle L, \vee, \wedge, \otimes, \rightarrow, \setminus, {}^*, 0, 1\rangle$ where $\langle L, \vee, \wedge, \otimes, \rightarrow, 0, 1\rangle$ is a complete residuated lattice, * is a hedge (a "very true" function (Belohlavek & Vychodil, 2006)) and $\setminus$ is a binary operation satisfying the following adjointness property: $a \setminus b \leq c$ if and only if $a \leq b \vee c$ for all $a, b, c \in L$. As a consequence, $a \setminus b = \bigwedge\{c \in L \mid a \leq b \vee c\}$. These operations are pointwise extended to fuzzy sets.

In particular, in this work we use the a discretization of the unit interval $L = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ and we consider $\otimes$ as a left-continuous t-norm (e.g. the Łukasiewicz or the Gödel t-norm), $\rightarrow$ as its residuated implication, * as the identity mapping and

$$a \setminus b = \begin{cases} a, & \text{if } a > b, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Henceforth, for simplicity's sake, we will describe FASL using this particular framework.

A dataset (a fuzzy formal context in the FCA terminology) is a tuple $\mathbb{K} = \langle G, M, I\rangle$ where G and M are sets of objects and attributes respectively, and $I \in L^{G \times M}$ is the incidence relation that is a fuzzy/graded relation between objects and attributes. Given a dataset $\mathbb{K}$ and two fuzzy subsets of attributes $A, B \in L^M$, we say that $A \Rightarrow B$ is (fully) true in $\mathbb{K}$, denoted as $\mathbb{K} \models A \Rightarrow B$, when the following property holds for all $x \in G$:

$$\bigwedge_{y \in M} (A(y) \rightarrow I(x, y)) \leqslant \bigwedge_{y \in M} (B(y) \rightarrow I(x, y)) \tag{2}$$

In words, the degree to which any object x has (all the attributes from) B is at least as high as the degree to which x has (all the attributes from) A.

The axiomatic system in FASL is defined as follows: for all $A, B, C, D \in L^M$ and $c \in L$,

[Ax]	infer $A \cup B \Rightarrow A$	(*Axiom*)
[Mul]	from $A \Rightarrow B$ infer $c \otimes A \Rightarrow c \otimes B$	(*Multiplication*)
[Sim]	from $A \Rightarrow B$ and $C \Rightarrow D$ infer $A \cup (C \setminus B) \Rightarrow D$	(*Simplification*)

In [Mul], we use $c \otimes A$ to denote the so-called c-multiple of $A \in L^M$ which is a fuzzy set such that $(c \otimes A)(x) = c \otimes A(x)$ for all $x \in M$ (i.e., the degrees to which $x \in M$ belongs to A is multiplied by a constant degree $c \in L$).

As usual, a formula $A \Rightarrow B$ is said to be *provable* from a basis of implications Σ, denoted by $\Sigma \vdash A \Rightarrow B$, if there is a sequence of implications $\varphi_1, \ldots, \varphi_n$ called a *proof* such that φ_n is $A \Rightarrow B$, and for each φ_i we either have $\varphi_i \in \Sigma$ or φ_i is inferred (in one step) from some of the preceding formulas using [Ax], [Mul], or [Sim]. Bases Σ_1 and Σ_2 are called *equivalent*, denoted $\Sigma_1 \equiv \Sigma_2$, if we have $\Sigma_1 \vdash \varphi$ iff $\Sigma_2 \vdash \varphi$, for all implication φ.

The soundness and completeness are ensured when we assume that M is finite. In addition, inference rules in FASL provide equivalences allowing simplification of sets of implications: for any $A, B, C, D \in L^M$,

(**DeEq**) $\{A \Rightarrow B\} \equiv \{A \Rightarrow B \setminus A\}$;
(**UnEq**) $\{A \Rightarrow B, A \Rightarrow C\} \equiv \{A \Rightarrow B \cup C\}$;
(**SiEq**) If $A \subseteq C$ then $\{A \Rightarrow B, C \Rightarrow D\} \equiv \{A \Rightarrow B, A \cup (C \setminus B) \Rightarrow D \setminus B\}$.

The following notions are crucial to the results presented in this paper.

Definition 3.1. Given a dataset $\mathbb{K}$, a set of implications Σ is said to be a *basis for* $\mathbb{K}$ if, for all implication $A \Rightarrow B$, we have $\mathbb{K} \models A \Rightarrow B$ iff $\Sigma \vdash A \Rightarrow B$.

Given a basis of implications Σ and a fuzzy set of attributes $A \in L^M$, the *closure of* A (*with respect to* Σ), denoted by A^+, is defined as the greatest fuzzy set in M such that $\Sigma \vdash A \Rightarrow A^+$. A is called Σ-*closed* if $A^+ = A$.

Note that since both $\mathbf{L}$ and M are finite, the closure A^+ exists. Namely, for all B_i such that $\Sigma \vdash A \Rightarrow B_i$ ($i \in I$), we get $\Sigma \vdash A \Rightarrow \bigcup_{i \in I} B_i$ by a repeated application of (**UnEq**). Closures in sense of Definition 3.1 can be used to characterize provability:

Theorem 3.1. *If Σ is a basis of implications and $A, B \in L^M$, then* $\Sigma \vdash A \Rightarrow B$ iff $B \subseteq A^+$.

In Belohlavek et al. (2016), based on these results, we proposed a new automatic reasoning method for fuzzy attribute implications that may be used to solve the classic-style problems of computing a closure and deciding entailment as well as a conceptually new problem of computing degrees of entailment. The method utilizes the above equivalences and replaces formulas by equivalent but simpler ones, overcoming the drawbacks of other potentially applicable rules. As demonstrated there by the experimental evalua-

tion, the methods are feasible from the computer point of view to almost the same extent as the classical methods.

In the following, for illustration purposes, we describe our recommender system in the field of medical diagnosis. First, we will take a dataset (fuzzy formal context) where patients are the objects and attributes can be symptoms (elements introduced in the dialog) or diseases (the items to be recommended). The original information can be extremely personalized since we use a multi-valued approach and a grade can be assigned to each symptom for each patient.

In summary, our method works as follows: in each step of the conversation, the user interacts with the system providing new symptoms and the algorithm iteratively applies the fuzzy closure operator to enrich the set of symptoms until a disease-column is included in the closed set of attributes, successfully ending the conversation providing a diagnosis as the recommendation.

The application of the closure operator provides a limited set of symptoms, strongly related with the conversation in its current stage, and also narrow the search space guiding in this way the next steps in the conversation.

We have introduced a feedback in the conversation to test if the so-called critiquing paradigm provides some benefits in terms of efficiency or accuracy of our recommender system. This feedback constitutes a depuration of the elicitation provided by the user. In an intermediate stage, the new attributes appearing in the closure are presented to the user in case he considers to increase the graded inferred for the symptoms.

The conversational process is described in the work flow showed in Fig. 1 and it can be briefly described with the following steps:

1. The system asks the user to provide a symptom and a degree associated with it: $(d_x|x)$ where $x \in M$ and $d_x \in L$.
2. It computes the closure $(d_x \mid x)^+$ and its associated reduced set of implications Σ.
3. If the closure contains an attribute identifying a disease, then a diagnosis has been produced. The system stops the process and provides the disease as the recommendation.
4. Otherwise, the recommender asks for a feedback to the user. The symptoms included in the set $(d_x \mid x)^+ - \{(d_x \mid x)\}$ are showed to the user, giving the opportunity to improve their grades. If the user wants to upgrade some of them, a new cycle of the dialogue begins, going to Step 2.
5. If the user declines to provide a feedback, agreeing with the information provided, then new symptoms have to be introduced to continue with the conversation, going to Step 1.

4. Results

In this section, we show the application of the proposed framework, building a recommendation system for the differential diagnosis of schizophrenia with real-world data. First, we present the dataset used in the experiments, then, we define the metrics used to measure the performance of our method. Finally, we describe the range of experiments performed and the obtained results.

4.1. The dataset

In the recent years, an increasing number of initiatives have appeared to share, curate, and study certain prevalent brain pathologies. Among these pathologies, schizophrenia is of the highest interest, and public, curated repositories, such as SchizConnect (Wang et al., 2016), have been released.

SchizConnect is a virtual data repository, integrating and mediating data from other schizophrenia-related databases, such as COBRE (Aine et al., 2017), which collect neuroimaging, psychological, neurological and clinical information. SchizConnect allows to retrieve data about the patients that fulfill some conditions introduced as a query to the database. Using this interface, a subset of the COBRE dataset has been retrieved, by querying SchizConnect for 105 patients with neurological and clinical symptoms. We also collected their corresponding diagnosis.

Among the clinical attributes in the dataset, one can find:

- *Calgary Depression Scale for Schizophrenia* (Addington, Addington, & Schissel, 1990), 9 items (attributes) assessing the level of depression in schizophrenia, differentiating between positive and negative aspects of the disease.
- The *Positive and Negative Syndrome Scale* (Kay, Fiszbein, & Opler, 1987), a set of 29 attributes measuring different aspects and symptoms in schizophrenia.
- The *Simpson-Angus Scale* (Simpson & Angus, 1970), 6 items to evaluate Parkinsonism-like alterations, related to schizophrenia, in an individual.
- The *Structured Clinical Interview for DSM-III-R Personality Disorders* (First, Spitzer, Gibbon, & Williams, 1997), with 9 variables related to the presence of signs affecting personality.
- The diagnosis for each individual: it can be *schizophrenia strict* or *other diagnosis* (which includes schizoaffective and bipolar disorders). These diagnoses are mutually exclusive, thus only one of them is assigned to each patient.

In summary, the dataset consists in the previous 53 attributes related to signs or symptoms, and 2 attributes related to diagnosis. This makes a dataset with 105 objects (patients) and 55 attributes to explore. The symptom attributes are multi-valued: for a given attribute (symptom), the available grades range from *absent* to *extreme*, with *minimal, mild, moderate, moderate severe* and *severe* in between. Thus, all attributes can be considered fuzzy and graded.

4.2. Performance metrics

Experiments in this section correspond to a two-fold purpose: ensure the validity of the recommendations generated by our proposal, comparing with other methods, and present various strategies to optimize the conversational process.

On the one hand, it is necessary to be able to compare the potential of the proposed conversational system as a mechanism to generate appropriate recommendations.

In this sense, the problem of generating a recommendation on a dataset like the one used in this work is similar to that of the prediction of the value of the class variable (*diagnosis*) in a classification problem.

Thus, the classical performance metrics (based on the contigency table) related to classification problems are suitable for the comparison of our proposal to other methods:

- Accuracy: fraction of instances correctly classified.
- Sensitivity: true positive rate (or 1 minus the false positive rate).
- Specificity: true negative rate.
- Precision: also called *positive predictive value*, is the fraction of positive instances among the retrieved instances, that is, the fraction of true positive cases retrieved by the system with respect to the total amount of positive cases.

In order to compute these quantities, we have considered as *positive* class the *strict schizophrenia* diagnosis whereas *negative* class means schizoaffective diagnosis. All these measures are bounded between 0 and 1. Values closer to 1 indicate a better performance of a method.

These measures can be used to compare our proposal to other recommender systems and other machine learning methods focused on classification. We intend to demonstrate the performance

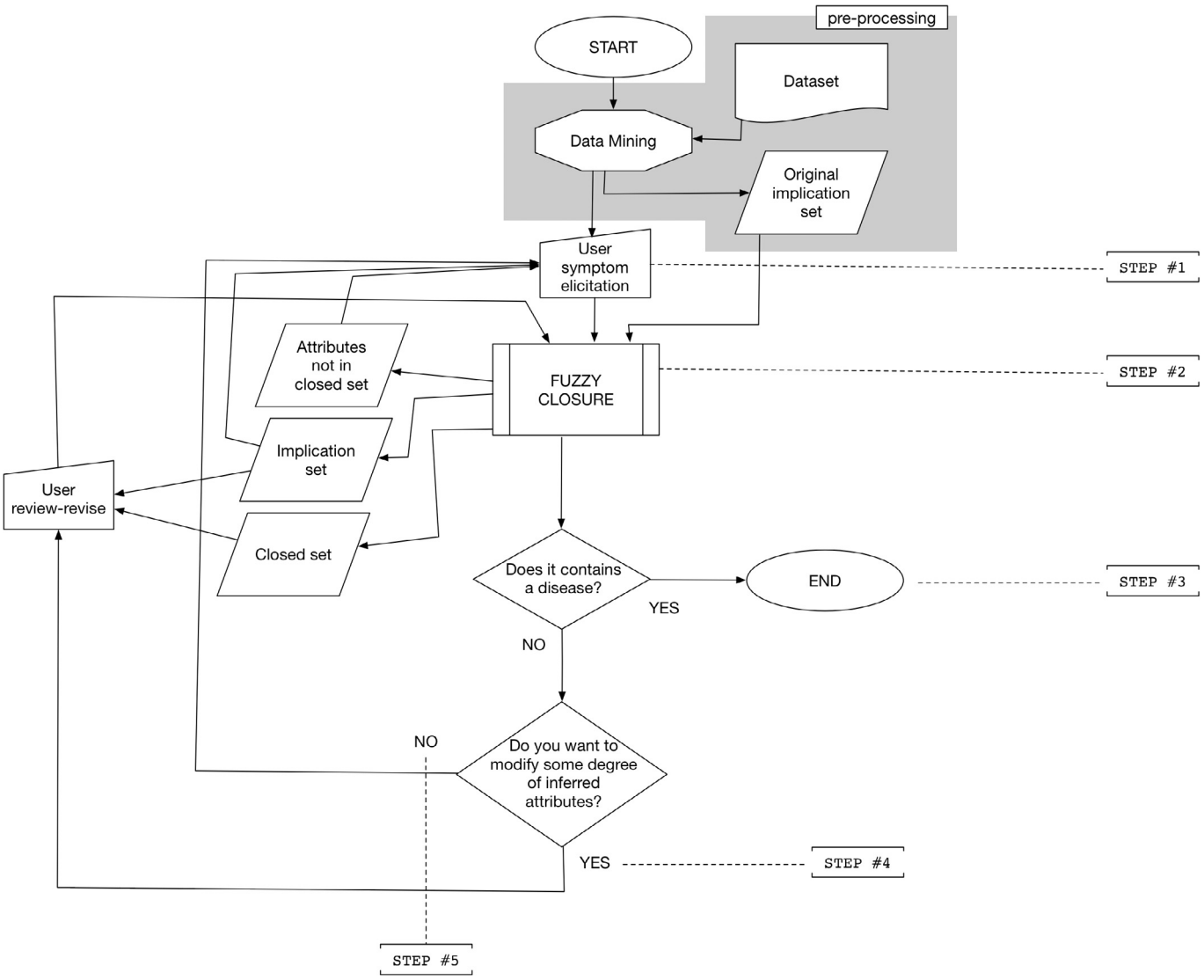

Fig. 1. Workflow of the conversational critiquing recommender.

of our proposal in terms of correct recommendations, using these measures as a basis.

On the other hand, we intend to understand how the conversational process can be optimized, by comparing various strategies, both to generate critiques and to elicit attributes in each step of the dialogue.

In order to compare various strategies in the generation of conversations in our proposal, it is necessary to use specific metrics for conversational systems.

One of the most basic tests to measure the interaction of a recommender system with the user is the evaluation of the *length of the dialogue* (Benito-Picazo et al., 2018; McSherry, 2001). This can be identified with the number of cycles in the conversation needed to obtain a diagnosis. Its usefulness comes from the fact that it outlines the interaction flow between the user and the system.

As we describe in the previous section, the conversation ends when the closure includes a disease, providing a diagnosis. The number of elicitations (Step 1 in Section 3) that the user has provided during the dialogue up to the recommendation is named the *number of steps* (N).

Apart from the previous measure, to study the reduction of the complexity of the problem as the conversation develops, two complementary measures have been used:

- The reduction in the number of rules available after each step. At step i, the proposed system builds the closure (the logical consequent with respect at the implication set in that step) of the set of attributes elicited up to that moment (including critiqued attributes, if necessary). Since the execution times of computing the closure depends on the number of implications, a greater reduction of implications implies faster convergence of the conversation and also a more refined exploration of the attribute space.
- The reduction in the number of attributes to explore. After i steps in the conversation, there is no need to re-elicit any attribute already elicited or critiqued, thus only a subset of the attributes is actually explored by the system at the next elicitation. The lower the number of attributes to explore, the faster the convergence, implying also a reduction of the search space.

These two measures are related each other. In the case of binary attributes, the set of attributes that can be explored after step i is actually given by the set of attributes present in the left hand side of the implications that remain after applying the simplification logic at that step. When using fuzzy attributes, although there is not a one-to-one correspondence between remaining attributes

and attributes in the left hand side of the implications, in practice we can see that using the fuzzy simplification logic induces similar reduction in the number of attributes to explore.

4.3. Experiments

As we described in the work flow of the method (see Fig. 3), the method requires a data preparation stage, where implications are extracted from the context (dataset). For this task, the well-known NextClosure algorithm for graded attributes has been used to retrieve the Duquenne-Guigues basis of implications, extended to use fuzzy graded attributes (Belohlavek, 2002).

In the construction of the basis, the algorithm generates some implications with zero support (that is, such that the left hand side of the implication, the premise, does not occur in the dataset). One feature of such implications is that they contain all attributes in the given context, so their interest is purely theoretical, but in practice they do not provide useful information. Thus, once obtained the implication set, all implications with zero support are removed. This two tasks complete the data preparation stage. We remark that this stage has to be executed just once for each dataset.

In our experiment, the original Duquenne-Guigues basis consists of 20,663 implications. After removing those ones with zero support, only 15,700 are actually considered in the simulations.

In order to carry out the experiments, a dataset consisting of 1000 observations generated following the same statistical distribution of the starting data, described above, has been constructed. They are observations not included in the original dataset, but following their same statistical distribution.

For each class or possible recommendation (in the example, each possible diagnosis), the joint statistical distribution of all attributes grades is determined, and new data is then generated following said distribution. This ensures that the statistical patterns present in the original dataset are taken into account, and that the n-dimensional attribute space is well-represented.

Next, we describe the experiments to compare the proposed method with other methods and, later, some experiments to understand the behavior of the method in terms of the best critiquing and elicitation strategy.

4.3.1. Comparison to other systems

The first set of experiments is intended to compare this proposal with other recommendation systems and other classification methods based on Machine Learning. For this, we will use as measures of the performance of each method those mentioned above in relation to accuracy, precision, sensitivity and specificity when the task is to determine the correct recommendation (diagnosis) for a given input (a new subject from the validation dataset).

Below, we present a list of other recommendation systems with which we have compared:

- **User-based collaborative filtering** (UBCF), the traditional CF method, which may suffer from serious problem in scalability, and **item-based collaborative filtering** (IBCF), which is proposed to build offline an item-item similarity matrix for rating prediction (Adomavicius & Tuzhilin, 2005). For each of these 2 collaborative filtering methods, we have considered tow modalities, depending on the similarity function employed: *cosine distance* and *Pearson's correlation*.
- **Alternated least squares** (ALS) (Zhou, Wilkinson, Schreiber, & Pan, 2008), a recommender for explicit ratings based on latent factors, calculated by alternating least squares algorithm.
- **LIBMF** (Chin et al., 2016), an open source initiative to approximate the incomplete rating matrix using the product of two matrices in a latent space, computing the factorization in parallel.
- **Singular value decomposition** (SVD) and Funk SVD, recommenders based on SVD approximation of the ratings matrix with column-mean imputation.

Table 1
Comparison of the current proposal to other recommender systems and machine learning methods.

	Accuracy	Sensitivity	Specificity	Precision
ALS	0.360	0.333	0.380	0.290
IBCF (Cosine)	0.555	0.475	0.615	0.483
IBCF (Pearson)	0.770	0.466	1.000	1.000
LIBMF	0.491	0.901	0.181	0.455
SVD	0.376	0.515	0.271	0.349
SVDF	0.431	1.000	0.000	0.431
UBCF (Cosine)	0.608	0.967	0.335	0.524
UBCF (Pearson)	0.525	0.783	0.330	0.470
C5.0	0.674	0.636	1.000	1.000
PART	0.883	0.847	0.950	0.970
JRip	0.752	0.814	0.688	0.731
Random Forest	0.953	0.924	1.000	1.000
xgboost	0.818	0.963	0.713	0.706
k-nn	0.589	0.603	0.544	0.815
Proposal	0.982	0.996	0.948	0.955

The aim of these recommender systems is to provide an estimation of the rating of the *diagnosis* attributes. Then, the class assigned to each individual is given by the diagnosis attribute with maximal rating.

Among Machine Learning systems, we have used:

- ***k*-nearest neighbours** (Altman, 1992), a well-known method which compares a new instance with the whole training dataset and classifies it according to the classes of the k nearest training instances in the n-dimensional attribute space.
- Decision trees and rule induction: **C5.0** (Kuhn & Johnson, 2013; Quinlan, 1993), **PART** (Frank & Witten, 1998), and **JRip** (repeated incremental pruning to produce error reduction) by Cohen (1995).
- Classical **random forests** (Breiman, 2001; Wright & Ziegler, 2017).
- **eXtreme Gradient Boosting** (xgboost) (Chen & Guestrin, 2016): an implementation of gradient boosted decision trees designed for speed and performance, which is currently one of the most used methods due to its consistent high performance.

These methods provide directly the classification needed as a recommendation.

The original dataset is used as training set in all the methods. The validation dataset is then employed to get recommendations and compute the classification metrics. The results of the comparison are shown in Table 1. In that Table, it can be seen that, for the problem at hand, Machine Learning algorithms perform consistently better than recommender systems in many cases. This is specially certain for algorithms such as PART, random forests and extreme gradient boosting, that can be considered as the best-performing in that category, achieving more than 80% accuracy in the problem, as well as having the other metrics high as well.

With respect to our proposal, it achieves the top accuracy in the table, with the random forest very close. Comparing the metrics, it can be seen that our method provides recommendations as good, at least, as the best machine learning method.

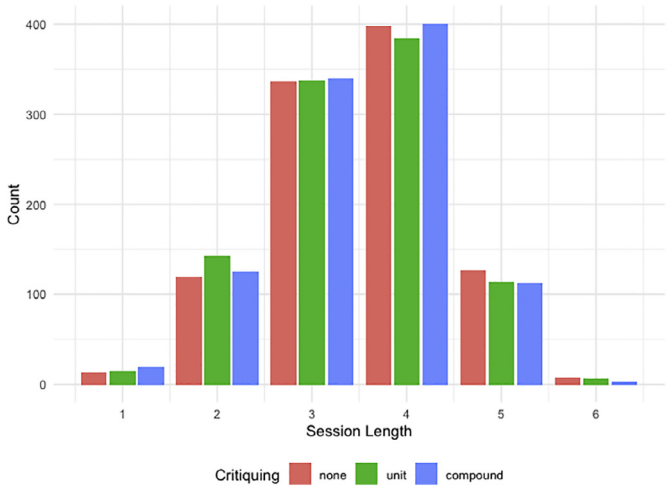

Fig. 2. Proportion of experiments that finished the dialogue after the given number of steps, for each type of critique.

4.3.2. Effect of the critiquing phase

In this section, our aim is to understand the best strategy for the conversational system, with respect to the use of the critique phase.

Particularly, we focus on studying the effect of the critiquing strategy. To this end, we have compared the proposed logic-based conversational system with and without critique. According to Chen and Pu (2012b), there are two main critiquing strategies:

- **Unit critiquing**: quantity- or quality-based feedback for a single attribute.
- **Compound critiquing**: feedback on multiple attributes at once.

Note that critiquing in our proposal is defined as the modification by the user of the degree of one or many attributes of those found by the system by applying the closure operator of the simplification logic.

The first thing to note is that, in our experiments, critiquing has no effect on the classification accuracy of the method. In fact, accuracy, sensitivity, specificity and precision remain the same with all critiquing strategies.

Next, we focus on the effect of these strategies on the metrics that quantify the course of a conversation: session length, the reduction in the number of implications and the reduction of attributes to explore after each step of the conversation.

In order to test the effect of critiquing in the conversation metrics, each validation instance has been processed 50 times by our method and the obtained metrics have been averaged.

As a result, we have found that the session length with unit critiquing is slightly lower than that without critiquing (3.498 vs 3.42 steps, averaged) but this difference is statistically significant ($p < .5 \cdot 10^{-4}$). This means that *unit critiquing* saves conversation cycles, but the dynamics induced by the simplification logic are already so optimized that the advantage of using critiquing is not so evident. Furthermore, in our experiments, there are no statistically significant differences between no critiquing and compound critiquing (3.45 conversation steps), although the latter has a lower average session length. In Fig. 2, we show a bar plot of the session length in our experiments.

The other two metrics, number of implications in each step and number of attributes to explore in each step, confirm this result and its interpretation. There are no significant differences in the reduction of both the number of implications and the number of attributes to explore. Fig. 3 shows the evolution of the average number of remaining implications and attributes in sessions with different types of critiquing strategies.

Both the number of remaining implications and the number of remaining attributes, as mentioned earlier, are intrinsically related, as can be deduced from Fig. 3, where lines corresponding to different critiquing strategies overlap and can not be distinguished, presenting the same average behavior in the course of a conversation, even for different critiquing options. This confirms the idea that the underlying simplification logic is able to optimize the search and reduce the attribute exploration needed to arrive at a recommendation. More details about this result will be given in Section 5, below.

4.3.3. Elicitation strategies

In addition to studying the effect of the critiquing phase in the proposed system, we propose to compare different elicitation mechanisms that may reflect an user's behavior.

In particular, several elicitation strategies have been defined in the present framework:

- **Random**: the user elicits a random attribute from those available at the current step. Since we are working with fuzzy graded attributes, the selected attribute must have positive degree.
- ***z*-score**: the user elicits the attribute whose degree deviates most from the mean of the attributes' values in the training dataset. That is, if x_i is the degree of attribute i, and μ_i and σ_i are, respectively, the mean and standard deviation of attribute i in the original dataset, then the user selects the attribute which maximizes $|\frac{x_i-\mu_i}{\sigma_i}|$.
- **Maximum degree**: The user selects the attribute with maximum degree.
- **Variable importance**: After performing a logistic regression in the original dataset, the absolute value of the t-statistic for each model parameter is used as variable importance (Siegel, 2016). The user elicits the attribute with higher importance.
- **Logistic coefficients**: The coefficients of the logistic regression model are used to estimate the user's elicitation preference.

Random elicitation supposes neither knowledge about the recommendation problem nor about the input to the system. This strategy, along with no critiquing, will be the baseline with which to compare the other ones.

Both z-score and maximum-degree strategies use information about the input (a new subject to diagnose), combined with simple *a priori* knowledge: z-score uses the statistical distribution of individual attributes in the original dataset to decide which attributes in the input deviate most, and therefore, the system should take care of before. The maximum-degree elicitation implicitly assumes that higher grades in an attribute are more relevant for the recommendation (a diagnosis). Both strategies try to simulate the elicitation behavior of an expert (a clinician in our example) with knowledge about the application domain. It may be hypothesized that these strategies could lead to an optimized conversation process.

The variable importance and logistic coefficients elicitation methods are used in the machine learning field as measures of attribute relevance and their main use is attribute selection, that is, determining the essential features or variables in a dataset with respect to a given problem. They use knowledge about the original dataset but they are independent of the input, that is, different inputs will have the same elicitation *preference* as indicated by the strategy.

We study the effect of these elicitation methods on the conversational metrics defined earlier, since, as commented in the previous section, classification metrics remain the same, in our experi-

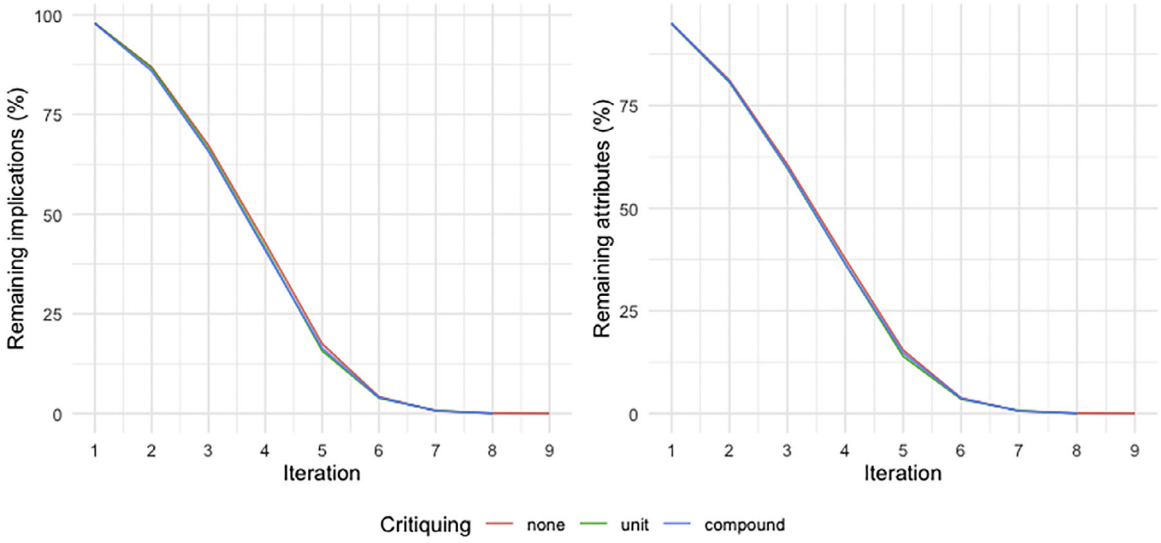

Fig. 3. Comparison of the evolution of the fraction of remaining implications (left) and remaining attributes (right) after each step of the conversation, for the three critiquing strategies.

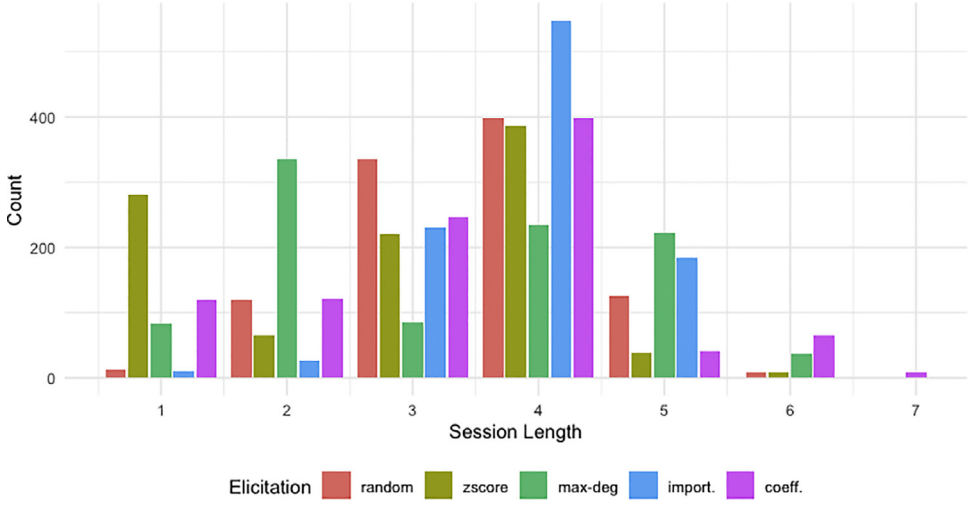

Fig. 4. Proportion of experiments that finished the dialogue after the given number of steps, for each type of elicitation.

ments, when we use different critiquing and elicitation setups. We used the same experimental procedure as described in the previous section.

We found that there exist statistically significant differences among all the elicitation strategies when comparing session lengths. Fig. 4 shows the distribution of validation cases with respect to its session length, grouping by the elicitation method used. Random elicitation had an average of 3.49 steps of length. The fastest sessions where obtained by z-score (2.83 steps), maximum degree (3.26 steps) and logistic coefficients (3.33 steps) elicitation methods. Interestingly, variable importance presented longer sessions (3.84 steps) on average than the rest of elicitation mechanisms. If we consider unit critiquing in addition to the different elicitation methods, we achieve a decrease of the average session length when using variable importance (3.81 steps) and logistic coefficients (3.26 steps) methods. In Table 2, a simple comparison of average session lengths is presented for all configurations tested in our experiments.

With respect to the reduction in the number of implications and attributes to explore after each conversation iteration, Fig. 5 shows the different behaviors of the system when using the proposed elicitation methods. The elicitation methods that more effectively reduce the search space and optimize the conversation process by reducing implications and attributes to explore are the z-score and maximum-degree strategies, closely followed by random and logistic coefficients elicitation methods. The strategy of using variable importance as elicitation preference obtained the worst results in the first steps, meaning that the measure is not well-suited for this task.

It is also confirmed the relationship between the decrease in the number of implications and the number of attributes in each step.

5. Discussion

In this section, we provide some discussion on the results obtained with our experiments.

We have compared our proposal with other recommender systems and machine learning methods. As mentioned in the Introduction, the problem we have presented presents some difficulties to recommendation systems, since the both data structure and availability are different to the standard setting. These difficulties have led to a poor performance of well-known recommender sys-

Table 2
Average session length in our experiments depending on the critiquing and elicitation strategies.

Elicitation:	Random	*z*-score	Max. degree	Var. importance	Logistic coeff.
No critiquing	3.49	2.83	3.26	3.84	3.33
Unit critiquing	3.42	2.83	3.26	3.81	3.26
Compound critiquing	3.45	2.83	3.26	3.84	3.33

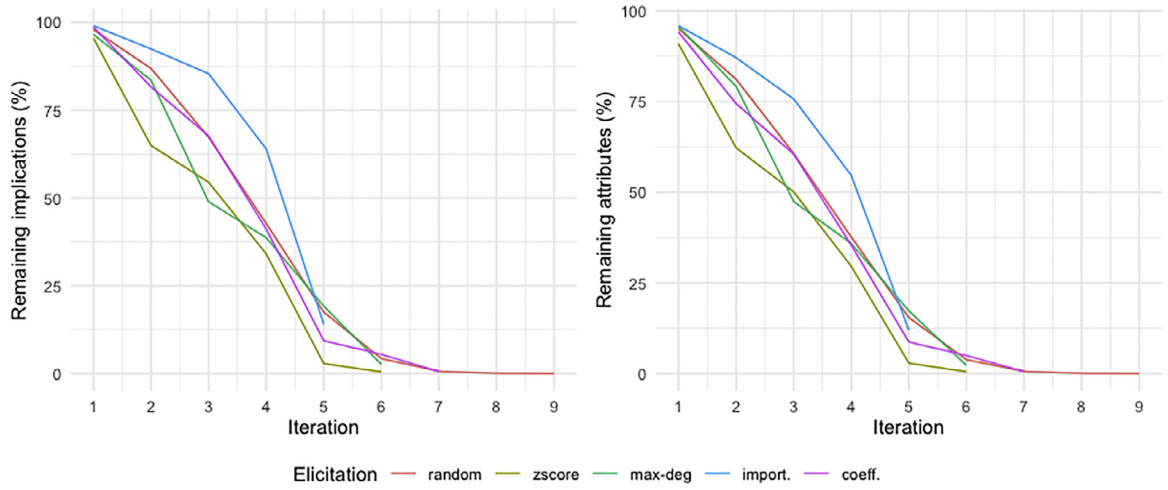

Fig. 5. Comparison of the evolution of the fraction of remaining implications (left) and remaining attributes (right) after each step of the conversation, for the elicitation strategies.

tems (UBCF, IBCF, and based on matrix factorizations) with this dataset.

Machine learning techniques, however, are better suited for this task and have achieved higher accuracy metrics. Some of them (random forests and eXtreme Gradient Boosting) are considered state-of-the-art in many fields, particularly in classification and regression tasks. Our proposal is able to achieve a better accuracy than those techniques and it can be said that it performs in aggregate as well as both methods.

This indicates that the logic tools used as engine to reason from data is a promising basis for the development of new recommendation techniques. Furthermore, other techniques based on statistical inference and interpretations, although performing well, lack explicability and interpretability, thus the use of logic tools is more appropriate.

We have also conducted experiments to compare conversation dynamics depending on two factors (the presence or absence of critiquing phase, and the mechanism to elicit attributes at each step of the conversation). The use of critiquing provides very discreet improvements, in contrast to previous studies (Chen & Pu, 2012b), mainly because the underlying logic is capable of reasonably completing (using the notion of semantic closure) the information accumulated throughout the conversation, thus requiring minimal corrections by the user. The experimental results confirm that the evolution of the conversation is not affected if critiquing is allowed or not.

Another strategy has also been studied to guide the conversation, establishing objective criteria to generate elicitations, simulating different levels of knowledge about the domain of the problem, from completely random (no knowledge) to basing the elicitation on statistical properties of the data, where a greater knowledge about the application domain is assumed.

It has been shown that those elicitation mechanisms that take into account knowledge about the problem and apply it to decide important attributes of the input into the system achieve conversations with fewer cycles of interaction compared to random elicitation, demonstrating that knowledge of the problem complements the knowledge deduced by FCA's own methods. Interestingly, the mechanism based on the determination of the importance of the variables in the original dataset achieves worse results than random elicitation. This indicates that the statistical importance of an attribute may not be a good indicator of which variables allow a more efficient conduct of the conversation.

On the other hand, the results also show that the complexity of the task of generating a recommendation is reduced by using the logic of simplification, measured by the number of applicable rules and the number of attributes to explore at each step of the conversation. An effective reduction in these parameters manages to alleviate the problem of exploring attributes, not by attacking the problem of high dimensionality at once, but by steps guided by logic.

As a general remark, it can be said that the use of the simplification logic leads to an improved mechanism for conversational recommenders, which makes better use of the information and knowledge implicit in the data, and serves to guide the conversation more efficiently.

6. Conclusions, impact and further research

Formal Concept Analysis has been used as an interesting tool to develop recommendations. Normally, the methods applied are based on clustering taking advantage of the concept lattice computed in FCA, or methods based on matrix. FCA extracts from the datasets concepts and implications and we face the design of recommender systems using automated methods for implications based on the Simplification Logic. The use of logic as the core of recommendation engine is the novelty of our proposal.

Specifically, our approach has been developed in the Fuzzy Formal Concept Analysis framework. Graded implications are extracted from the dataset as the background knowledge linking the fuzzy attributes (symptoms and diagnosis). We have proposed a closure operator based on the Fuzzy Attribute Simplification Logic as the reasoning engine to guide the conversational method. The

closure operator provides an enrichment of the set of symptoms until a disease is found in the closed set of attributes.

These formal tools in the framework of FCA have allowed the development of a conversational recommender system to make medical diagnosis strongly based on fuzzy logic with reasoning capabilities. More specifically, the conversational strategy has been enriched using the so-called *incremental critiquing*, and more specifically, the recommend-review-revise, strategy.

We have shown the performance and advantages of our approach with the development of a recommendation system for the differential diagnosis of schizophrenia with real-world data. From the dataset selected, we have extracted 20,663 graded implications and we have run 1000 simulations of the execution of the conversational recommender system with and without the critiquing stage. The results indicate that the critiquing strategy is able to accelerate the dialogue, reducing the number of required steps to convergence, while being able to infer more information and further reduce the dimension of the problem. An extensive comparison with the main recommender and Machine Learning systems used in the literature for this kind of dataset has been done. We show that our method is very promising and we improve the results with respect the main metrics of these other recommenders. Particularly, we show that our proposal is, at least, comparable to the best-performing machine learning systems with respect to classification metrics: our proposal achieves a 98% accuracy in the test problem, also with high values for sensitivity and specificity, followed by random forests with a 96% accuracy. Also, we have tested several conversational strategies, depending on critiquing and elicitation mechanisms, and found that although critiquing is not as relevant as in other works (Chen & Pu, 2012b), due to the logic tools employed, the elicitation strategy may help in guiding the conversation in a more efficient manner.

As future work, we will apply collaborative filtering strategies to our proposal. The mix of both strategies, providing a conversational collaborative filtering strategy, seems to be promising and allows us to tackle, in a unified approach, the accuracy and efficiency issues.

Although for the current work, the diseases are mutually exclusive, the system developed is prepared and may work even better when there are comorbidities, that is, several diseases may appear together with different grades. This possibility could help the expert to distinguish between complex and related diagnostics of diseases with even incomplete intersections of symptoms.

The framework have been designed in such a way that it could be used as a ICT service that can interoperate with electronic health records (EHRs), providing recommendations as a patient's EHR is updated. Thus, the use of the closure operator would help to infer signs and symptoms which may be unnoticed or undiscovered, thus filling in the gaps and missing data in the EHR. Interestingly, the proposed system, in that situation, would be extended to be able to rise alerts about possible unnoticed signs.

We also have in mind to incorporate techniques based on sentiment analysis for critiquing-based recommender systems in the line of Chen, Yan, and Wang (2019).

Regarding how to take advantage of the hidden knowledge in the dataset in order to guide the conversational process, we consider that FCA can bring some interesting ideas, such as the use of the implicit knowledge contained in the concept lattice, the use of minimal generators, attribute exploration techniques, and other implication bases which may reduce the computational complexity of the task.

To conclude, we plan to do an online experiment with real users. Presumably, user preference over feature can be modeled, and user feedback can provide more information to update this model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

P. Cordero: Conceptualization, Validation, Formal analysis, Investigation, Supervision, Writing - review & editing. **M. Enciso:** Methodology, Validation, Visualization, Project administration, Funding acquisition, Writing - review & editing. **D. López:** Software, Validation, Investigation, Resources, Data curation, Visualization, Writing - original draft. **A. Mora:** Validation, Formal analysis, Investigation, Visualization, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgments

This work has been partially supported by the projects TIN2017-89023-P and PGC2018-095869-B-I00 of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

Addington, D., Addington, J., & Schissel, B. (1990). A depression rating scale for schizophrenics. *Schizophrenia Research, 3*(4), 247–251.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734–749.

Afolabi, A. O., & Toivanen, P. (2018). Recommender systems in healthcare: Towards practical implementation of real-time recommendations to meet the needs of modern caregiving. In *Handbook of research on emerging perspectives on healthcare information systems and informatics* (pp. 323–346).

Aine, C., Bockholt, H. J., Bustillo, J. R., Cañive, J. M., Caprihan, A., Gasparovic, C., … Lauriello, J., et al. (2017). Multimodal neuroimaging in schizophrenia: description and dissemination. *Neuroinformatics, 15*(4), 343–364.

Alqadah, F., Reddy, C. K., Hu, J., & Alqadah, H. F. (2015). Biclustering neighborhood-based collaborative filtering method for top-n recommender systems. *Knowledge and Information Systems, 44*(2), 475–491.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175–185.

Anelli, V. W., & Noia, T. D. (2019). *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 3001–3002).

Aufaure, M.-A., & Le Grand, B. (2013). Advances in FCA-based applications for social networks analysis. *International Journal of Conceptual Structures and Smart Applications, 1*(1), 73–89.

Belohlavek, R. (2002). Algorithms for fuzzy concept lattices. In *Proceeding of fourth international conference on recent advances in soft computing* (pp. 200–205). Nottingham, United Kingdom: Nottingham Trent University.

Belohlavek, R., Cordero, P., Enciso, M., Mora, Á., & Vychodil, V. (2016). Automated prover for attribute dependencies in data with grades. *International Journal of Approximate Reasoning, 70*, 51–67.

Belohlavek, R., & Vychodil, V. (2006). Attribute implications in a fuzzy setting. *Lecture Notes in Computer Science, 3874*, 45–60.

Benito-Picazo, F., Enciso, M., Rossi, C., & Guevara, A. (2018). Enhancing the conversational process by using a logical closure operator in phenotypes implications. *Mathematical Methods in the Applied Sciences, 41*(3), 1089–1100.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems, 46*, 109–132.

Borrás, J., Moreno, A., & Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications, 41*(16), 7370–7389.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Cakir, O., & Aras, M. E. (2012). A recommendation engine by using association rules. *Procedia - Social and Behavioral Sciences, 62*, 452–456.

Calero Valdez, A., & Ziefle, M. (2019). The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies, 121*, 108–121.

Castellanos, A., De Luca, E., Garcia-Serrano, A., & Cigarran Recuero, J. M. (2015). Time, place and environment: Can conceptual modelling improve context-aware recommendation?. In *Proceedings of the fifth workshop on context-awareness in recommendation and retrieval*.

Chemmalar Selvi, G., Lakshmi Priya, G. G., & Joseph, R. B. (2019). *A FCA-based concept clustering recommender system*: 1.

Chen, L., & Pu, P. (2012a). Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction, 22*, 125–150.

Chen, L., & Pu, P. (2012b). Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction, 22*(1-2), 125–150.

Chen, L., Yan, D., & Wang, F. (2019). User perception of sentiment-integrated critiquing in recommender systems. *International Journal of Human-Computer Studies, 121*, 4–20. Advances in Computer-Human Interaction for Recommender Systems

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *KDD* (pp. 785–794). ACM.

Chin, W.-S., Yuan, B.-W., Yang, M.-Y., Zhuang, Y., Juan, Y.-C., & Lin, C.-J. (2016). Libmf: A library for parallel matrix factorization in shared-memory systems. *The Journal of Machine Learning Research, 17*(1), 2971–2975.

Christakopoulou, K., Beutel, A., Li, R., Jain, S., & Chi, E. H. (2018). Q&r: A two-stage approach toward interactive recommendation. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. In *KDD '18* (pp. 139–148).

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 815–824).

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning proceedings 1995* (pp. 115–123). Elsevier.

Davis, D. A., Chawla, N. V., Christakis, N. A., & Barabási, A.-L. (2010). Time to care: A collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery, 20*(3), 388–415.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. (1997). *User's guide for the Structured clinical interview for DSM-IV axis I disorders SCID-I: Clinician version*. American Psychiatric Pub.

Folino, F., & Pizzuti, C. (2010). A comorbidity-based recommendation engine for disease prediction. In *2010 IEEE 23rd international symposium on computer-based medical systems (CBMS)* (pp. 6–12). IEEE.

Frank, E., & Witten, I. H. (1998). Generating accurate rule sets without global optimization. In J. Shavlik (Ed.), *Fifteenth international conference on machine learning* (pp. 144–151). Morgan Kaufmann.

Ganter, B., & Wille, R. (1999). *Formal concept analysis: Mathematical foundations*.

Guo, G. (2012). Resolving data sparsity and cold start in recommender systems. In *Proceedings of the 20th International conference on user modeling, adaptation, and personalization* (pp. 361–364).

Hors-Fraile, S., Rivera-Romero, O., Schneider, F., Fernandez-Luque, L., Luna-Perejon, F., Civit-Balcells, A., & Vries, H. d. (2018). Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: A scoping review. *International Journal of Medical Informatics, 114*, 143–155.

Ignatov, D. I., & Kuznetsov, S. O. (2008). Concept-based recommendations for internet advertisement. *ArXiv, abs/0906.4982*.

Ikemoto, Y., Asawavetvutt, V., Kuwabara, K., & Huang, H.-H. (2019). Tuning a conversation strategy for interactive recommendations in a chatbot setting. *Journal of Information and Telecommunication, 3*(2), 180–195.

Jannach, D., Shalom, O., & Konstan, J. (2019). *Towards more impactful recommender systems research*: 2462.

Jooa, J., Bangb, S., & Parka, G. (2016). Implementation of a recommendation system using association rules and collaborative filtering. *Procedia Computer Science, 91*, 944–952.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin, 13*(2), 261–276.

Khanian Najafabadi, M., Naz'ri Mahrin, M., Chuprat, S., & Sarkan, H. M. (2017). Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. *Computers in Human Behavior, 67*, 113–128.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*: 26. Springer.

Lafta, R., Zhang, J., Tao, X., Li, Y., & Tseng, V. S. (2015). An intelligent recommender system based on short-term risk prediction for heart disease patients. In *2015 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT): 3* (pp. 102–105).

Laranjo, L., G Dunn, A., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., … Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association, 25*(9), 1248–1258.

Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems, 74*, 12–32. doi:10.1016/j.dss.2015.03.008.

McSherry, D. (2001). Minimizing dialog length in interactive case-based reasoning. In *Proceedings of the 17th international joint conference on artificial intelligence: 2* (pp. 993–998).

Medina, J., Pakhomova, K., & Ramírez-Poussa, E. (2017). Interpreting and analyzing a location-based social network by fuzzy formal contexts. In *2017 IEEE Symposium series on computational intelligence (SSCI)* (pp. 1–6).

Mezni, H., & Abdeljaoued, T. (2018). A cloud services recommendation system based on fuzzy formal concept analysis. *Data & Knowledge Engineering, 116*, 100–123.

Montenegro, J. L. Z., da Costa, C. A., & da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications, 129*, 56–67.

Mora, Á., Cordero, P., Enciso, M., Fortes, I., & Aguilera, G. (2012). Closure via functional dependence simplification. *International Journal of Computer Mathematics, 89*(4), 510–526.

Musto, C., Narducci, F., Lops, P., de Gemmis, M., & Semeraro, G. (2019). Linked open data-based explanations for transparent recommender systems. *International Journal of Human-Computer Studies, 121*, 93–107. Advances in Computer-Human Interaction for Recommender Systems

Narducci, F., de Gemmis, M., Lops, P., & Semeraro, G. (2018). Improving the user experience with a conversational recommender system. In *International Conference of the Italian Association for Artificial Intelligence* (pp. 528–538).

Nenova, E., Ignatov, D. I., & Konstantinov, A. V. (2013). An FCA-based boolean matrix factorisation for collaborative filtering. *CEUR Workshop Proceedings, 977*, 57–73.

Nilashi, M., Ibrahim, O. b., & Ithnin, N. (2014). Hybrid recommendation approaches for multi-criteria collaborative filtering. *Expert Systems with Applications, 41*(8), 3879–3900.

Phan, L. P., Huynh, H. H., & Huynh, H. X. (2017). User based recommender systems using implicative rating measure. *International Journal of Advanced Computer Science and Applications, 8*(11).

Priyogi, B. (2019). Preference elicitation strategy for conversational recommender system. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 824–825).

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Reilly, J., McCarthy, K., McGinty, L., & Smyth, B. (2005). Incremental critiquing. *Knowledge-Based Systems, 18*(4), 143–151. AI-2004, Cambridge, England, 13th-15th December 2004

Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). An extensive study on the evolution of context-aware personalized travel recommender systems. *Information Processing and Management, 57*(1).

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2010). *Recommender systems handbook*. Springer-Verlag.

Siegel, A. (2016). *Practical business statistics*. Academic Press.

Simpson, G., & Angus, J. (1970). A rating scale for extrapyramidal side effects. *Acta Psychiatrica Scandinavica, 45*(S212), 11–19.

Thong, N. T., & Son, L. H. (2015). Hifcf: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis. *Expert Systems with Applications, 42*(7), 3682–3701.

Tran, T. N. T., Atas, M., Felfernig, A., Le, V. M., Samer, R., & Stettinger, M. (2019). Towards social choice-based explanations in group recommender systems. In *Proceedings of the 27th ACM conference on user modeling, adaptation and personalization* (pp. 13–21).

Vesin, B., Ivanović, M., Klašnja-Milićević, A., & Budimac, Z. (2012). Protus 2.0: Ontology-based semantic recommendation in programming tutoring system. *Expert Systems with Applications, 39*(15), 12229–12246.

Wang, L., Alpert, K. I., Calhoun, V. D., Cobia, D. J., Keator, D. B., King, M. D., … Turner, M. D., et al. (2016). Schizconnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *NeuroImage, 124*, 1155–1167.

Wiesner, M., & Pfeifer, D. (2014). Health recommender systems: Concepts, requirements, technical basics and challenges. *International Journal of Environmental Research and Public Health, 11*, 2580–2607.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*(1), 1–17.

Wu, G., Luo, K., Sanner, S., & Soh, H. (2019). Deep language-based critiquing for recommender systems. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 137–145).

Zhang, X., Xie, H., Li, H., & Lui, J. C. S. (2019). Toward building conversational recommender systems: A contextual bandit approach. arXiv:1906.01219, abs/1906.01219.

Zhou, Y., Wilkinson, D., Schreiber, R., & Pan, R. (2008). Large-scale parallel collaborative filtering for the netflix prize. In *International conference on algorithmic applications in management* (pp. 337–348). Springer.