User Modelling
An Overview

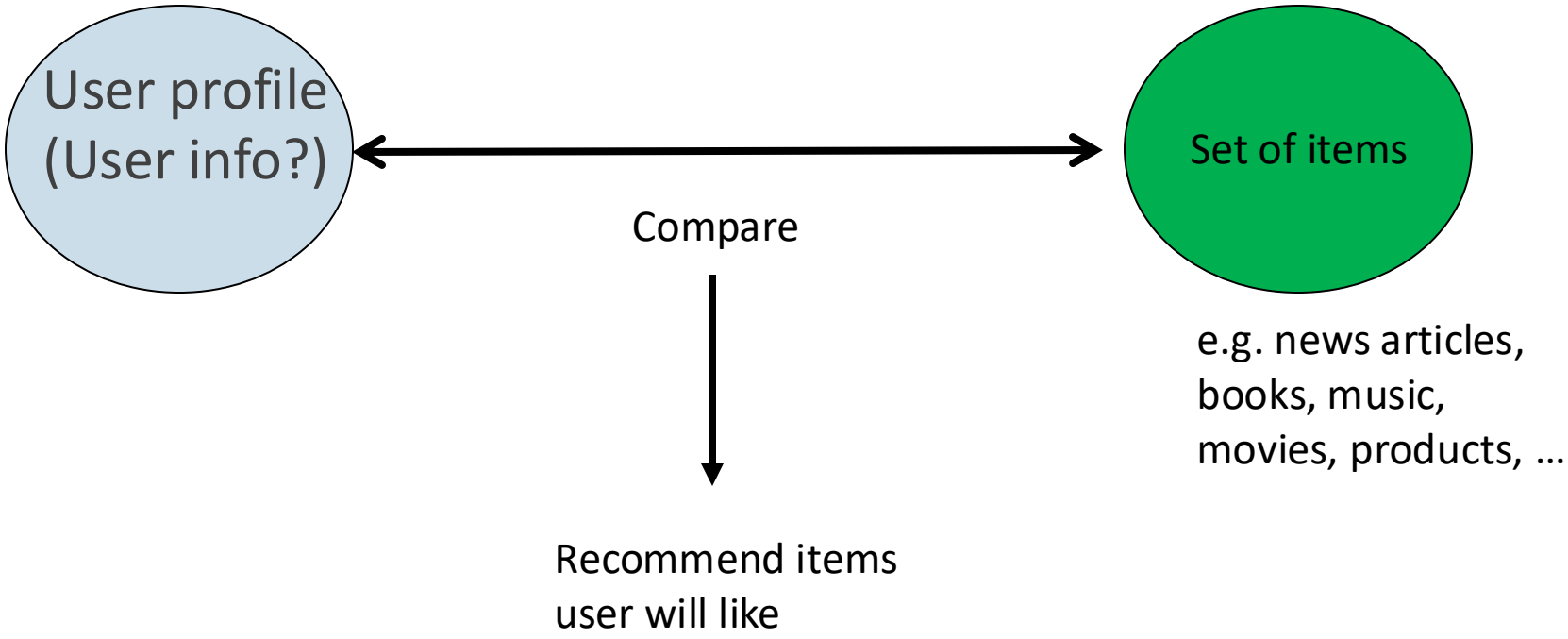
MEI

CONSTANTINO MARTINS, CATARINA FIGUEIREDO, DULCE MOTA AND FÁTIMA RODRIGUES

Disclaimer

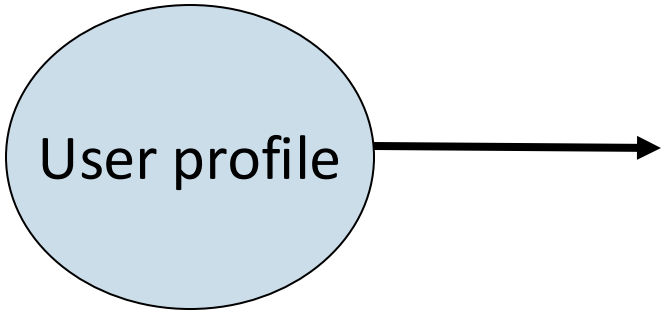
- **Some of this material/slides are adapted from several:**
  - Presentations found on the internet;
  - Papers
  - Books;
  - Web sites
  - …

RS - general Idea?



Personalisation: General Idea

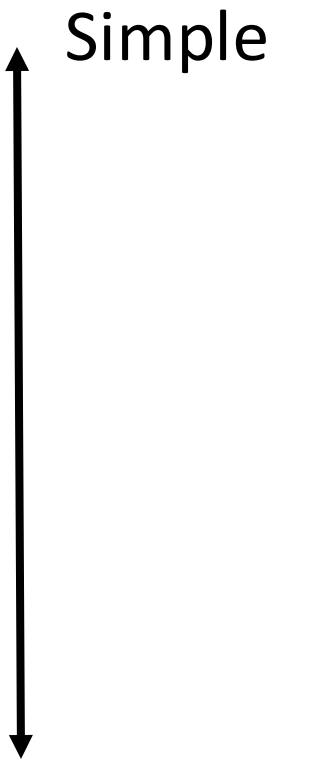
Personalization = user adaptive systems?



interaction is adapted based on data about an individual user
Eg personal websites, personalized tutoring, personalized recommendations, etc.

User Profile

- Demographical info
  - Age, gender, location, …
- Interests, preferences, expertise level, …
- Purchase records, observed behavior
- Ratings
- …
- Complete “lifelog”



User Modelling – User Profiling

- Important role in recommendation processes
- Models represent the user profile or a model of user interests and/or preferences
- User model is generally represented in the form of a user profile which captures the personal preferences of the users in terms of the user's knowledge about the object or subject in which they are interested

The user profile allows users to be modelled

Source of user profile

- Information about the user's preferences or interest in items can be **explicit or implicit**
- Entered **explicitly** by user (example: questionnaire)
- Gathered **implicitly** by system
  - ✓Observing/recording person's behavior
  - ✓Learning/inferring interests/preferences/level…
- Combination of both approaches

- Another dimension: **public/private**

Location of User Profile

- Centralized:
  - ✓Generic
  - ✓Device & application independent
  - ✓Easier to apply generalization across users
- Distributed
  - ✓Mobile use
  - ✓Better privacy
- Mixed forms

User Model

- A user model is composed by a set of characteristics that adjust the content, presentation and navigation to each user
- These characteristics can be:
  - Domain Dependent Data (DDD)
  - Domain Independent Data (DID)

Domain Dependent Data (DDD)

- Is related with system responses tailored according to the domain knowledge of a user
- Direct Dialogue or **explicit**: users to input and share their knowledge (for example, using questionnaires or forms) and mechanisms to process the inserted data to correctly measure user knowledge regarding the domain.
- Indirect Acquisition **or implicit**: Indirect acquisition method allows the system to assess user knowledge indirectly according to how the user performs different actions

Domain Independent Data (DID)

- Is composed of two elements:
  - The psychological model
  - The generic user profile model
- One of the advantages of this data is that it is supposedly unaltered, allowing the system to understand some of the characteristics it should adapt to

Examples of UM characteristics

Model	Profile	Characteristics	Descriptions/Examples
Domain Independent Data	Generic Profile	Personal Information	Name, email, password, etc.
		Demographic Data	Age, Gender, etc.
		Patient Background	Smoker, Pregnant, etc.
		Health data	User heal data
		Deficiencies: visual or others	Sees well, uses eyeglasses, etc.
		Domain of application	Localization of the user, etc.
		Inheritance of the characteristics	Creation of stereotypes that allow to classify the user
		Knowledge (Background Knowledge)	A collection of knowledge translated in concepts. Possibility of a qualitative, quantitative or probabilistic indication of concepts and knowledge acquired for the user
	Psychological Profile	Cognitive Capacities	
		Traits of Personality	Psychological profile (introvert, extrovert, etc.)
		Personal Preferences	Likes and Dislikes
		Inheritance of characteristics	Creation of stereotypes that allow to classify the user
Domain Dependent Data		Objectives	Questionnaires to determine user objectives
		Complete description of the navigation	Kept register of each page accessed
		Knowledge acquired	A collection of knowledge translated in concepts.
		Item Intake	Data related to patient item intake
		Context model	Data related with the environment of the user
		Aptitude	Definition of the capacity to use the system
		Task Preferences	Definition of the individual preferences with the objectives to adapt the navigation and contents

User Profiling process
Three main phases

MEI

CONSTANTINO MARTINS, CATARINA FIGUEIREDO, DULCE MOTA AND JOAQUIM SANTOS

Phase 1 - Information collection

- The system needs relevant information about the user's preferences or interests
- Basically, the systems can gather user interests or preferences from user feedback
- Feedback can be **explicit** or **implicit**

Phase 1 - Information collection

- Explicit user information collection approaches depend on users inputting personal information
- This information can be acquired directly via forms or questionnaires, or by asking users to rate items, or by tracking users' queries words
- For example, many sites collect user preferences by providing personalised services to users and then directly asking them to give personal information to create a profile

Phase 1 - Information collection

- Explicit information can include:
  - ✓Demographic information (e.g., gender, educational background, age, location, and occupation)
  - ✓Data about interests and preferences (e.g., topics of interest, tastes, preferred products and brand preferences)
  - ✓Opinion-based information (e.g., reviews, comments, and feedback)

Phase 1 - Information collection

- Explicit ratings data is widely used to profile users' preferences operate by using explicit ratings data (Netflix, which utilises movie ratings to generate popular movie suggestions for customers)

Problems of explicit information?

Phase 1 - Information collection

- **implicit user information** is based on **user behaviour**
- The implicit user information or implicit user feedback can be collected through web usage logs, click streams, browsing histories, purchase records, and content or structural information from visited web pages
- Browsing histories are a common source of implicit information, extracted user's browsing contents of each web page in a session to compute user's real-time preference.

Phase 1 - Information collection

- With Web 2.0, some new kinds of user information can be used as implicit user information, such as tags, comments, images, videos, posts, and click-streams. This data provides rich information about the relationship among users
- For example, the keywords on tags can be used to capture the user's topic interests (Amazon uses the usage logs of users to recommend books to their customers)

Problems of implicit information?

Difficulty to convert user behaviour into user preferences, as the accuracy depends on whether the user behaviour is interpreted correctly
For example, users might buy items such as music for someone else

Phase 1 - Information collection

- **Ethics and Privacy** concerns may cause some users to withhold information or behave differently when logged in to the system
- However, the advantage of user profiling lies in the access to both implicit and explicit user preference information
- Determining new users' preferences is challenging because **limited information** is available, and even that may be **inaccurate**

Phase 2 - Profile construction and representation

- The profiles can likewise be generated from either **implicit** user data (e.g., sets of keywords, web usage data, content and structural information about visited web pages, user ratings data, and demographic information) or **explicit** user data (e.g., questionnaires or interviews with the user).
- In most recommender systems, user ratings for items are widely utilised to indicate their item preferences; this is called a **rating-based user profile**

Phase 3 - Exploiting information in a user profile to provide personalised services

- After a user profile is constructed, it is then used to provide personalised services in different areas, such as personalised recommender systems, personalised searches, queries, and trust-aware recommender systems
- Three main methods approach: content-based, collaborative, and hybrid

User Modelling Techniques

MEI

CONSTANTINO MARTINS, CATARINA FIGUEIREDO, DULCE MOTA AND JOAQUIM SANTOS

Stereotypes

- One of the easiest and most common techniques for building models of other people is the evocation of stereotypes
- Stereotypes were first introduced in the literature related to User modelling by Elaine Rich in 1979, and it was brought with the necessity to define a “useful mechanism for building models of individual users on the basis of a small amount of information about them”

Stereotypes

- In order to correctly define and use stereotypes it is necessary to collect and use two kinds of information:
  - The first required information is related to the stereotypes themselves which includes the information of different collections of clusters of characteristics or facets
  - Facets may include characteristics such as interests, level of expertise, preferences, behaviors, or any other attribute considered significant for the purpose of the system
  - These facets depend on the domain and purpose of the system
  - These different facets will result and describe different groups of users

Stereotypes

- The second kind of information is related to the use of triggers which correspond to the occurrence of different events and that in turn will activate appropriate stereotypes
- For example, if a user performs an advanced task while using the system, an “expert user” trigger could be activated

Linear models

- Are one of the most common techniques
- These models are easy to build and understand; they are efficient and assume probabilistic data as plausible effects
- Has been a successfully employed theory so far
- They generally use weighted sums or means of frequently accessed items to conclude user interests, and, in the case of product applications described previously, can infer the likelihood for new unknown items satisfy the users

Markov Model

- A Markov Model follows a structure very similar to a Linear Model and consists of a set of states, a set of probabilities which determine the likelihood of transition between these states and, for each state, a set of observation/probability pairs
- For example, a Markov Model could be used to predict user most frequent actions while using the system by looking at his past performed actions

Bayesian Network

- A Bayesian Network is a directed acyclic graph where nodes denote variables and the arcs connecting nodes represent causal links from parent nodes to child nodes
- Each node is associated with a conditional probability distribution which assigns a probability to each possible value of this node for each combination of values of its parent nodes
- Examples of Bayesian Network models could be to predict the most adequate type of suggestions for a user according to the type of action being performed, or to predict error rates while the user is using the application

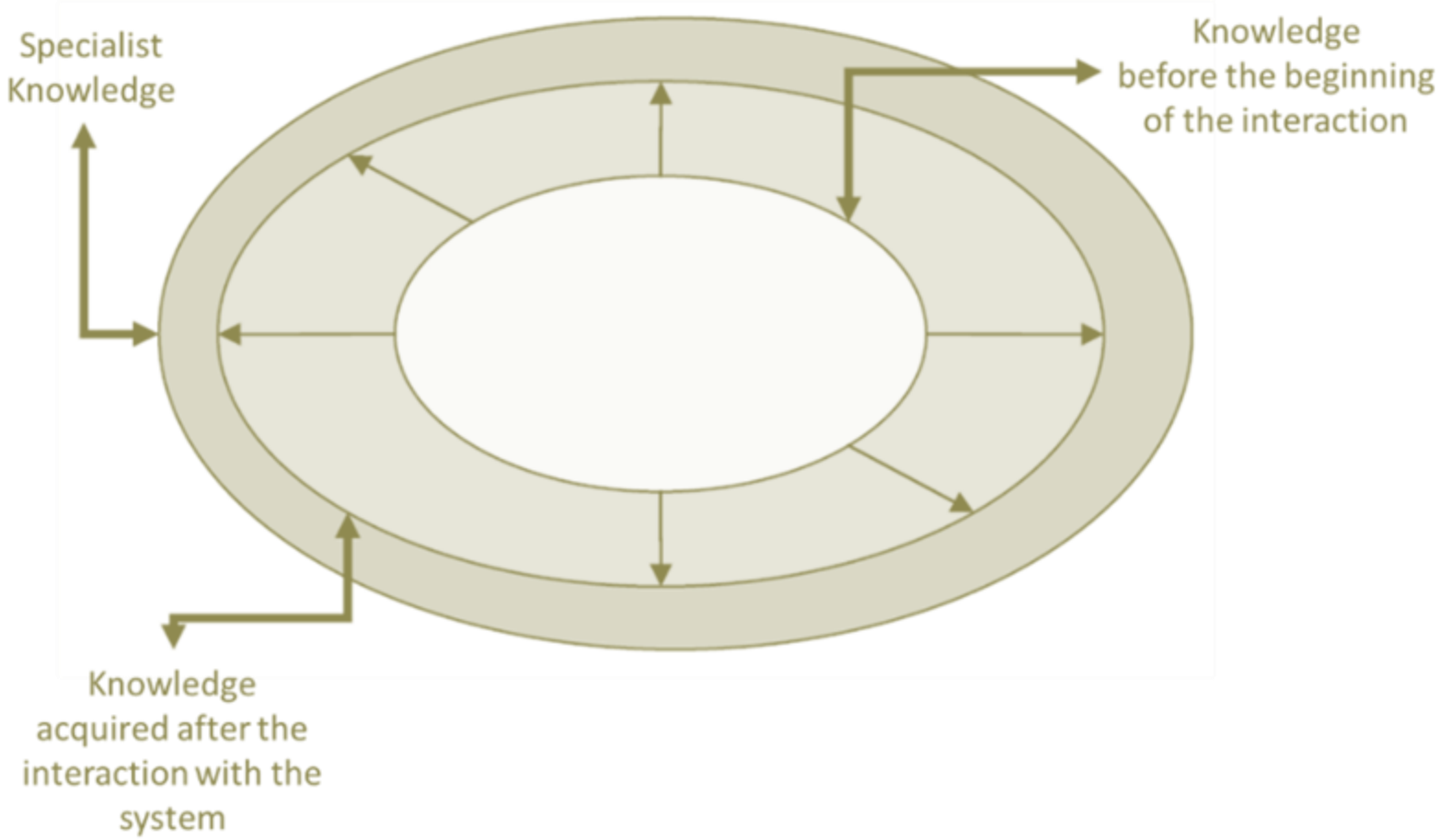
Behavioral adaptation

- Behavioral adaptation is referenced as the most suitable to be used and can be implemented in two ways:
  - ✓Overlay
  - ✓Perturbation

Overlay Model

- An overlay model assumes that the user's knowledge is a subset of the domain knowledge
- The overlay model consists of (a subset of) the concepts from the underlying domain model. For each concept, the overlay model contains data that represents (an estimation of) the individual user’s knowledge about or interest in this concept (or some other relationship with this concept)
- The expression of the knowledge level of each concept is dependent on the Domain Model itself: this value can be binary (knows or ignores), qualitative (good, average, weak, etc.) or quantitative (the probability of knowing or not, a real value between 0 and 1, etc.).

Overlay Model



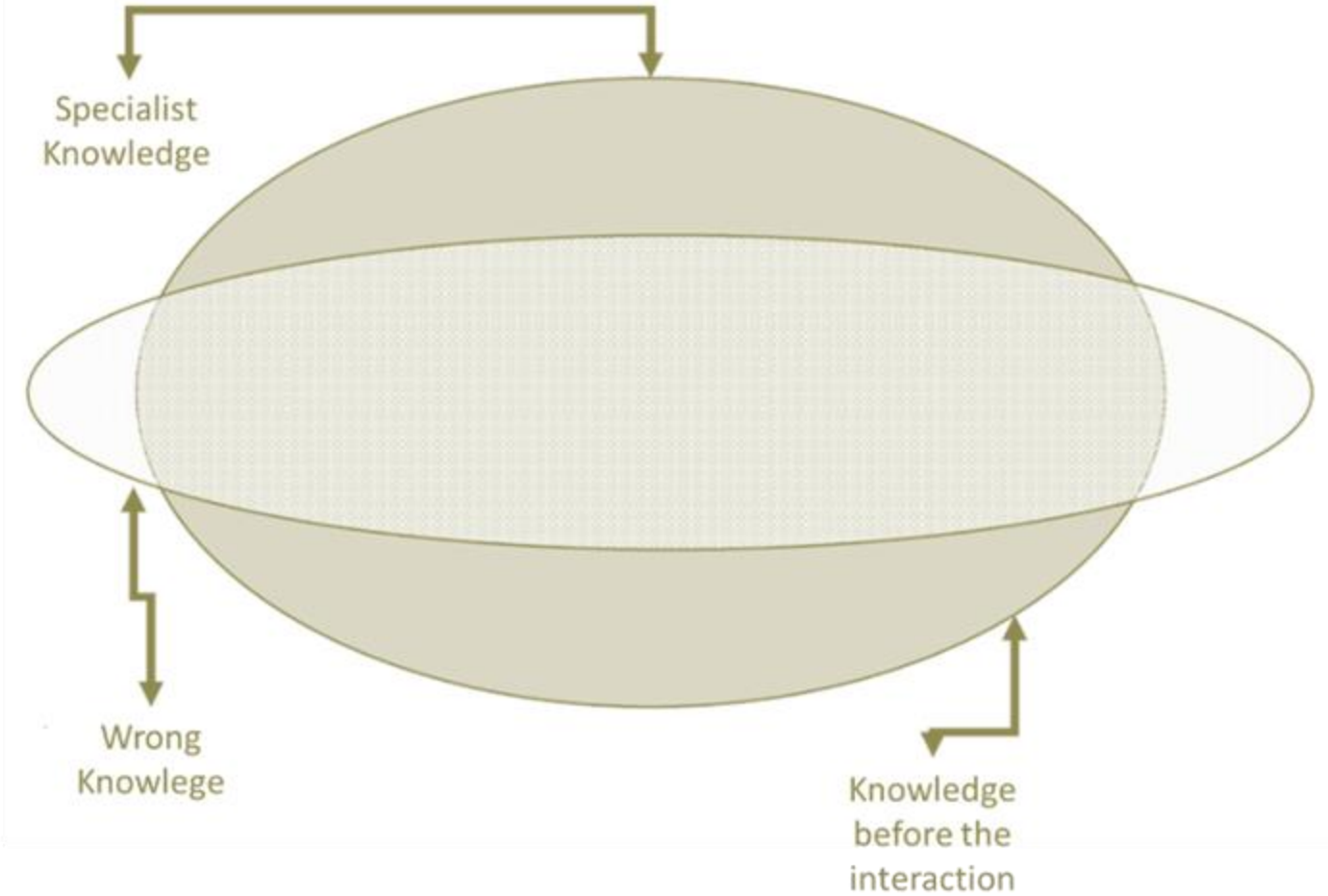
Perturbation model

- The perturbation model can represent user beliefs that the overlay model cannot handle
- A perturbation user model assumes that the beliefs held by the user are similar to the knowledge the system has, although the user may hold beliefs that differ from the system's in some areas
- These differences in the user model can be viewed as perturbations of the knowledge in the domain knowledge base

Perturbation model

- Thus, the perturbation user model is still built with respect to the domain model but allows for some deviation in the structure of that knowledge
- Perturbation model represents users as the subset of expert's knowledge plus their mal-knowledge
- This method considers that the knowledge and the student aptitudes are a perturbation of the specialist knowledge, and not a subset of his knowledge (as in the previous model)
- This method can be used to represent knowledge that is beyond the Domain Model defined by the specialist.

Perturbation model



Decision trees

- Decision trees are also a straightforward technique to use, and probably the easier to understand visually
- The trees have nodes representing the different attribute values or choices, spreading until a solution, or inference is found
- Generally, decision trees have the disadvantage of needing expert knowledge to be created and to be evolved. They represent knowledge limited in time and do not support new situations, which makes them high maintenance

Association rules

- Generally applied to supermarket shopping carts and are also named basket analysis algorithms, and their operation mode is elementary.
- Some examples are: those who buy baby bottles also buy milk, or those who buy chicken also buy beer
- The **Apriori algorithm** is one of the most used techniques. It uses a heuristic, which allows it to avoid the combinatory explosion issue, by discarding rules whose items do not have enough case support
- The name "Apriori" refers to the apriori principle, which is a fundamental premise of the algorithm. This principle states that if a set of items is frequent, then all its subsets are also frequent

Clustering

- K-Means is the most recognized clustering algorithm that attempts to build an initially known number of clusters, by iteratively relating each item with its closer cluster, using the K-Nearest neighbour heuristic and redefining each cluster

Data mining - Classification

- Classification — This technique tries to classify new items according to the classification of previous items
- It analyses attributes and finds the ones that will better contribute to create the knowledge associated with the classification process
- Generally, the representation that results from classification algorithms

More techniques

- Neural networks are one of the most recent techniques used in user modelling
- Text mining: One of the biggest challenges of text mining algorithms is to correctly deal with all the nuances and vocabularies peculiarities, such as different meaning words written the same way
- NLP
- LLM?

Comparison Between User Modelling Techniques

Technique	Advantages	Disadvantages
Linear models	• Simple to use and understand • Efficient • Lots of application domains • Easy to modify	• Not suitable for complex knowledge representations
Decision trees	• Extremely easy to read • Good performance, mainly in the case of binary trees • Can tackle cold-start issues because doesn't need initial knowledge	• Require expert knowledge • Hard to maintain and change
Neural networks	• Good performance • Can evolve over time autonomously	• Can take more time that we wished to converge to optimal results
Classification	• Can result in intuitive and useful decision trees • Can assist in decision-making process	• Needs a substantial amount of data to be efficient
Clustering	• Can discover invisible data groups • Can detect isolated cases, if that is an objective	• Challenging to deal with isolated cases • Difficult to define the ideal number of clusters
Association rules	• Can detect invisible item associations • Can assist in decision-making process	• Can result in unimportant, illogical or useless associations
Text Mining	• Only way to extract knowledge from text • The way to cope with content-based filtering	• Textual information means a totally complex domain to correctly explore
Bayesian networks	• Efficient • Represents both initial and future facts • Evolves autonomously	• Needs expert knowledge for the initial assumptions

References

Serão colocadas depois da entrega do primeiro trabalho