Semantic Reasoning: A Path to New Possibilities of Personalization^{*}

Yolanda Blanco-Fernández, José J. Pazos-Arias, Alberto Gil-Solla,
Manuel Ramos-Cabrer, and Martín López-Nores

Department of Telematics Engineering, University of Vigo, 36310, Spain
{yolanda,jose,agil,mramos,mlnores}@det.uvigo.es

Abstract. Recommender systems face up to current information overload by selecting automatically items that match the personal preferences of each user. The so-called content-based recommenders suggest items similar to those the user liked in the past, by resorting to syntactic matching mechanisms. The rigid nature of such mechanisms leads to recommend only items that bear a strong resemblance to those the user already knows. In this paper, we propose a novel content-based strategy that diversifies the offered recommendations by employing reasoning mechanisms borrowed from the Semantic Web. These mechanisms discover extra knowledge about the user's preferences, thus favoring more accurate and flexible personalization processes. Our approach is generic enough to be used in a wide variety of personalization applications and services, in diverse domains and recommender systems. The proposed reasoning-based strategy has been empirically evaluated with a set of real users. The obtained results evidence computational feasibility and significant increases in recommendation accuracy w.r.t. existing approaches where our reasoning capabilities are disregarded.

1 Introduction

Recommender systems provide personalized advice to users about items or services they might be interested in. Currently, these tools are gaining momentum in the Digital Revolution, helping people efficiently manage content overload and reducing complexity when searching for relevant information.

To fulfill these personalization needs, three main components are required in a recommender system: (i) a database where the available items are stored, (ii) personal profiles where the users' preferences are modeled, and (iii) recommendation strategies aimed at selecting personalized suggestions for each individual. The first such strategy was the so-called *content-based filtering*, which suggests to a user items similar to those he/she liked in the past. In spite of its accuracy, this technique is limited due to the employed similarity metrics. These metrics are based on rigid syntactic approaches

^{*} Work funded by the Ministerio de Educación y Ciencia (Gobierno de España) research project TSI2007-61599, by the Consellería de Educación e Ordenación Universitaria (Xunta de Galicia) incentives file 2007/000016-0, and by the Programa de Promoción Xeral da Investigación de la Consellería de Innovación, Industria e Comercio (Xunta de Galicia) PGIDIT05PXIC32204PN.

that only detect similarity between items that share all or some of their attributes [1]. Consequently, traditional content-based approaches lead to *overspecialized suggestions* including only items that bear a strong resemblance to those the user already knows (i.e. items with attributes defined in his/her profile).

To fight overspecialization, researchers devised a new strategy named *collaborative filtering*, based on offering to each user items that were appealing to others with similar preferences (named *neighbors*). Collaborative filtering reduces the effects of overspecialization by considering other users' interests, but it also causes new limitations, such as scalability problems, difficulties to select each user's neighborhood when the available preferences are sparse (commonly named *sparsity problem*), and privacy concerns related to the confidentiality of the users' personal data (see [1] for details).

Bearing in mind the severe drawbacks of the collaborative solutions, we propose a novel *content-based strategy* that exploits the main strengths of this personalization paradigm and overcomes the overspecialized nature of its recommendations. For that purpose, our strategy diversifies the offered suggestions without resorting to other users' preferences, thus protecting their privacy. Specifically, we fight syntactic limitations of the existing content-based approaches by employing two reasoning techniques borrowed from the Semantic Web field: the so-called *semantic associations* [3] and *Spreading Activation techniques* (henceforth, SA techniques) [7]. Instead of using the traditional syntactic similarity metrics, these associations trace semantic bonds between the user's preferences and the items available in the recommender system, which are previously formalized in a domain ontology along with their semantic annotations. Next, SA techniques efficiently explore these semantic relationships and discover new knowledge related to the users' interests. This knowledge permits our strategy to compare in a more flexible way the user's preferences with the available items, thus offering more accurate recommendations. Although the adopted reasoning mechanisms have been widely used in the Semantic Web [3,14,15], their internals must be adapted to fulfill personalization requirements of a recommender system. So, these mechanisms must allow to: (i) learn automatically new knowledge about the users' preferences from their feedback, and (ii) adapt dynamically the strategy as these preferences evolve.

In spite of the generality of our reasoning-based approach, in this paper we have adopted a specific context with the goal of describing in detail its use in a domain where the information overload is noticeable. Specifically, we have exploited the reasoning capabilities of our content-based strategy in order to enhance the recommendations offered to viewers of the Interactive Digital TV (IDTV). Today, TV viewers are exposed to overwhelming amounts of information, and challenged by the plethora of interactive functionality provided by the current digital receivers. As there are hundreds of channels with an abundance of programs available, it is likely that appealing TV programs go unnoticed. To assist these viewers, it is possible to take advantage of the personalization capabilities provided by a TV recommender system, which sifts through the myriad of programs available in the digital stream and selects those that match the viewers' preferences by using our reasoning-based strategy.

This paper is organized as follows: Sect. 2 describes the two key elements in our reasoning framework: (i) the ontology where the domain knowledge is formalized, including the available TV programs and their semantic descriptions, and (ii) the user

modeling approach employed to create the users' profiles. Next, Sect. 3 describes how the semantic associations and SA techniques are exploited in our content-based strategy. Then, a sample example where a set of TV programs are suggested to a given viewer is presented in Sect. 4. The tests carried out to validate our reasoning-based approach are explained in detail in Sect. 5. Finally, Sect. 6 draws some conclusions and points out possible lines of further work.

2 Domain Ontology and User Modeling

2.1 The Domain Ontology

Two elements are needed to formalize the IDTV domain by an ontology: (i) the semantic descriptions of the TV programs that can be suggested, and (ii) a language expressive enough to represent the concepts (i.e. classes and their instances) and relationships (i.e. hierarchical links and properties) identified in the domain. In our approach, the semantic descriptions have been extracted from TV-Anytime metadata specifications [6], whereas the OWL (DL) language has been selected due to its expressive capability, which allows to formalize concepts and expressions not supported in RDF and RDFS.

Starting from TV-Anytime metadata, we have defined and included in our OWL ontology several hierarchies of classes and properties, as well as specific instances of them, as shown in the TV ontology depicted in Fig. 1. The considered TV programs (identified by unique IDs) have been automatically extracted from the Internet Movie DataBase (IMDB) and the BBC web server¹, and are represented as specific instances belonging to a hierarchy of genres organized in several levels (e.g. *fiction*, *leisure*, *romance*, etc.), as shown at top of Fig. 1. The main attributes of these programs (e.g. involved credits, topics and places, intended audience, intention, etc.) are also instances related to them by labeled properties. These attributes also belong to hierarchically organized classes. As some of these classes are already defined in existing conceptualizations, we have imported ontologies about different domains such as sports, geographical information, credits involved in TV programs (e.g. actors), among others².

2.2 Our User Modeling Approach

Our approach models the user's profiles by reusing the knowledge available in the domain ontology, that is why we named them *ontology-profiles*. Specifically, we propose a semantic model for each user that gives information about: (i) the TV programs that were appealing or uninteresting for him/her (named *positive* and *negative preferences*, respectively), (ii) their main attributes, and (iii) the genres under which these programs are classified in the TV ontology (see at the top of Fig. 1). This user modeling approach provides a formal representation of the users' preferences, permitting to reason about them and discover additional knowledge about their interests. Such knowledge permits

¹ See <http://www.imdb.com> and <http://backstage.bbc.co.uk/data/7DayListingData> for details.

² These ontologies were extracted from the DAML repository located in www.daml.org/ontologies and converted to the OWL language by means of a tool developed by the MINDSWAP Research Group (see <http://www.mindswap.org/2002/owl.shtml> for details).

Fig. 1. Excerpt from classes, properties and instances in our TV ontology

to compare, in a more effective way, the users' preferences with the available items, thus leading to personalization processes more accurate than the traditional syntactic approaches [1]. In this regard, note that our ontology-profiles greatly improve other flat lists-based approaches which are not well structured to favor the discovery of new knowledge (see [4] for details).

Fulfilling the goals of our personalization strategy requires identifying the interest of the user in both TV programs defined in his/her profile and their attributes and genres. Specifically, these *Degrees Of Interest* (named DOI indexes and belonging to $[-1, 1]$) can be explicitly stated by the user or automatically inferred from his/her viewing behavior (e.g. programs accepted or rejected after recommendations, viewing time for each suggested program, etc.). Once the DOI indexes of each program in the user's profile have been established, we compute the indexes corresponding to their attributes and to

the genres under which these programs are classified in the ontology. This computation mechanism -omitted here due to space limitations- is explained in detail in [5].

Although other ontology-based proposals have been devised in literature, our user modeling approach differs to a great extent from these existing works. As an example, note the Quickstep system proposed by Middleton in [10], which suggests research papers according to the users' interests. The main difference between our work and Quickstep is related to the knowledge used for modeling purposes. In fact, Quickstep uses a simple taxonomy of research categories for representing the papers each user appreciates, whereas our proposal exploits the whole knowledge formalized in the ontology, permitting to carry out reasoning processes that discover extra information about users' preferences. The same limitation can be identified in the system proposed in [16], which recommends books according to the user preferences. There, the knowledge discovery is based on analyzing just only hierarchical relationships, thus hampering more complex inference processes as those pursued in our work.

3 Our Reasoning-Based Strategy

Our personalization strategy suggests programs that are semantically associated with the contents the viewer has liked in the past, improving the syntactic similarity metrics adopted in the traditional content-based methods. Specifically, our strategy consists of two stages -named filtering phase and recommendation phase, respectively-, which are sketched next and fully described in Sect. 3.1 and Sect. 3.2.

- *Filtering phase.* This stage selects in the OWL ontology instances of classes and properties that are relevant for the user, by considering his/her personal preferences. Next, our reasoning-based approach infers semantic associations among the selected entities identifying specific TV programs. These hidden associations -which we borrow from [3]- are discovered from the hierarchical links and properties defined in the domain ontology.
- *Recommendation phase.* The inferred knowledge is processed in the second phase by employing SA techniques. This intelligent mechanism works as concept explorer, as it detects concepts that are closely related to the user's preferences by exploring the entities and semantic associations inferred during the filtering phase.

3.1 Filtering Phase

Firstly, our strategy locates in the domain ontology the programs that were (un)appealing to the user (defined in his/her profile). Next, it traverses successively the properties bound to these programs until reaching new class instances (nodes referred to programs, actors, topics...) in the ontology. To guarantee the computational feasibility, we have developed a controlled inference mechanism that works as follows. As new nodes are reached from a given instance, our approach firstly quantifies their relevance for the user. Then, the nodes whose *relevance indexes* are lower than a specific threshold³ are ignored, in a such

³ The value of this threshold depends on both the domain ontology and the recommender system that adopts our content-based strategy. In our tests in DTV field, we have used values around 0.65.

way that our inference mechanism continues traversing successively only the properties that permit to reach new nodes from those that are significant for the user (according to his/her profile). Consequently, our strategy explores solely entities of interest for the user, thus filtering those that probably do not provide knowledge useful for the personalization process.

In our filtering mechanism, the more significant the relationship between a given node and the user's preferences (either positive or negative preferences), the more relevant this node. In order to measure this relevance value, we have developed a technique that takes into account diverse ontology-dependent *filtering criteria*. Some of these criteria –described in detail in [5]– are sketched next.

1. *Length of the chain of properties that enables to reach the considered node starting from the user's preferences.* Specifically, the longer this property sequence⁴, the lower the relevance index of the node, as its relationship to the user's preferences is less significant due to the presence of many intermediate nodes.

Example: Let us consider that a user has enjoyed the documentary *Learn about WW I* shown in Fig. 1. Here, it is possible to find the property sequence *Learn about WW I - World War I - Paths of Glory - Stanley Kubrick - Eyes Wide Shut*. In this case, the relevance index of the program *Paths of Glory* is greater than the index of *Eyes Wide Shut*, as the relationship between *Paths of Glory* and *Learn about WW I* is more significant than the relationship between this documentary and *Eyes Wide Shut*. In other words, the relation “movie about the World War I” is more relevant than “movie whose director has directed movies about the World War I”.

2. *Existence of hierarchical relationships between the node and the user's preferences.* The relevance of a node is increased when it is possible to find a common ancestor between it and the user's preferences in the ontology hierarchies.

Example: Let us consider again the user who has liked the war documentary mentioned in the previous example, whose topic is represented in our ontology by the instance *World War I*. In this case, the filtering phase increases the relevance index of other instances that share the common ancestor *War Topics* with the class instance *World War I* (e.g. *Vietnam War*).

3. *Existence of implicit relationships between the node and the user's preferences detected by concepts from graph theory.* In graph theory [8], the betweenness among three nodes is high when in the most of paths existing between the first and the second node, the third node is also included. So, from a high value of betweenness, it follows that the involved nodes are strongly related. In our approach, these nodes are the user's preferences and the class instance whose relevance is measured.

Example: Let us consider a user who has liked the movies *Vanilla Sky* and *Jerry Maguire* with *Tom Cruise* as leading actor. In this case, the relevance index of the instance *Born on the 4th of July* gets higher, as this movie is closely related to the user's preferences. In fact, as shown in Fig. 1, the node *Tom Cruise* is included in all the paths established between *Born on the 4th of July* and the two movies defined in the user's profile.

⁴ The length of a sequence is defined as the number of properties included in it.

Once the nodes related to the user's interests (and the properties linking them to each other) have been selected, our strategy infers semantic associations between the instances referred to TV programs. Specifically, we adopt three associations that have been defined by Anyanwu and Sheth in [3]:

- *ρ -path association*. In our approach, two programs are *ρ -pathAssociated* when they are linked by a chain or sequence of properties in the ontology. For instance, in Fig. 1, it is possible to trace a sequence between the documentary *Learn about WW I* and the movie *Paths of Glory* by means of the *World War I* instance.
- *ρ -join association*. Two programs are *ρ -joinAssociated* when their respective attributes belong to the same class in the domain ontology. For instance, in Fig. 1 there exists a *ρ -join* association between the documentary *Welcome to Tokyo* and the movie *Last Samurai*, as both programs are bound to different cities in Japan (*Tokyo* and *Kyoto*, respectively, which are classified as *Japanese cities* in Fig. 1).
- *ρ -cp association*. Two programs are *ρ -cpAssociated* when they share a common ancestor in the genre hierarchy defined in the ontology. For instance, note that all the movies depicted at the top of Fig. 1 are *ρ -cpAssociated* by the ancestor *Non Fiction Contents*.

By means of the filtering and knowledge inference processes, our approach has built a network for the user, whose nodes are the instances of classes selected during the filtering phase, and whose links are both the properties joining these instances in the ontology and the semantic associations inferred from it. The knowledge represented in this network is explored during the second phase of the strategy by exploiting the inference capabilities provided by SA techniques, which are one of the most-used processing frameworks for semantic networks.

3.2 Recommendation Phase

We emphasize the use of SA techniques as a computational mechanism able to: (i) explore efficiently the relationships among the nodes interconnected in the user's network (henceforth *SA network*), and (ii) infer from them knowledge useful for the recommendation process by detecting concepts closely related to the user's preferences. According to the guidelines established in [7], these techniques work as follows:

- The nodes of the network have an implicit relevance, named *activation level*. Besides, each link joining two nodes has a *weight*, in a such way that the stronger the relationship between both nodes, the higher the assigned weight. Initially, a set of nodes are selected and their activation levels are spread until reaching the nodes connected to them by links (named *neighbor nodes*).
- The activation level of a reached node is computed by considering the levels of its neighbors and the weights assigned to the links that join them to each other. Consequently, the more relevant the neighbors of a given node (i.e. higher their activation levels), and the stronger the relationship between the node and its neighbors (i.e. higher the weights of the links between them), the more relevant this node.

- This spreading process is repeated successively until reaching all the nodes of the network. Finally, the highest activation levels correspond to the nodes that are closest related to those initially selected.

The SA techniques have been widely used in the fields of searching and information retrieval [15,12,9]. However, to combine their inferential capabilities with the personalization requirements of a recommender system, it is necessary to extend the existing approaches. The modifications we propose affect mainly to two issues: (i) the kind of links traditionally modeled in the SA network, and (ii) their weighting process.

- On the one hand, the links considered in traditional approaches model only simple and direct relationships⁵, thus disregarding during the spreading activation process a huge amount of knowledge hidden behind more complex relationships. In order to fight this limitation, our SA network models both the properties defined in the ontology and the semantic associations inferred in the filtering phase. This way, the links corresponding to the associations allow to spread the relevance of the user's preferences until reaching programs appealing to him/her, which would go unnoticed in traditional SA-based approaches.
- On the other hand, as weights of the links depends only on the strength of the relationship between the connected nodes, these values remain static in existing SA approaches. Bearing in mind the purposes of a recommender system, our SA-based strategy must also consider the user's preferences during the weighting process, in a such way that this process adapts to changes in his/her interests.

In our approach, the proposed strategy activates in the user's network the nodes referred to the programs defined in his/her profile, and assigns them an initial activation level equal to their respective DOI indexes. Next, it is necessary to weight conveniently the links of the network, which represent both the explicit knowledge formalized in the ontology (i.e. properties), and the implicit knowledge discovered from it (i.e. semantic associations). As we mentioned in the previous section, our strategy adjusts dynamically these values as the user's preferences evolve over time. In fact, the weight of a link joining two nodes in our user's network is computed by considering their respective relevance indexes, which are measured during the filtering phase. As a result, our approach leads to a highly positive weight when the two linked nodes and the user's positive preferences are strongly related, and to a negative weight when the relationship is established to his/her negative preferences. This way, as the user's preferences change, the weights of the links in his/her SA network are conveniently modified and, hence, the elaborated recommendations are also updated.

According to the traditional SA techniques, once the propagation process has reached all the nodes in the user's network, the highest activation levels correspond to TV programs satisfying two conditions: (i) their neighbor nodes are also relevant for the user (that is why their high activation levels), and (ii) they are closely related to the user's preferences (that is why the high weight of the links). For that reason, these nodes identify the TV programs finally suggested by our content-based strategy.

⁵ For instance, in information retrieval, a link between two nodes referred to terms indicates their co-occurrence in a document (see [7] for details).

As a conclusion, note that the spreading process employed in our reasoning-based approach has three main advantages in the personalization field:

- Firstly, our strategy is able to discover that a TV program is appealing to the user even when its attributes are not defined in his/her profile. Thanks to the SA techniques, this program is relevant if it is semantically associated with the user's preferences. Consequently, our reasoning-based strategy offers diverse recommendations, beyond the overspecialized suggestions offered by the traditional syntactic content-based techniques.
- Secondly, our reasoning mechanisms consider both the positive and negative preferences of the user. Whereas the interests help to identify contents appealing to the user, the negative preferences decrease the activation levels of the nodes to which are related (either explicitly by means of properties, or implicitly by semantic associations). This way, our strategy prevents from suggesting programs associated with those the user did not like.
- Lastly, note that our approach not only favors the knowledge reusing, but also permits the user's network to adapt easily to changes in his/her preferences. This way, as these interests evolve, the filtering phase selects new nodes, properties and semantic associations, and incorporates them into the current network of the user.

4 A Sample Scenario

The example described in this section shows the differences between our reasoning-based recommendations and those offered by traditional content-based strategies. Due to space limitations, we consider only the brief excerpt from our TV ontology shown in Fig. 1. However, this restriction does not prevent from highlighting the associations used by our SA techniques for the selection of personalized recommendations.

In this scenario, we consider that the target user is *U*, who has enjoyed the documentaries *Welcome to Tokyo* and *Learn about World War I*, and the movies *Vanilla Sky* and *Jerry Maguire* starring *Tom Cruise*. As for *U*'s negative preferences, we assume that this user liked neither *Morgan Freeman* (supporting actor in *Million Dollar Baby*), and *Game of Death* (a movie about *martial arts* with *Bruce Lee* as leading actor).

Filtering Phase: Selecting Instances Relevant for *U*

Firstly, our strategy selects in the domain ontology instances that are relevant for the user *U* by considering his/her personal preferences. For that purpose, the strategy locates in the ontology the nodes referred to the programs defined in *U*'s profile, and explores successively the nodes joined to them by properties. For instance, from the node identifying *U*'s favorite actor (*Tom Cruise* in Fig. 1), it is possible to reach the instances *Born on the 4th of July* and *The Last Samurai*. Considering the second filtering criterion, our strategy selects the two reached instances, as they share common ancestors with *U*'s positive preferences. In fact, as shown in the hierarchy at the top of Fig. 1, *Born on the 4th of July* and *Jerry Maguire* are *Drama* movies, and *The Last Samurai* and *Vanilla Sky* are classified as *Action* movies.

Next, our filtering phase continues exploring the instances linked to the two previously selected nodes (i.e. *Vietnam War* from the node *Born on the 4th of July*, and *Kyoto* from *The Last Samurai*). In this case, both instances are relevant for U , as this user has appreciate other instances belonging to their classes in the ontology (specifically, *World War I* belonging to the *War Topics* class, and *Tokyo* belonging to *Japanese cities*).

We also search for class instances related to the user's negative preferences. These instances are relevant during the personalization processes, as they help to identify programs the user will probably not enjoy. Among such nodes, note *Danny the Dog*. As shown in Fig. 1, this movie not only involves an actor U does not like (*Morgan Freeman*), but also it is about *Karate*, a topic that seems to be unappealing to this user, who has not liked the movie about martial arts entitled *Game of Death*. This knowledge will be used by our SA-based reasoning to elaborate recommendations for U .

Filtering Phase: Inferring Semantic Associations Between TV Programs

Once the instances relevant for U have been identified, our strategy infers associations between the programs included in the selected property sequences (see Table 1).

Table 1. Property Sequences and Semantic Associations inferred for the user U

Sequence Properties	Semantic Associations
<i>Learn about WW I - World War I - Paths of Glory</i>	ρ -path (<i>Jerry Maguire</i> , <i>Born on 4th July</i>)
<i>Jerry Maguire - Tom Cruise - Born on 4th July - Vietnam War</i>	ρ -join (<i>Welcome to Tokyo</i> , <i>The Last Samurai</i>)
<i>Vanilla Sky - Tom Cruise - The Last Samurai - Kyoto</i>	ρ -join (<i>Learn about WW I</i> , <i>Born on 4th July</i>)
<i>Welcome to Tokyo - Tokyo</i>	ρ -cp (<i>Vanilla Sky</i> , <i>The Last Samurai</i>)
<i>Danny the Dog - Karate</i>	ρ -cp (<i>Danny the Dog</i> , <i>Game of Death</i>)
<i>Game of Death - Kung Fu</i>	ρ -path (<i>Danny the Dog</i> , <i>Million Dollar Baby</i>)

The next step is to build the user U 's SA network, by including as nodes the class instances selected by the filtering process, and as links both the properties that join these instances to each other in the ontology, and the associations inferred from it. Once the links have been weighted, U 's network (in Fig. 2) is processed by SA techniques, which reason about the represented knowledge to select the personalized recommendations.

Fig. 2. Network used by SA techniques to select content-based recommendations for U

Recommendation Phase: Suggesting TV Programs by Means of SA Techniques

After spreading the activation levels of U 's preferences until reaching all the nodes in the SA network, our strategy suggests the TV programs with the highest levels. According to what we explained in Sect. 3.2, these programs receive links from other contents which are appealing to the user U . This way, our content-based approach suggests to U the movies *Paths of Glory*, *Born on the 4th of July* and *The Last Samurai*, by employing the associations inferred between these contents and his/her positive preferences.

- **Paths of Glory:** The activation level corresponding to this movie is increased thanks to the weight spread from the program *Learn about WW I*. The link between the two movies in the user's SA network is due to the fact that both contents are about the *World War I* in which U is interested. For that reason, the relevance of *Learn about WW I* reaches the suggested movie *Paths of Glory*.
- **Born on the 4th of July:** The activation level of this movie gets higher thanks to the links from two programs relevant for U : *Learn about WW I* and *Jerry Maguire*. The war topic turns the documentary into a program appealing to the user, whereas the movie *Jerry Maguire* is relevant as it involves U 's favorite actor (*Tom Cruise*).
- **The Last Samurai:** As shown in Fig. 2, *Vanilla Sky* and *Welcome to Tokyo* inject positive weights in *The Last Samurai* node. Both programs are specially relevant for U , thus increasing the activation level of *The Last Samurai*, a movie the user will appreciate due to two reasons: (i) his/her favorite actor takes part in it, and (ii) the movie is set in a city of Japan, a country that seems to be interesting for U in view of the documentary about history and customs in Tokyo this user has liked.

Our SA network permits also to discover programs unappealing to U , which are identified by the low activation levels obtained after the spreading process. For instance, note the movie *Danny the Dog*, whose level is decreased by the negative weights received from the nodes *Morgan Freeman* and *Game of Death* (U 's negative preferences).

In conclusion, our content-based approach suggests not only programs with the *same* attributes defined in U 's profile, but also contents that are significantly related to his/her preferences. So, out of the available movies starring U 's favorite actor, we suggest those bound to war topic and Japan, two issues especially relevant for U according to the programs *Learn about WW I* and *Welcome to Tokyo* in his/her profile.

5 Testing

We have implemented a prototype of recommender system for testing purposes including: (i) the OWL ontology about TV domain, (ii) our user modeling technique based on *ontology-profiles*, and (iii) the proposed content-based strategy. Our experimental evaluation involved this prototype and 400 undergraduate students from University of Vigo. These users rated 400 programs extracted from our ontology, by assigning values in $[-1,1]$ (positive and negative values for appealing and uninteresting programs, respectively). A brief description was included for each program, so that all the users could judge even contents unknown for them.

Our goals were: (i) to evaluate the accuracy of our reasoning-based recommendations, and (ii) to compare this approach (henceforth *Asso-SA*) with other existing techniques that are devoid of our semantic inference capabilities. Due to space limitations,

here we include just only two well-known approaches defined in the field of personalization. The first one –proposed by O’Sullivan *et al.* in [13]– combines content-based and collaborative filtering with discovery of association rules to measure similarity between programs; the second approach –proposed by Mobasher *et al.* in [11]– devises a collaborative strategy enhanced by the semantic descriptions of the suggested items.

5.1 Approach Based on Association Rules

This approach (henceforth *Rules*) mixes a collaborative phase with a content-based stage to select personalized recommendations for a set of users. The collaborative phase compares (program to program) each user’s preferences against the profiles of the remaining users. Then, it extracts his/her k most similar neighbors. The similarity between contents is based on identifying association rules between them. These rules are automatically extracted by the Apriori algorithm [2], which works on a set of training profiles containing several programs. So, given two programs A and B , a rule $A \Rightarrow B$ means that if a profile contains A , it is likely to contain the program B as well. In fact, each rule $A \Rightarrow B$ has a confidence value interpretable as the conditional probability $p(B/A)$. This way, the similarity between A and B is computed as the confidence of the rule involving both contents, as long as this value is greater than a threshold confidence.

Once the k most similar profiles to the user’s one have been identified, the approach *Rules* applies the content-based phase. Here, the programs in these neighbors’ profiles (and unknown for the user) are ranked according to their relevance for this user and the top- N_1 are returned for recommendation. A program is more relevant when: (i) it is very similar to those contained in the user’s profile, (ii) it occurs in many neighbors’ profiles, and (iii) these neighbors’ profiles are strongly correlated to the user’s preferences.

5.2 Semantically Enhanced Collaborative Filtering

The goal in the approach of Mobasher *et al.* (henceforth *Sem-CF*) is to predict the rating of the user U on a given item t (in our case, a TV program). For that purpose, the *Sem-CF* approach selects the N_2 contents in U ’s profile which are most similar to the program t . Next, the level of interest of the user U in t is predicted as the sum of the ratings given by U on the N_2 contents more similar to t . Each rating is weighted by the corresponding similarity between t and the N_2 selected programs.

The similarity between two items i_p and i_q is computed by combining linearly two components: (i) the ratings of the users who have rated both items in their profiles (denoted by $RateSim(i_p, i_q)$), and (ii) the semantic attributes of the compared items (denoted by $SemSim(i_p, i_q)$). In order to compute $SemSim$ component, Mobasher *et al.* build a matrix S where each row refers to each one of the n items available in the recommendation process, and each column corresponds to the values of their respective semantic attributes. Once this matrix has been created, the *Sem-CF* approach applies techniques based on Latent Semantic Indexing (LSI) to reduce its dimension, resulting in a much less sparse matrix S' . Finally, Mobasher *et al.* apply the mathematical techniques described in [11] and obtain from S' a $n \times n$ square matrix in which an entry i, j corresponds to the semantic similarity between the items i and j ($SemSim(i, j)$).

5.3 Test Data

The preferences of the 400 users were divided into two groups: *training* and *test users*.

Training users (40%). The programs rated by these users with positive DOI indexes were employed to build the so-called *training profiles*. From these profiles, the similarity between two specific contents is computed in both the approach *Rules* and *Sem-CF*. In the rule-based work, these profiles are used to train the Apriori algorithm, whose task is to discover association rules between the programs contained in them and to quantify their similarity. Regarding *Sem-CF*, note that the *RateSim* component for two given programs is measured by selecting the training users who have rated both contents in their profiles, and computing the correlation Pearson-r between their respective ratings.

Test users (60%). The three evaluated strategies were executed by considering the preferences of these users (defined in the so-called *test profiles*). To identify these preferences, we proceeded as follows. Out of the initially rated 400 programs, we selected for each user those 10 ones most appealing to him/her and those 10 he/she found less interesting (according to their ratings), and we built their respective profiles. In this process, a low overlapping between the programs rated by these users was obtained, leading to a great sparsity level (89%). The remaining 240 programs rated by these users and their DOI indexes (named hereafter *evaluation data*) were hidden in order to be compared to the programs offered by each approach.

5.4 Methodology and Accuracy Metrics

Our evaluation is organized as follows. Firstly, the evaluated strategies were executed for each test user. Each strategy suggested 20 TV programs and as configuration parameters, we have used $k = 30$ neighbors and $N_1 = 20$ programs in *Rules*, $\alpha = 0.6$ and $N_2 = 10$ programs in *Sem-CF*, and finally, a filtering threshold of 0.65 in *Asso-SA*⁶.

Next, we compute for each user the values of the employed accuracy metrics: recall and precision. The recall of a strategy for a given user is the ratio of programs rated as interesting in his/her evaluation data (i.e. with positive DOI index) that were suggested by this strategy. The precision is the ratio of contents recommended by this technique that have a positive DOI in the user's evaluation data. Lastly, the average and variance of the recall and precision were computed over the 240 test users, to detect possible abrupt dispersions between the values measured for each user and the average value.

5.5 Discussion on Experimental Results

The results from Fig. 3 indicate that the reasoning capabilities of our approach increase greatly the recommendation accuracy, offering diverse suggestions the users appreciate and excluding the typical drawbacks of the two evaluated collaborative approaches (*Rules* and *Sem-CF*). In spite of high sparsity level in the used data (89%), our approach selected, in average, the greatest number of programs interesting to the test users (highest recall), and the lowest account of contents they found unappealing (highest precision). This improvement is due to the combination of semantic associations and SA

⁶ These values were chosen after several experiments because they offered the most accurate recommendations in each evaluated approach.

Fig. 3. Average (left) and variance (right) of recall and precision for the evaluated approaches

techniques as reasoning mechanisms. As shown in Fig. 3, these techniques discover programs of interest for the test users, which go unnoticed in the other two evaluated approaches devoid of reasoning capabilities.

Instead of reasoning about the knowledge of a domain ontology, the approach *Rules* extracts association rules between programs by considering just only the occurrence patterns detected by Apriori from the training profiles. In our tests, there was a low overlap among the programs rated by the training users, thus obtaining a reduced number of significant rules. Consequently, it was difficult to select accurately the k most similar neighbors for each user, thus offering recommendations less precise than ours.

Something similar happened in the approach *Sem-CF*, where many programs appealing to the test users could not be suggested due to both the intrinsic limitations of collaborative solutions and the absence of semantic reasoning processes. On the one hand, the low overlap among the programs rated by the test users affected to the computation of the *RateSim* component, hampering the creation of the users' neighborhoods in an accurate way. On the other side, the *SemSim* component only detected similarity between programs with the same attributes, thus disregarding the semantic associations established between them. Besides, according to what we described in Sect. 5.1 and Sect. 5.2, we checked that both *Rules* and *Sem-CF* suggested only programs defined in the test users' profiles. This fact caused the lower recall values of the two approaches w.r.t. our strategy, since many programs could not be suggested as they had not been rated by the users involved in our experiments.

6 Conclusions and Further Work

In this paper, we have proposed a personalization strategy that overcomes limitations unresolved in traditional content-based recommender systems, by resorting to reasoning mechanisms borrowed from the Semantic Web. Specifically, our strategy is based on: (i) filtering from a domain ontology instances that are irrelevant for the user by a threshold-based process, (ii) inferring semantic associations between his/her preferences and the

resulting instances, and (ii) exploring efficiently the discovered knowledge by Spreading Activation techniques to select the personalized recommendations. Our strategy is clearly promising for its generality, as it is not exclusively bound to a specific application domain. For that reason, the proposed technique is flexible enough to be reused in other recommendation domains, becoming a good starting point to implement diverse personalization services for the Semantic Web users.

Our content-based strategy has been implemented in a prototype of TV recommender system and evaluated in a scenario with real users. The obtained results show computational feasibility and significant increases in the accuracy of our suggestions w.r.t. traditional syntactic approaches, which are devoid of our reasoning capabilities.

Our further work is focused on two aspects. On the one hand, we are developing a mechanism to weight automatically the filtering threshold used in our strategy. Broadly speaking, this threshold will be dynamically adjusted as the user's preferences change, and its value will depend on both the domain ontology and the feedback provided by the user regarding the past recommendations. On the other hand, we also plan to continue the experimental evaluation of our strategy by involving the 80000 subscribers of the cable networks of Spanish operator R⁷, over which a recommender system using our reasoning-based strategy is being deployed.

References

1. Adomavicius, G., Tuzhilin, A.: Towards 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), 739–749 (2005)
2. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
3. Anyanwu, K., Sheth, A.: ρ -Queries: Enabling Querying for Semantic Associations on the Semantic Web. In: 12th International World Wide Web Conference, pp. 115–125 (2003)
4. Ardissono, L., Kobsa, A., Maybury, M.: Personalized Digital Television: Targeting Programs to Individual Viewers. Kluwer Academic Publishers, Dordrecht (2004)
5. Blanco, Y., Pazos, J.J., Gil, A., Ramos, M., López, M.: A Flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. *Knowledge-based Systems* (in press), <http://dx.doi.org/10.1016/j.knosys.2007.07.004>
6. Broadcast and On-line Services: Search, select, and rightful use of content on personal storage systems. ETSI TS 102 822 series (2003)
7. Crestani, F.: Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review* 11(6), 453–482 (1997)
8. Diestel, R.: Graph Theory. Springer, Heidelberg (2000)
9. Huang, Z., Chen, H.: Applying associative retrieval techniques to alleviate sparsity in collaborative filtering. *ACM Transactions on Information Systems* 22(1), 116–142 (2004)
10. Middleton, S.: Capturing Knowledge of User Preferences with Recommender Systems. PhD thesis, University of Southampton (2003)
11. Mobasher, B., Jin, X., Zhou, Y.: Semantically Enhanced Collaborative Filtering on the Web. In: Web Mining: applications and techniques, pp. 57–76. IDEA Group Publishing (2004)

⁷ <http://www.mundo-r.com>

12. O'Hara, K., Alani, H., Shadbolt, N.: Identifying Communities of Practices: Analyzing Ontologies as Networks to Support Community Recognition. *Information Systems*, 89–102 (2002)
13. O'Sullivan, D., Smyth, B., Wilson, D., McDonald, K.: Improving the Quality of the Personalized EPG. *User Modeling and User-Adapted Interaction* 14, 5–36 (2004)
14. Perry, M., Janik, M., Ramakrishnan, C., Arpinar, B., Sheth, A.: Peer-to-Peer Discovery of Semantic Associations. In: *Workshop on P2P Knowledge Management*, pp. 1–12 (2005)
15. Rocha, C., Schawabe, D., Poggi, M.: A Hybrid Approach for Searching in the Semantic Web. In: *13th International World Wide Web Conference*, pp. 374–383 (2004)
16. Ziegler, C., Lausen, G.: Exploting Semantic Product Descriptions for Recommender Systems. In: *2nd Semantic Web and Information Retrieval Workshop*, pp. 25–29 (2004)