Information Sources Selection Methodology for Recommender Systems Based on Intrinsic Characteristics and Trust Measure

Silvana Aciar[1], Josep Lluis de la Rosa i Esteva[1] and Josefina López Herrera[2]

Agents Research Lab, University of Girona
Campus Montilivi (Edif. PIV) 17071. Girona, Catalunya, Spain.
E-mail: {saciar, peplluis}@eia.udg.es
Departament of Llenguatges i Sistemes Informátics, Universitat Politècnica de Catalunya
E-mail: jlopez@lsi.upc.edu

Abstract

The main objective of this paper consists of improving recommendation effectiveness. An effective recommendation is one made to a user who is satisfied with the recommendation and motivated to buy the product in the future. The effectiveness of the recommendations depends on the information available for the filtering methods to predict if a product will please a user or not. In order to obtain this objective more specific objectives have been defined:

- Define a methodology to select the best structured source of information for the recommender systems. They must be defined criteria that allow sources to be qualified or evaluated to obtain the most relevant and reliable.
- Propose a mechanism to retrieve information about user preferences available on the Internet.

Introduction

The work developed in this paper presents an in-depth study and provides innovative solutions in the field of recommender systems. The methods used by these systems to carry out recommendations, such as Content-Based Filtering (CBF), Collaborative Filtering (CF) and Knowledge-Based Filtering, require information from users to predict preferences for certain products. This may be demographic information (gender, age and address), evaluations given to certain products in the past or information about their interests. There are two ways of obtaining this information: users offer it explicitly or the system can retrieve the implicit information available in the purchase or search history. For example, the movie recommender system MovieLens (http://movielens.umn.edu/login) asks users to rate at least 15 films on a scale of * to ***** (awful, ... , must be seen). The system generates recommendations based on these evaluations. When users are not registered into the site and it has no information about them, recommender systems like Amazon (http://www.amazon.com) make recommendations according to the site search history or recommend the best selling products. Nevertheless, these systems suffer from a certain lack of information (Adomavicius 2005). This problem is generally solved with the acquisition of additional information; users are asked about their interests or that information is searched for in the various sources available. The solution proposed in this paper is to look for that information in various sources, specifically those that contain implicit information about user preferences. These sources can be structured like databases with purchasing information or they can be unstructured sources like review sites where users write their experiences with, preferences for and opinions about a product they buy or possess.

We have found the fundamental problems to achieve this objective to be:

- The identification of sources of suitable information for recommender systems.
- The definition of the criteria that allow the comparison and selection of the most suitable sources.
- Retrieving information from unstructured sources.

In this sense, the proposed paper has developed:

1. A methodology that allows the identification and selection of the most suitable sources. Criteria based on the characteristics of the sources and a measure of confidence has been used to solve the problem of identifying and selecting sources.
2. A mechanism available on the Web to retrieve unstructured information from users. Text mining techniques and ontologies have been used to extract information and structure it appropriately for use by the recommenders.

These contributions allow us the achievement of two important objectives:

1. Improving the recommendations using alternative sources of information that are relevant and reliable.
2. Obtaining information found about users and implicitly available on the Internet

This document is organized as follow: Section 2 presents our contribution and the result of a study carried out to analyze the methods employed to make the recommendations and the problem and its solutions. Measures that provide information about the relevance and reliability of a source have been defined and used in a methodology to select structured sources with suitable information for the recommenders. The methodology and the experiments are presented in Section 3. Review websites can be powerful sources of information about user preferences. A mechanism to acquire this information and structure it in an appropriate way to be used by the recommender systems is presented in Section 4. This mechanism of information retrieval is especially useful for making recommendations to new users of the system who know little or nothing about it. Finally, Section 5 presents the conclusions of the paper, including a list of publications and conference contributions.

Our Contribution

All recommender systems use one or many of the recommendation methods such as Content-Based Filtering (CBF), Collaborative Filtering (CF) or Knowledge-Based Recommendation. These systems need input in the form of user information to be able to make the recommendations. In this way two key problems have been identified.

a) Lack of information: when there is not enough (sparsity) or none at all (Cold Start) user information. The second problem can be the result of three situations: the arrival of a new user, the recommendation of a new product which has not been evaluated by any users or the creation of a recommender system which depends on new users and as yet unevaluated products (Schein et al. 2002). Many solutions have been proposed for these problems, including recommendation of the most popular item, the use of social networks, the use of questionnaires, etc. In the case of a new item, it is recommended to users who liked it or bought similar items. Each system solves this problem in the most appropriate way, depending on the method used.

b) Acquisition of the preferences of unstructured sources: information from users is acquired explicitly when they themselves offer the information, or implicitly when the system monitors behaviour through either navigation records or purchase records. Completing a questionnaire or evaluating a certain quantity of products according to a scale of values is very often tedious and intrusive for the user (Adomavicius 2005). There is, however, a source that could be widely used to obtain this information, namely, web pages where users can freely post what they think about a product. Collecting this type of information is a difficult task which has not yet been solved. Part of the problem lies in the complexity of extracting information from a text. Until now only one work has been found that used this source to argue and justify the recommendations made (Ricci and Wietsma 2006)(Wietsma and Ricci 2005).

To solve these problems, we propose searching for information in sources with implicit information from users; whether in databases containing information about purchases made by them or review websites with opinions about products from different domains. The purchase databases are structured sources while the reviews constitute unstructured sources. Taking into account these clarifications, this paper proposes:

1. A methodology to select the best source of structured information, called ACQUAINT. It makes known the characteristics of the sources that offer information and their relevance to make the recommendations. As in daily life when a person is introduced to us, at first sight, and depending on the intrinsic characteristics of that person, we can know what he or she is like, even though that first impression changes over time and to the extent that we interact with him or her. This methodology will allow us to know if the sources, based on their intrinsic characteristics, are relevant for recommendations, and will also let us know if the sources are reliable based on the results obtained every time the source is used in the recommendations.
2. A mechanism to retrieve unstructured information available on the Internet. This method has been called URR (User's Reviews Retrieval). User reviews available on the Internet are a powerful source of information used by recommenders to obtain evaluations of the products and in that way solve problems related with the lack of information. The retrieval of this information implies the definition of a structure to represent the most important information from the product reviews.

The next sections will present in more detail each of the contributions made during work on this paper.

ACQUAINT Methodology

Currently there are no methods to automatically indicate which sources of information are the most appropriate for recommendations. This work proposes a methodology that measures the suitability of existing sources with regard to the necessities of a recommender system and the search for information about users. As has been seen until now in the previous section, many of the existing recommenders acquire information about users either explicitly, although it is a little tedious and requires user effort, or implicitly by monitoring their behaviour. The latter is not tedious for users, but only certain information is known: that which is provided by them in this domain. This methodology lets one finish or attempt to know users better through a search for information in other sources and other domains. This methodology has been defined specifically for sources of structured information. The steps that compose it are listed below and are shown in Figure 1.

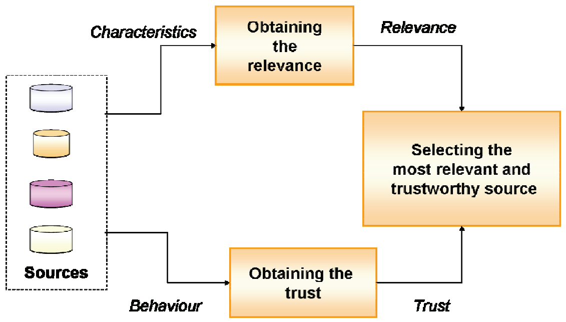

Figure 1: Methodology to select relevant and trustworthy sources for a recommender system

1. Obtain a set of characteristics representative of the information contained in the sources. These characteristics must allow the most relevant source to be compared with others before being chosen, and must be intrinsic to the sources. According to the Spanish Royal Academy, intrinsic means intimate, essential, and for that reason the characteristics of the sources must be obtained automatically from the data of the sources themselves without any human intervention.
2. Obtain a measurement to select the most reliable source. Trust in a source is obtained from the results of the recommendations made previously with this source.
3. Choose the most suitable source. An algorithm has been sued to decide, in a justified way, which sources are the most adequate for the recommendations. It uses the relevance measure from step 1 and the confidence measure from step 2.

Obtaining the relevance of the sources

Sources should be measured in some way to know if they contain or not useful information for recommender systems. Their characteristics should have certain properties such as:

- Being representative of the information needed to make the recommendations.
- Allowing the comparison and selection of sources.

The previously mentioned characteristics used to evaluate the quality of the sources are not representative of the information needed to make the recommendations. That is, they do not indicate the number of users or demographic information or information about the relevant attributes which are necessary to make the recommendations. However, they do function as criteria for buying and selecting sources, although not applicable to the recommenders. For these reasons we have redefined some of the measures mentioned previously, adapting them for the recommender systems and we have defined new measures that allow us to know if a source contains the information requirements for these systems. What follows is a list of each one of the characteristics that we will measure to know the relevance of sources to be used in the recommender systems.

Completeness: Given the set U of users of a recommendation domain, the completeness of a source S is the quantity of users of U within S, known as $|C|$, divided by the quantity of users $|U|$.

$$Completeness(S) = \frac{|C|}{|U|}$$

Diversity: The diversity of a source S is equal to the entropy H.

$$Diversity(S) = H$$

H is calculated as:

$$H = -\sum (p_i \log_2 p_i)$$

Adapted to the recommender systems each pi is calculated as follows:

$$p_i = \frac{n_i}{N}$$

Where n_i is the number of users included in the group i and N is the total quantity of users in source S. The users can be grouped according to gender, age, etc.

Frequency: The frequency of interactions of a source S is the result of the weights w_i, given for each category f_i, multiplied by $|f_i|$ which represents the quantity of users within each category, divided by the quantity of users of S which is N.

$$Frequency\ (S) = \frac{\sum w_i * |f_i|}{N}$$

Timeliness: The Timeliness of a source S is the sum of the weights w_i, given for each category p_i, multiplied by $|p_i|$, which represents the quantity of users within each category p_i, divided by the quantity of records of S, which is N.

$$Timeliness\ (S) = \frac{\sum w_i * |p_i|}{N}$$

Relevant Attributes: Given the set D of relevant attributes to make the recommendations, the quantity of relevant attributes of a source S is the quantity of attributes of D within S,$|B|$, divided by the quantity of attributes $|D|$.

$$\mathrm{Re}levant\ Attributes(S) = \frac{|B|}{|D|}$$

Finding a source that is more complete, more diverse, that has more interactions, is the most timeliness and has all the attributes necessary for the recommendations is improbable. Some sources will have better characteristics than others and there will be still others with poorer characteristics than them. For that reason, when choosing a source, each of the characteristics must be considered and have a weight assigned to it according to how important it is for making the recommendations. For example, if the source required must be timeliness but its diversity does not matter as much, the "Timeliness" characteristic will have greater weight than the "Diversity" characteristic when the source is selected.

Relevance: The relevance (R) of a source S is the sum of the values c_i of each of the characteristics j multiplied by the weight w_i assigned to each of these characteristics divided by the quantity of characteristics $|N|$.

$$R(S) = \frac{w_j * c_j}{|N|}$$

Obtaining the trust of the sources

The trust of the sources is defined as the probability with which sources are evaluated to use their information. This trust value is obtained from observations of the past behaviour of the sources. Trust mechanisms have been applied in various fields such as e-commerce (Noriega, Sierra and Rodríguez 1998), recommender systems (O'Donovan and Smyth 2005) (Massa and Avesani 2004) and social networks (Yu and Singh 2003). In our work, trust is used to evaluate the reliability of the source (S) based on the record of successful or unsuccessful recommendations made with information from that particular source, and there is a trust value for each one of the sources. The information required to compute the degree of success of the recommendations is saved. This information is then used to evaluate recommendations made with information from a source as "successful" or "not successful", indicating as the Result = 1 and Result = 0, respectively. The success of a recommendation is evaluated using one of the measures of evaluation of the recommendations (Herlocker et al. 2004). With the information about the successful recommendations, the measure of trust defined by Jigar Patel (Patel et al. 1998) is applied. They define the value of trust in the interval between [0,1], 0 meaning an unreliable source and 1a reliable source. The trust of a source S is computed as the expected value of a variable B_s given the parameters α and β . B_s is the probability that S has relevant information. This value is obtained using the next equation.

$$T(S) = E[B_s / \alpha\beta]$$

E is computed as follows:

$$E[B_s / \alpha\beta] = \frac{\alpha}{\alpha + \beta}$$

The parameters α and β are calculated as:

$$\alpha = m_s^{1:t} + 1 \qquad \beta = n_s^{1:t} + 1$$

Where m_i^{1t} is the number of successful recommendations using source S and n_i^{1t} is the number of unsuccessful recommendations and t is the time of the interaction.

Selecting the most suitable source

The most suitable and reliable sources are chosen to make the recommendations. A selection algorithm has been defined to make the choice automatically. The algorithm is made up of 3 elements:

1. A set (S) of candidate sources.
2. A selection function *Selection* to obtain the most relevant and reliable sources. This function uses the values of relevance R(s) and trust T(s) of the sources as parameters.
3. A solution set (F) containing the sources selected (F ⊂ S).

With every step the algorithm chooses a source of S, let us call it s. Next it checks if the s F can lead to a solution; if it cannot, it eliminates s from the set S, includes the source in F and goes back to choose another. If the sources run out, it has finished; if not, it continues.

Algorithm to select relevant and trustworthy source

```
Algorithm (S: Set of candidates sources)
F := ∅ ;
while (S <> ∅) do
   if Selection(R(s),T(s)) > threshold then
      F := F∪ s;
   end if
end while
return F;
```

The selection function has as parameters the relevance R(s) and trust T(s). This function returns a value between 0 and 1, and is obtained through equation:

$$selection\ (R(s), T(s)) = R(s) * T(s)$$

Experiment 1: Caprabo

This case study was conducted using information about consumers' buying behaviour from a very well known supermarket in Girona (Spain), Caprabo (http://www. caprabo.es/). This test case will help us prove the suitability and effectiveness of the characteristics defined in this work. The databases contain real information about 4137 clients, products and their purchases made in the period 2002-2003. All these purchases were made either on the Internet (online) or in the supermarket (offline). The common users of the various databases are identifiable because each one has the same user identifier in all databases. Table 1 shows the values of the characteristics of the eight sources from Caprabo.

Characteristics	S1	S2	S3	S4	S5	S6	S7	S8
Relevant attributes	0.80	0.50	0.20	1.00	0.60	1.00	0.10	0.80
Completeness	0.10	0.60	0.30	0.30	0.57	0.33	0.30	0.70
Diversity (Z)	0.13	0.11	0.12	0.14	0.71	0.24	0.25	0.23
Diversity (F)	0.33	0.67	0.67	0.73	0.07	0.56	0.56	0.49
Diversity (H)	0.20	0.20	0.21	0.11	0.20	0.19	0.19	0.26
Frequency	0.23	0.40	0.25	0.20	0.50	0.30	0.20	0.45
Timeliness	0.25	0.40	0.42	0.15	0.47	0.35	0.50	0.30

Table 1: Characteristics of the sources from Caprabo

Then the recommendations were made using only the selected sources based on their relevance and trust and were made using all information sources, Figure 2 and Figure 3 show the result obtained.

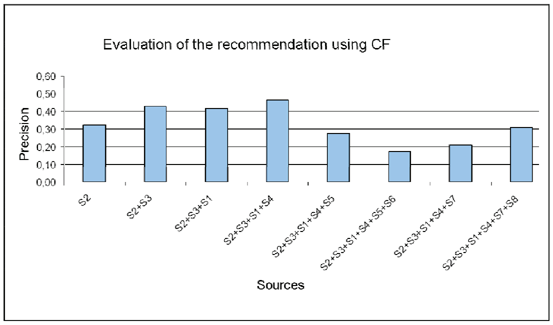

Figure 2: Precision of the recommendation including all the sources of information

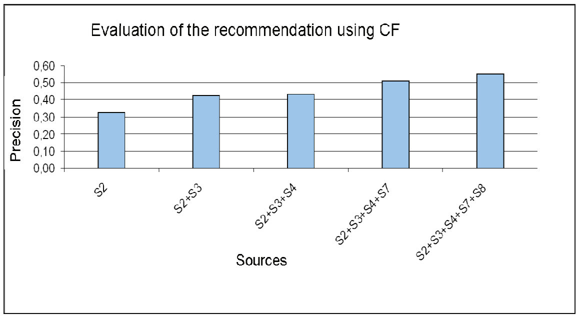

Figure 3: Precision of the recommendation including only the selected sources by ACQUAITN

Experiment 2: Amazon

The basic idea is to exploit the reviews to obtain information to build our data set to test our approach. We have retrieved reviews about CDs, DVDs, Magazines and Books composing four information sources. The information collected is resumed in Figure 4. Table 2 shows the values for each of the characteristics obtained when applying the equations defined in Section 3.1. The table also shows the relevance of each of these sources.

Data set:

Domains: Book, CD, DVD, Magazine
Source S1: Information from Books domain
Rating = 732
Users = 124
Books =699
Source S2: Information from CDs domain
Rating = 188
Users =40
CDs = 179
Source S3: Information from DVDs domain
Ratings =225
Users = 45
DVDs = 212
Source S4: Information from Magazines domain
Ratings = 72
Users = 35
Magazines =42

Figure 4: Data from Amazon.com used in the experiments

	F1	F2	F3	F4
Completeness	1,00	0,15	0,79	0,32
Diversity	0,50	0,12	0,65	0,15
Frequency	0,63	0,90	0,40	0,70
Timeliness	0,79	0,23	0,65	0,24
Relevant Attributes	1,00	0,40	0,30	0,30
R(s)	0,39	0,19	0,28	0,18

Table 2: Characteristics of the sources from Amazo.com

The graph in Figure 5 is used to compare the precision of the recommendations made with the sources selected by ACQUAINT and the precision of the recommendations made with: only information from source F1; information from all the sources; and information from the optimal combination of sources. In this case, the graph also indicates that the recommendations made with information from source F1 obtains less precise results than if the recommendations are made with a combination of information from different sources. The most precise recommendations were obtained with information from the optimal combination of sources. However, the precision obtained when making the recommendations with the selected sources based on R and T is the same as that obtained with the optimal combination of sources.

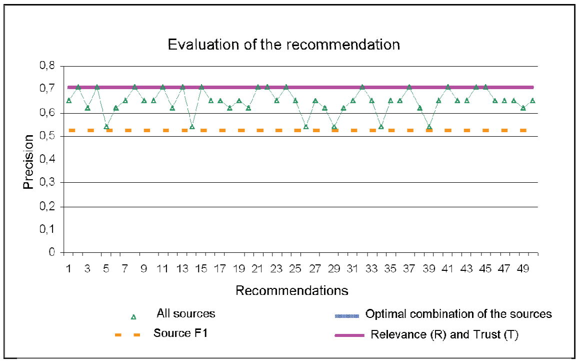

Figure 5: Result of experiments using data from Amazon.com

User's Reviews Retrieval (URR)

Obtaining evaluations from users is another of the problems recommender systems have to tackle in order to produce more effective recommendations (Adomavicius 2005). In the previous sections, we presented a methodology that enables the system to select the data sources that may have information about these users in order to improve the precision of the recommendations. This methodology has been tried out in structured data sources. However, there are internet-based sources of non-structured data that contain useful information for recommender systems. One problem when applying ACQUAINT methodology to these kinds of sources is how to structure this information. In this chapter, we present a mechanism for retrieving and structuring the user preference data that is available in web pages. Figure 6 shows the overall process structure.

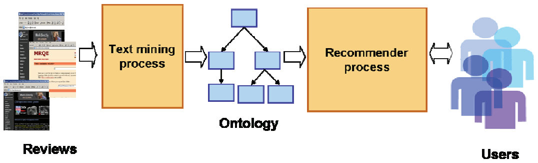

Figure 6: Process to retrieval user information from unstructured sources

In this part of the research have been make three mayor contributions:

1. An ontology to translate the information from the reviews into structured form that is suitable for processing by the recommender system.
2. An automatic ontology mapping process using text mining techniques at a sentence level.
3. A ranking mechanism for prioritizing the product quality with respect to the consumer level of expertise and the rating given to some features of the product has been developed. A set of measures such as Opinion Quality (OQ), Feature Quality (FQ), Overall Feature Quality (OFQ) and Overall Assessment (OA) have been defined to select the relevant reviews and provide the best recommendation in response to a user request.

Once the product opinions mining base is populated, we employ text mining techniques to extract useful information from review comments. In order to make reviews information useful for the recommendation process, it has to be translated into a structured form and communicated to the recommender process in a form suitable for generating recommendations. We have developed and employed an ontology to translate opinions' quality and content into a form suitable for utilisation by the recommender process. The ontology contains two main parts: Opinion Quality and Product Quality, which summarise the consumer skill level and the consumer experience with the product in the review, respectively. Figure 7 show the structure of the ontology defined to structured user's reviews.

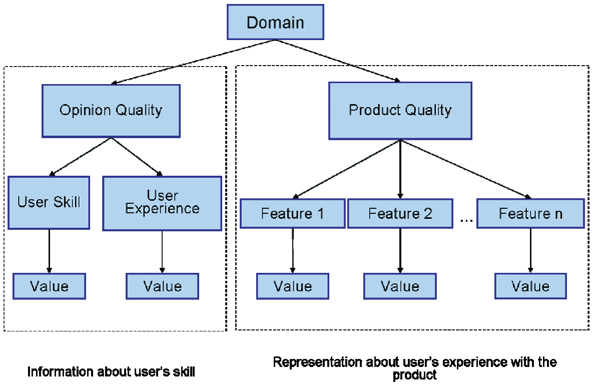

Figure 7: Structure of the Ontology used in the Recommendation from Consumer Opinions Applications

The text mining process maps the review comments into the ontology. Figure 8 shows the component of the text mining process that we have used.

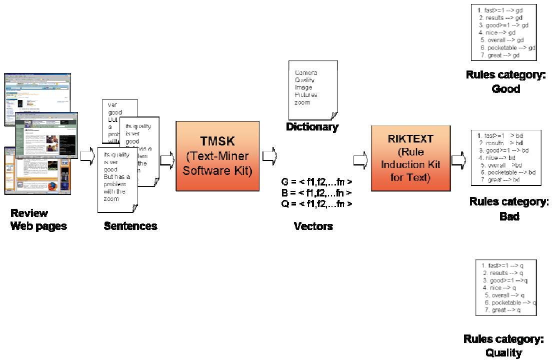

Figure 8: Inputs and Outputs for Classifier Consumer's Re-views Process

A ranking mechanism operates with over the data stored in the ontology. It prioritises that information with respect to the consumer level of expertise in using the product in consideration. The recommendation is made based on the data in the ontology. Therefore, the recommendation quality depends on the accurate mapping of the proper knowledge from the semantic features in the review comments into the ontology structure. Figure 9 and Figure 10 present the recommendations made for a user in the domain of digital cameras.

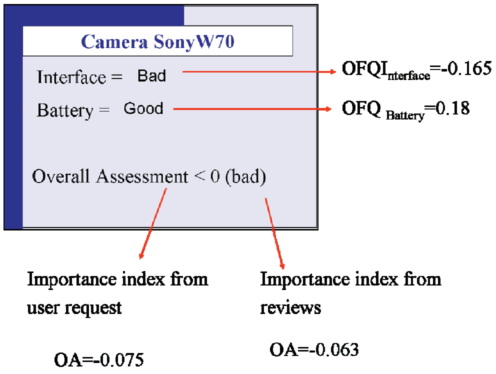

Figure 9: Final recommendation answer to the user request generated from consumer's opinions about digital cameras

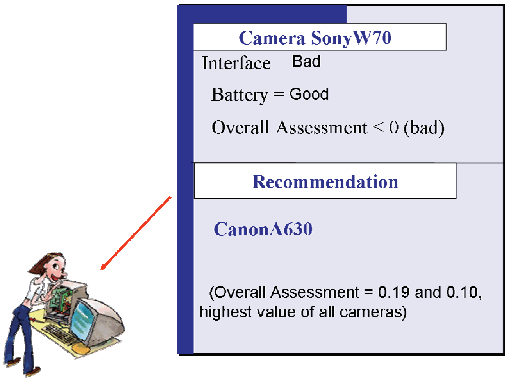
Camera SonyW70
Interface = Bad
Battery = Good
Overall Assessment < 0 (bad)
Recommendation
CanonA630
(Overall Assessment = 0.19 and 0.10, highest value of all cameras)

Figure 10: Recommendation in response for a user request from consumers' opinions

Details of the implementation of the contributions can be found in (Aciar et al. 2007). For constraint of space we only show some Figures illustrating the contribution.

Conclusions

This paper focuses on the field of recommender systems. In this field, there have been many advances in research. But despite all this research, improvements need to be made to recommender systems in order to make the recommendations more effective and applicable to real life (Adomavicius 2005). The success of a recommendation method depends on the amount of data available to make the evaluations. The lack of data gives rise to the so-called "cold start" problems when there is no user data with which to make the first recommendation and the problem of "Sparsity" when there is insufficient user preference data in relation to a product in order to make recommendations to the user.

The search for and selection of relevant and trustworthy sources that allow us to get more user preference data is one of the subjects analyzed in this paper. In Section 3, we presented the ACQUAINT methodology which uses two criteria to select the most suitable sources: the relevance (R) and trustworthiness (T) of the sources. The relevance is obtained on the basis of the intrinsic characteristics of the sources. These characteristics must be representative of the data required to find out whether or not the sources contain user data. The trustworthiness of a source is a rating that represents the degree of success of the recommendations made with data from that source in the past. Applying this methodology in two case studies showed that the set of characteristics defined is representative of the data contained in the sources. The results obtained have shown that, recommendations made with information from several data sources each selected on the basis of these two criteria (R and T) are more precise and approach optimal levels. The measurements are general and easy to apply. In addition to this methodology, we have also proposed a mechanism that enables us to retrieve and structure the internet-based user data. Once this information is structured, the ACQUAINT methodology is applied. In Section 4, we present the mechanism for retrieving and structuring this kind of information. This mechanism was implemented in the acquisition of user preferences from web pages where users can introduce opinions on particular products. Text mining techniques were used to collect this user preference data. The data was structured using ontologies. In short the contributions are:

1. A set of intrinsic characteristics has been created which indicate whether or not a source contains information that the recommenders need.
2. A measure of the relevance of the sources, based on the characteristics defined in point 1, has been defined.
3. A measurement has been applied that enables us to know how trustworthy a source may be. This measurement is calculated according to the success of previous recommendations using data from that source.
4. An algorithm has been created that the system uses to select, from a set of candidate sources, the most

relevant and trustworthy on the basis of the measures defined in the previous points

5. Ontology has been defined to structure internet-based user preference data.
6. A process has been created which, using text mining techniques, allows the system to acquire data available on webpages on the experience and evaluations of the users.

References

Aciar S., Zhang D., Simoff S. and Debenham J. *I*nformed Recommender: A Recommender System That Bases Recommendations on Consumer Product Reviews. Accepted to be published in *IEEE Intelligent Systems. Special Issue on Recommender Systems* - May/June 2007.

Adomavicius, G. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering,* Vol. 17, No. 6, pp. 734-749.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Information Systems,* Vol. 22, No 1, pp. 5-53.

Massa, P. and Avesani, P. 2004. Trust-aware collaborative filtering for recommender systems. In Proceedings of the International Conference on Cooperative Information Systems (CoopIS'04). pp. 492-508.

Noriega, P., Sierra, C., and Rodríguez, J. A. 1998. The fishmarket project. reflections on agent-mediated institutions for trustworthy e-commerce. Workshop on Agent Mediated Electronic Commerce (AMEC-98, Seoul).

O'Donovan, J. and Smyth, B. 2005. Trust in recommender systems. Proceedings of the 10th international conference on Intelligent user interfaces. pp. 167-174.

Patel, J., Teacy, W. T. L., Jennings, N. R., and Luck, M. 2005. A probabilistic trust model for handling inaccurate reputation sources. *Lecture Notes in Computer Science*. Vol. 3477/2005. pp. 193-209.

Ricci, F. and Wietsma, R. T. A. 2006. Product reviews in travel decision making. *Information and Communication Technologies in Tourism 2006*. pp. 296-307.

Schein, A., Alexandrin, P., Lyle, H., and David, M. 2002. Methods and metrics for cold-start recommendations. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 253-260.

Wietsma, R. and Ricci, F. 2005. Product reviews in mobile decision aid systems. Workshop on Pervasive Mobile Interaction Devices, in conjunction with Pervasive 2005, PERMID 2005. pp. 15-18.

Yu, B. and Singh, P. 2003. Searching social networks. Proceedings of Second International Joint Conference on Autonomous Agents and Multi-Agent Systems, pages 65-72.