Integrating information sources for recommender systems

Silvana Aciar [a,1], Josefina López Herrera [a] and Josep Lluis de la Rosa [a]
[a] *Agents Research Laboratory*
University of Girona
{saciar, peplluis}@eia.udg.es, lopez.herrera@udg.es

Abstract. This paper presents a Multi-agent System and a Methodology to select and to integrate heterogenous and distributed information sources to make recommendations. A set of intrinsic characteristics has been defined. These characteristics allow having a description of the information contained in the sources to select the most relevant information sources. Ontologies are used to integrate the information from the selected sources. And a case study confirms our proposal.

Keywords. Recommender Systems, Information Integration, Negotiation, Ontology

1. Introduction

Today an essential research challenge is the development of large-scale agent-oriented information systems that can connect the right information with the right people at the right time [12]. This challenge has been exacerbated by the explosive increase of the information available in the web. Models and techniques for multi-agents systems, information retrieval and recommender systems have emerged as research approaches to address this problem. Recommender systems present relevant information to users according to previous patterns of information retrieval and individual user model [8] what has been used to deal with the information overload problem [12]. In e-commerce applications, recommender systems need a responsive, strategic network of interchange of information that can respond instantly to requirements of the users. They need access to different information sources to find the information necessary to make the best recommendation that satisfies every user requirements.

To improve the recommendation through the interchange of information the problems are:

- Select the information source with the most appropriate information to a recommender.
- Integration of the information: to set more knowledge from disperse data bases.

Nevertheless the problem of selecting and integrating information from other sources is a difficult task,[1][11] the complexity is made by :

[1]Correspondence to: University of Girona, Campus Montilivi Tel.: +34 972 41 8478; Fax: +34 972 41 80 98

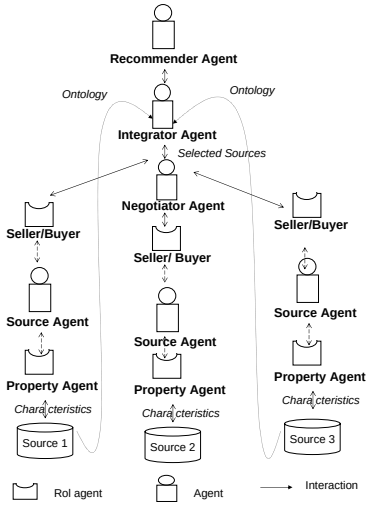

Figure 1. Multi-Agent System

- The dynamism of the sources.
- The geographic distribution.
- The heterogeneity of the sources.

A Multi-agent System and a Methodology are presented to solve this problem. This paper is organized as follows: In **Section 2** a Multi-agent System for selecting and integrating distributed and heterogenous information sources is presented. A methodology to select and integrate the information is described in **Section 3**. **Section 4** describes a case study. Finally, conclusions are presented in **Section 5**.

2. Multi-agent System

A Multi-agent system to select and integrate the information from distributed and heterogenous sources has been designed, it can be see in Figure 1.

A Multi-agent system is an Artificial Intelligence approach suitable to manage geographically distributed information [4] in which is necessary to have agents to mediate the differences between the components. The agents interact by a negotiation protocol for selecting the relevant sources, and the information is integrated with ontologies.

- Each information source is managed by a **Source Agent (SCA)**. It has different roles in the system, as **Property Agent (PA)** is the agent in charge of obtaining the description of the source. **Buyer or Seller Agents (BA/SA)** participate in the negotiation process by buying or selling information contained in the sources. The **BA** is a **SCA** agent that demands information and the **SA** agent offers information.
- **Negotiator Agent (NA)** is a mediator between the **BA** and the **SA**. It is responsible for selecting the sources which provide the most relevant information to make the recommendation.
- **Integrator Agent (IA)** integrates the information from the selected sources.

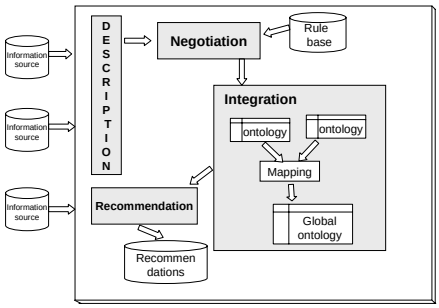

Figure 2. Functional view of the methodology

- **Recommender Agent (RA)**: makes the recommendation with information from the selected sources.

The agents execute the steps of the methodology explained in the next sections.

3. Methodology

This methodology attempts to provide access to information from multiple, distributed and heterogeneous information sources to make best recommendations. Figure 2 shows a functional view of the methodology.

3.1. Phase1: Description of the sources

The description of the information contained in the sources is interchanged among the agents to select the most appropriate source. A set of characteristics are defined to achieve:

- A representation of the information contained in the source;
- Criteria to compare and select a source.

In Figure 3 the characteristics defined for the present research are listed.

3.2. Phase 2: Negotiation

A negotiation protocol is applied to choose the relevant information source. The negotiation protocol is initiated by an BA agent. A user looking for a certain good or service contacts BA agent and provides it with all the necessary information, then:

1. The BA agent sends a message of requirement to the NA agent.
2. The NA agent sends a message of request about the description of the sources to all the SA agents in the system.
3. The SA agents answer with the description of the source (characteristics).
4. The NA agent selects the SA agents. The strategy applied by the NA is to choose the SA that offers the best information that satisfies the requirements of the BA. If none of the SA gives an acceptable offer the negotiation enters conflict. To solve this conflict the sources are selected by values of the characteristics near to an acceptable value.

Characteristics	Measure
Completeness: Number of users from one information source also found in another source	$Completene \quad ss = \frac{\sum (A \cap B)}{\sum A}$ $(A \cap B)$ = Users existing in both sources A, B = Users from one source of information
Diversity: Number of user groups. It allows the users to be to be grouped according to degree of similarity following a given criterion.	$H = -\sum (p_i * \ln p_i) \qquad p_i = \frac{n_i}{N}$ n_i = Number of users included in group i N = Total number of users in the source
Ontology: Semantic representation of the information contained in the sources	Number of relevant attributes for the recommendation, includes in the source
Timeliness: Update of the information about the users interactions.	$Timeliness = \frac{\sum wi * ci}{N}$ C_i = Number of user that purchased in a period of time i w_i = Weight of the period of time. N = Total number of user in the source
Frequency: Frequency of the user interactions	$Frequency = \frac{\sum wi * fi}{N}$ f_i = Number of user in a ratio of purchase frequency w_i = Weight of the a frequency of purchase N = Total number of user in the source

Figure 3. Source description

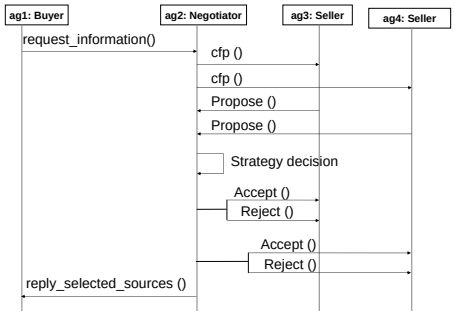

Figure 4. Negotiation protocol between the BA agent, NA agent and SA agents in this system

5. The NA answers to the BA with a list of the possible SA that has the source that contains information that satisfies his requirements.

The negotiation protocol is showed in the Figure 4.

3.3. Phase 3: Integrating the information from the selected sources

In addition to the capability of retrieving information from a large number of heterogenous sources is necessary an ontological approach to connect conceptually related information [7]. The RA agent can see a collection of physically distributed and heterogeneous data sources as relational databases structured according to a global ontology. The global ontology is specified by the mappings between the ontologies of each source. A "concept" in the global ontology is a subset of a cartesian product of a list of domains, i.e., if $D1, ..., Dn$ is a list of domains, then: $X \subseteq D1...Dn$ is a concept. The structure of a concept, X is described by a list of attributes as: $X = ((at_1, v_1); (at_2, v_2); ...; (at_n, v_n))$, i.e., Person= (("name", Juan), ("age", 25)) is the structure of a concept with two attributes. The concept can be formed with instances retrieved from one or more relevant data sources using a set of predefined queries. When the instances of a concept are fragmented across two or more ontologies. Thus, each information source stores values of a subset of attributes of the concept. It is assumed that the existence of a special concept

that is created a global ontology so that the corresponding fragments of each instance can be combined. Giving two concepts Y and Z into a new concept YZ involves combining each instance of Y with the corresponding instance of Z followed by taking the union of the instances of Y and instances of Z, $YZ = Y \bigcup Z$.

3.4. Phase 4: Making the recommendation

In a recommendation, relevant information is presented to users according to their preferences. Some methods are used to realize the recommendations. These methods are Collaborative Filtering [9], Content-based Filtering [2] and a hybrid approach between both methods [10]. This work is focussed on the selection and integration of the sources. When the relevant information sources have been selected it is possible to apply any of these methods.

4. Case study

Three data bases in the consumer package goods domain (retail) were used. The data bases are related tables that contain information of the retail products, 1200 customers and the purchases that they realized during the period 2001-2002. A data base S1 contains information about the purchases realized on Internet (Online). The data base S2 and S3 they contain information about the purchases realized at the store. The three data bases contain common customers. The experiments were based on the information from the table that contains information about the purchases realized. This table has 23 attributes, between which are identifier of the customers, identifier of the products purchased, import of the purchase, the quantity of units and date of the purchase. Basically these are the attributes used in this case of study.

4.1. Description of the sources

Figure 5 shows the values of the characteristics of each source. The values were obtained according to the equations defined in the section "*Description of the sources*" Figure 3. The characteristics shown in Figure 5 allow to know, in an abstract way , the information contained in the sources S1, S2 and S3. Observing the results the following conclusions are obtained: The source S1 contains the most relevant attributes for the recommendation, the most complete source is the source S2 and the source S3 is the most updated and the most diverse.

4.2. Selection of the sources

Once the characteristics of the sources were define, the selection is realized across the negotiation protocol between the agents of the system. The Buyer Agent (Source S1) and the SA agents (Source S2 and S3). The process of negotiation executed in this case of study gave a result that the selected source was the source S2. Once selected the source it waits that the results of the recommendation will be better, incorporating information from S2 into of S1. This is:

$$E(R(S1 + S2)) > E(R(S1)) \quad (1)$$

Characteristics	Source 1 (S1)	Source 2 (S2)	Source 3 (S3)
Ontology	0.80	0.50	0.20
Diversity			
Z (Zone)	0.13	0.11	0.12
F (Family)	0.66	0.67	0.67
S (Sex)	0.20	0.20	0.21
Completeness	0.10	0.60	0.30
Frequency	0.23	0.40	0.25
Timeliness	0.25	0.40	0.42

Figure 5. Intrinsic characteristics of the sources in the consumer packaged good domain

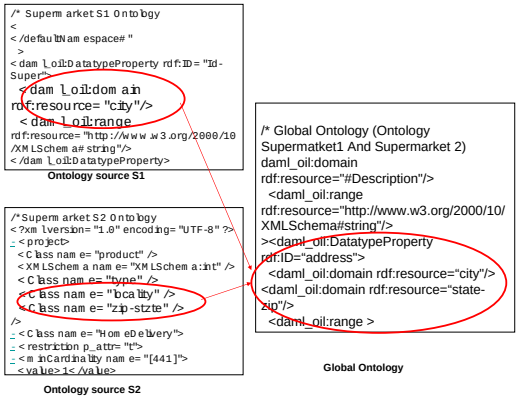

Figure 6. Global Ontology for S1 and S2

It hopes that the recommendations with information of the source S1 and information of the source S2 are better than the recommendations with only information of the source S1.

4.3. Integration of the information

In this first prototype the information to integrate is from the same domain (retail) although, for each source an ontology was created. The mapping through the query [3] is illustrated by an example. Assume a global concept corresponding called BEVERAGE with attributes ID, DESCRIPTION, TYPE. There are two data sources S1 and S2 with each ontology that contains information about BEVERAGE. In the global ontology the BEVERAGE concept is described in terms of global concepts associated with the sources S1 and S2 as follows:

Create Or Replace View BEVERAGE (Id, Name, Type)
As Select Id, Name, Type From S1supermarket.BEVERAGE
Union
Select Id, Name, Type From S2supermarket.BEVERAGE

The global ontology obtained from the mapping between the sources S1 and S2 is shown in Figure 6

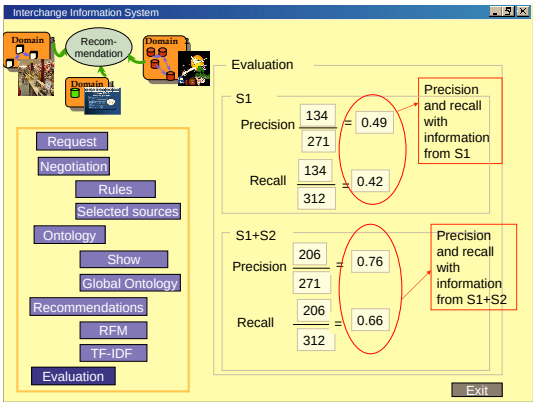

Figure 7. Evaluation of the recommendation

4.4. Recommendation

The TF-IDF (Term Frequency times Inverse Document Frequency) technique [6] has been used to make the recommendation. With the TF-IDF the relevance of the products for every customer can be established. A table is generated containing total frequencies of product using the next equation:

$$TF - IDF = T_{ik} * log_2(\frac{N}{n_k}) \tag{2}$$

Where, T_{ik} is the frequency of the product k in the purchase of customer i, n_k is the total number of customers that have purchased the product k and N is the total number of customers.

Two related criteria, normalized recall and normalized precision, have been used to evaluate the recommendation . The following equations [5] describe these measures:

$$Precision = \frac{Nrf}{nr} \tag{3}$$

$$Recall = \frac{Nrf}{nf} \tag{4}$$

Where Nrf is the number of recommended products that match with the purchases, Nr is the total number of recommended products and Nf is the total number of purchased products. Precision represents the probability of a recommendation to be successful. Recall measure represents the probability that a relevant product will be purchased. Applying the evaluation measures in the recommendations the result obtained are shown in the Figure 7.

The values of the precision and recall measures demonstrate the hypothesis E(R(S1+S2)) > E(R(S1)) can be supposed, because the method used to make the recommendation is the TF-IDF which is based on the purchase frequencies, and the source S2 has more purchases (Frequency measure) and is more updated (Timeliness measure) and

almost it includes all the clients of S1, then improvements were waited in the efficiency of the recommendation given by the precision and the recall.

5. Conclusion

This paper presents a SMA (Multi Agent System) and a methodology to select and to integrate information sources. The use of the intrinsic characteristics of the sources that describe the information contained in them let have a priori idea if a potential source can or can-not improve the efficiency of a given recommender system. The results obtained in the case study show that integrating information of a source with certain characteristics increases the efficiency of the recommender systems. The efficiency is obtained by the precision and recall measure. The future work considers to implement a mechanism of trust to infer the results of the recommendations from the characteristics of the sources.

References

[1] V. Arens, C. Y. Chee, C-N. Hsu, and C. A. Knoblock: Retrieving and integrating data from multiple information sources. International Journal on Intelligent and Cooperative Information Systems, 2(2),pp.127–158, 1993.

[2] M. Balabanovic and Y. Shoham: Fab: Content-Based, Collaborative Recommendation. Communications of the ACM,40(3),672,March,1997

[3] E. Mena, A. Illarramendi, V. Kashyap, and A. Sheth. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. International journal on Distributed And Parallel Databases (DAPD), 8(2),pp. 23–271, 2000.

[4] A. Moreno, A. Valls, J. Bocio : A multi-agent system to schedule organ transplant operations. Inteligencia Artificial, Revista Iberoamericana de IA. Special issue on Multi-Agent Sytems Development, n.13, pp.36–44, 2001

[5] G. Salton and M.J.McGill: Introduction to Modern Information Retrieval. McGraw-Hill Publishing Company, NewYork, NY, 1983.

[6] G. Salton. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, 1989.

[7] Semantic Web .Available at: http://www.w3.org/2001/sw/ 2004

[8] J.B. Schafer, J. Konstan and J. Riedl : Recommender Systems in E-Commerce. In: EC '99: Proceedings of the First ACM Conference on Electronic Commerce, Denver, CO, pp. 158–166. 1999

[9] U. Shardanand and P. Maes: Social Information Filtering. Algorithms for Automating Word of Mouth. SIGCHI Conference, 1995.

[10] B. Smyth and P. Cotter: A Personalized Television Listings Service. Communications of the ACM. ACM Press. 2000

[11] K. Sycara, K. Decker and M. Williamson : Modeling information agents: advertisement, organizational roles, and dynamic behavior, in Technical Report WS-9602, American Association for Artificial Intelligence, 1996.

[12] Y.Z. Wei, L. Moreau and N.R. Jennings: Recommender systems: A market-based design. In: Proceedings of International Conference on Autonomous Agents and Multi Agent Systems (AAMAS03), Melbourne pp.600–607, 2003