# Identifying Information from Heterogeneous and Distributed Information Sources for Recommender Systems

Silvana Aciar, Josefina López Herrera, and Josep Lluis de la Rosa

University of Girona
Campus Montilivi, edifice P4, 17071 Girona, Spain
{saciar,peplluis}@eia.udg.es, josefina.lopez@udg.es

**Abstract.** With easy access to World Wide Web the users are overloaded of information. Recommender systems have emerged as research approach to address this problem. The users want to find what they need, when they need it and under the conditions that they want.
These conditions drives to the recommender systems to access to different sources to find relevant information to recommend.
In this paper we presented a set of intrinsic characteristics to determine the relevance of the sources to make recommendations.
These characteristics allow to have a representation of the information contained in the sources that is relevant to recommend and a set of criteria to select the most relevant.
A multi-agent system has been designed to obtain these characteristics.
Preliminary results of recommendations made with the selected sources are presented in this paper.

## 1 Introduction

Today knowledge integration from various sources is gaining importance in areas such as Process Administration, Data Warehouse, e-commerce, Knowledge Administration and Marketing. The information integration allows more knowledge to be acquired from disperse information.

However, identifying, accessing and integrating information from various sources is a difficult task, made complex by [1][13]:

- the dynamism,
- the geographic distribution and
- the heterogeneity of the sources

The agent paradigm has emerged as tool to manage the complexity in the integration systems. It is a promising proposal for information recovery and a solution to distributed problems in which mechanisms of cooperation and coordination are necessary [7].

Today the quantity of available information is enormous, almost immeasurable and the complexity of the systems grows if the agents have to access to all the information sources [12] [4].

*© A. Gelbukh, R. Monroy. (Eds.)*
*Advances in Artificial Intelligence Theory*
*Research on Computing Science 16, 2005, pp. 55-64*

In this paper the agents select and integrate relevant sources for a recommender system. A recommender system receives as inputs the user preferences and generates recommendations of products or services according to its preferences [11].

A measurement based on the intrinsic characteristics of the sources is used to establish the relevance to select the information sources.

This paper is organized as follows: In **Section 2** the intrinsic characteristics and the relevance measurement used to select the sources are defined. The MAS (Multi-Agent System) designed to select the relevant information sources is presented in **Section 3**. In **Section 4** is presented the case of study with the obtained results. Finally, conclusions and future work are presented in **Section 5**.

## 2 Intrinsic Characteristics of the Information Sources

In this section a set of characteristics of the sources is defined that allow to establish a relevance value to select information sources to make the recommendations.

The set of characteristics provides:

- A representation of the information contained in each source, and
- a set of criteria to compare the sources and to select the most relevant.

In Table 1 the characteristics defined are listed.

### 2.1 Selecting the Information Sources

The selection of information sources is based on the relevance value obtained from the intrinsics characteristics. Relevance R(S) is defined as quantity of information that a source S can contain for a recommendation:

$$R(S) = \sum (f_i \times p_i \times [\frac{1}{\sum_{i=1}^{n}(f_i)}]) \quad (1)$$

Where $f_i$ = Weight of the characteristic i (*Completeness, Diversity, Frequency, ...*), $f_i >= 1$. The characteristic most relevant for the domain have the highest value. n = Total number of characteristic. $p_i$ = value of the characteristics i, it is obtained with the equations from Table 1.

### 2.2 Making the Recommendation

Information to make recommendations is only obtained from the information sources with the highest values of relevance.

Recommendation-based contents were used to make the recommendation in[2][10][8] [6]. The system processes information from various sources and attempts to extract characteristics and useful elements about their content. Content-based filtering techniques can vary according to their degree of complexity. For

| **Characteristics** | **Description** | **Measurement** |
|---|---|---|
| **Completeness** [14] | Number of users from one information source also found in another source | *Completeness* $= \sum(A \cap B) / \sum A$<br>$A \cap B$ = *Users existing in both sources*<br>A = *Users existing in the source A*<br>B = *Users existing in the source B* |
| **Diversity** [16] | Number of user groups The users are grouped according to a common criterion | *Diversity*= $H = \sum p_i * \log_2(p_i)$<br>$p_i = n/N$<br>n = *Number of users included in group i*<br>N = *Total number of users in the source* |
| **Ontology** | Semantic representation of the information *contained in the sources* | *Relevant Attributes* = $\frac{a}{A}$<br>a = *Number of relevant attributes i in the source*<br>A = *Number of relevant attributes j for the recommendation* |
| **Timeliness** [17] | Update of the information about the users interactions | *Timeliness* $= \sum w_i * c_i/N$<br>$c_i$= *Number of user that purchased i in a period of time i*<br>$w_i$ = *Weight of the period of time i*<br>N = *Total number of user in the source* |
| **Frequency** [17] | Frequency of the user interactions | *Frequency* $= \sum w_i * f_i/N$<br>$f_i$ = *Number of user in a ratio of purchase frequency i*<br>$w_i$= *Weight of the a purchase frequency*<br>N = *Total number of user in the source* |

**Table 1.** Intrinsic characteristics of the information sources

example, a search based on key words is one of the simplest. A more complex technique is the based on the extraction of semantic content from information contained in documents. In the present work the information to make recommendations is the user behaviour, i.e. the purchase frequency, the quantity purchased, the last date that a purchase was made and the product purchased.

For example, the following information is found in the data-base for user X:
Last date of purchase = 10/05/2005
Amount = 1000
Frequency = high
Product = computer

A computer would not be recommended to this customer because it is an infrequent product and the purchase date is very recent.

The RFM algorithm (Recency, Frequency, Monetary)was used to obtain knowledge of the user behaviour. This algorithm divide the users according to their last date of purchase, the frequency of their purchases and the quantity purchased. Based on this information, decisions concerning whether or not to make a recommendation to deserter customers can be taken. Analysis of behaviour is the key to obtain favourable responses to the recommendations made to customers.

## 3 Multiagent System to Select Relevant Information Sources

A MAS has been designed to select the relevant information sources. The MAS is shown in Figure 1

- Each information source is managed by a **Manager Agent (MA)** that has abstract information about the content of the source, defined in terms of the intrinsic characteristics. In addition, the agent is responsible for measuring the relevance of the source.
- The **Property Agents (PA)** are in charge of obtaining the characteristics. They are task agents [13] and interact directly with the source. There is an agent for each property of the source and their task is to measure the intrinsic characteristic.
- A **Selector Agent (SA)** is responsible for selecting the most relevant sources in order to make the recommendation.
- A **Recommender Agent (RA)** is an Interface agent [13]that interacts with the user. Once the sources are selected, it makes the recommendation.

The adaptability to the environment in a MAS is one of the most important characteristics. In this system the agents adapt themselves and learn from the environment, updating the values of the intrinsic characteristics of the sources. The agents periodically calculate these values, so data change, the values of the characteristics also change. For the selection phase, the relevance measure R(S) is calculated from these values, and is therefore also updated. Based on these

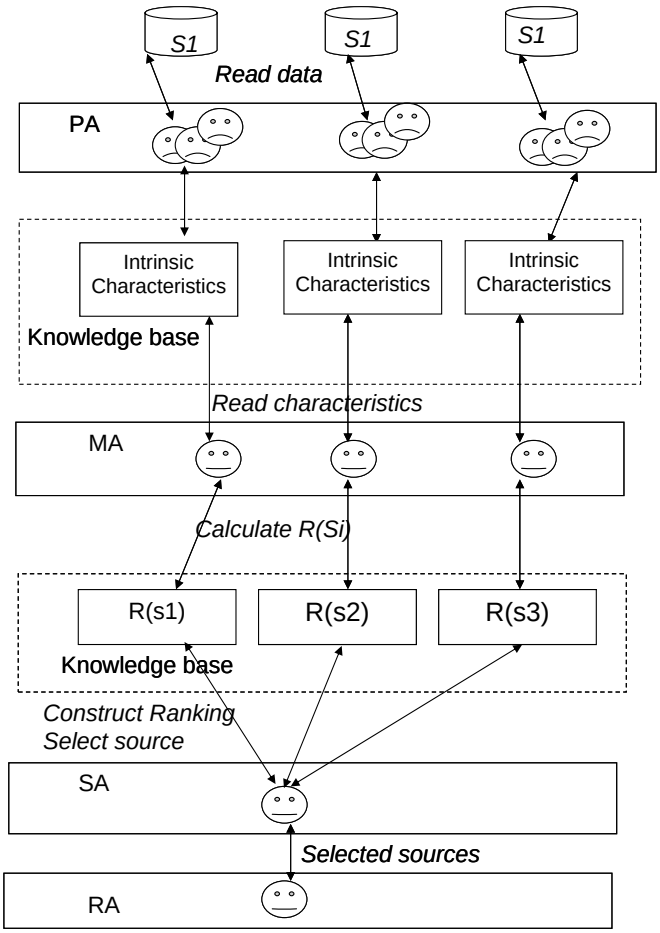


**Fig. 1.** Multi-Agent System to select relevant information sources

hypotheses, it can be said that MAS learn and adapt to their environment (preferences and tastes of the users). In general terms, planning in this kind of system can be viewed as a centralized planning for distributed plans [5].The agents do not work as a team as described in [3]. In complex societies, such as the administration of information from various sources, the amount of communication needed for agents to reach agreement on a joint decision would overcharge the system. This is the reason why the responsibility for taking decisions (selecting sources of information) is delegated to only one type of agent, the **SELECTOR AGENT**, and based on the relevance measure of the source. To obtain this objective, the other agents have to calculate the values of the characteristics allowing the relevance information provided by each source to be inferred.

## 4 Case Study

Eight data bases in the consumer package goods domain were used. These data bases contain information about clients, products and a historical record of the products purchased.

They contain real data from more than thousand clients and the purchases made by them during the period 2001-2002. There is data about clients who made purchases on the Internet (OnLine) and purchasing information about customers who personally shopped in the supermarket (Off-Line).

The different data bases have common users, who are easily identifiable because they have the same id in all the data bases.

At the beginning the agents have no knowledge of the sources. They do not know which is the most relevant to make the recommendations. The agents begin to randomly explore the information sources.

The MA from each source creates PA to obtain the measurements to calculate the relevance of each source. Table 2 shows the different values of the measures obtained by the PAs.

These measurements were obtained following the equations mentions in Table 1. To obtain the diversity measure user groups were established according to the zone in which they lived (Z), their sex (S) and the number of persons in the family (F).

The MA obtains the relevance measure R(S) for each source as can be seen in Table 3 .

With this measurement, the SA has knowledge of the relevance of each one of the sources to select the most relevant ones.

Once the recommendation is computed and the result obtained, the SA has knowledge of the relevance of the source and the result of the recommendation.

Figure 2 shows how the recommendation improvement increases as a result of integrate information only from relevant sources (S1, S2, S4, S5, S6, and S8).

In Figure 3 is shown the recommendation results from all sources (S1, S2, S3, S4, S5, S6, S7 and S8).

In this graphic can be observed that the result of recommendation decrease with the information integration from source less relevant (S3 and S7).

| | | **Sources** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** | **S8** |
| | $f_i$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | $p_8$ |
| **Characteristics** | | | | | | | | | |
| **Ontology** | | | | | | | | | |
| Relevant Attributes | 9 | 0,80 | 0,50 | 0,20 | 1,00 | 0,60 | 1,00 | 0,10 | 0,80 |
| **Diversity** | | | | | | | | | |
| Z | 3 | 0,13 | 0,11 | 0,12 | 0,14 | 0,71 | 0,24 | 0,25 | 0,23 |
| F | 3 | 0,66 | 0,67 | 0,67 | 0,73 | 0,07 | 0,56 | 0,56 | 0,49 |
| S | 3 | 0,20 | 0,20 | 0,21 | 0,11 | 0,20 | 0,19 | 0,19 | 0,26 |
| **Completeness** | 5 | 0,10 | 0,60 | 0,30 | 0,30 | 0,57 | 0,33 | 0,30 | 0,70 |
| **Frequency** | 7 | 0,23 | 0,40 | 0,25 | 0,20 | 0,50 | 0,30 | 0,20 | 0,45 |
| **Timeliness** | 5 | 0,25 | 0,40 | 0,42 | 0,15 | 0,47 | 0,35 | 0,50 | 0,30 |

**Table 2.** Intrinsic characteristics of the sources in the consumer packaged good domain

| **Sources** | S1 | S2 | **S3** | S4 | S5 | S6 | **S7** | S8 |
|---|---|---|---|---|---|---|---|---|
| **R(S)** | 0,39 | 0,44 | **0,29** | 0,45 | 0,49 | 0,50 | **0,27** | 0,52 |

**Table 3.** Relevance of the sources

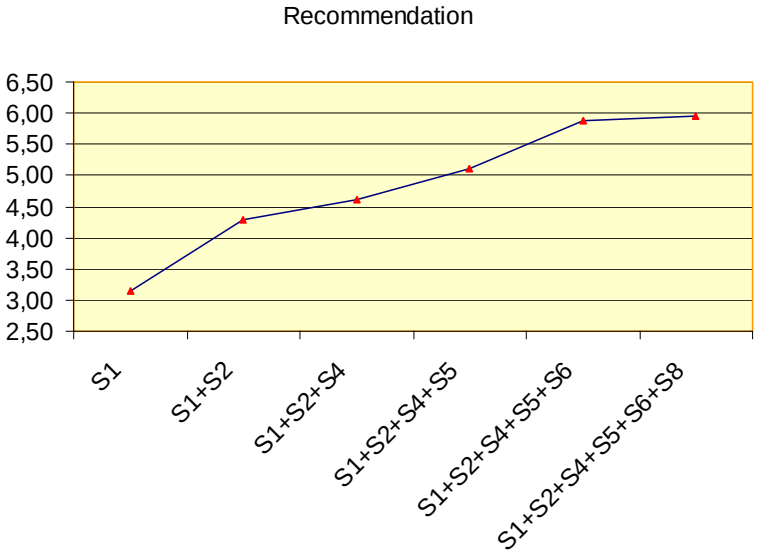


**Fig. 2.** Recommendation with the relevant sources

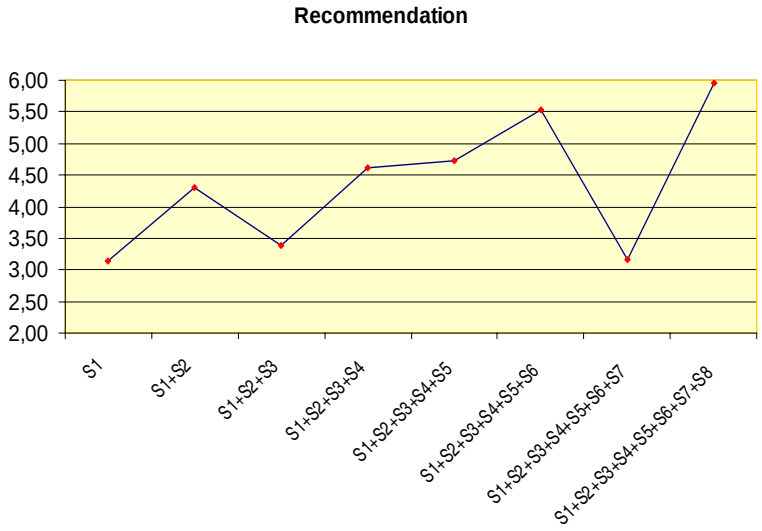


**Fig. 3.** Recommendation with all sources

## 5 Conclusion

This paper has introduced MAS to acquire and to integrate information from relevant information sources in order to improve the recommendation results. The sources of information are represented by the intrinsic characteristics. Based on them, the relevance of each source is obtained. Although very simple, this proposal produces significant results by improving the recommendation results. Future work should consider the need to compute recommendations integrating information from information sources in different domains, in which mapping between two or more ontologies of different domains will be necessary. It should also, evaluate the behaviour of the system when the information of a relevant source selected decays.

## References

1. Arens, C. Y. Chee, C-N. Hsu, and C. A. Knoblock.V: Retrieving and integrating data from multiple information sources. Inter-national Journal on Intelligent and Coop-erative Information Systems. 2(2).1993,27–158
2. M. Balabanovic and Y. Shoham.:Fab: Content-Based, Collaborative Recommendation.Communications of the ACM,40(3),6–72,March,1997
3. P. Cohen, H. Levesque, and J. H. Nunes: On Acting Together.Proceedings AAAI-90,1990.

4. Edmund H. Durfee and Jeffrey S. Rosenschein: Distributed Problem Solving and Multi-Agent Systems: Comparisons and Examples. Thirteenth International Distributed Artificial Intelligence Workshop, 1994,pages 94– 104
5. Fisher, K:Problem Solving and Planning. In Klusch, M., Fischer, K., and Luck, M.2000.Working Notes of The Second European Agent Systems Summer School, EASSS.
6. Lieberman H.: Letizia : An Agent That Assists Web Browsing. Proceedings of the International Joint Conference on Artificial Intelligence,Montreal.1995
7. A.Moreno, D. Isern: Accessing Distributed Health-Care Services Through Smart Agents.Proceedings 4th International Workshop on Enterprise Networking and Computing in Health Care Industry (HEALTHCOM'02).2002,pages 34–41
8. Moukas, A. and Maes,P: Amalthaea: An evolving multi-agent information filtering and discovering system for the WWW.Autonomous Agents and Multi-Agent Systems.1998
9. Perugini, S., Goncalves, M. A.,and Fox, E. A: Aconnection centric survey of recommender system research. Journal of Intelligent Information Systems,volume 23[1],2004
10. Salton, G: The SMART Retrieval System - experiments in automatic document processing.T. Perntice-Hall, Inc,1971. Englewood Cliffs
11. Schafer, J. B., Konstan, J. and Riedl, J.: Recommender Systems in E-Commerce. EC '99: Proceedings of the First ACM Conference on Electronic Commerce,Denver, 1999. Pages:158–166
12. Shah, T. Finin and J.Mayfield.:Information retrievel on the semantic web. 10th International Conferenc on Information and Knowledge Management. ACM Press, 2003.
13. Sycara, K., Decker, K., and Williamson, M.: Modeling information agents: advertisement, organizational roles, and dynamic behavior. Technical Report WS-9602, American Association for Artificial Intelligence, 1996.
14. F. Naumann and J. C. Freytag.:Completeness of Information Sources. Technical Report HUB-IB-135, Humboldt University of Berlin, 2000.
15. X. Jin, Y. Zhou, and B. Mobasher:Web Usage Mining Based on Probabilistic Latent Semantic Analysis. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'04),Seattle,2004.
16. R.J. Hilderman, H.J. Hamilton: Principles for mining summaries using objective measures of interestingness. ictai, p. 0072,12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00). 2000.
17. P. Cykana, A. Paul and M. Stern: DOD Guidelines on Data Quality Management. Proceedings of the Conference on Information Quality, Cambridge, MA, 154-171. 1996