Universitat de Girona

INFORMATION SOURCES SELECTION METHODOLOGY FOR RECOMMENDER SYSTEMS BASED ON INTRINSIC CHARACTERISTICS AND TRUST MEASURE

Silvana Vanesa ACIAR

ISBN: 978-84-690-8018-4
Dipòsit legal: GI-1186-2007

INFORMATION SOURCES SELECTION METHODOLOGY FOR RECOMMENDER SYSTEMS BASED ON INTRINSIC CHARACTERISTICS AND TRUST MEASURE

PhD Thesis

by

Silvana Vanesa Aciar

Supervised by
Dr. Josep Lluis de la Rosa i Esteva
and
Dra. Josefina López Herrera

May, 2007
Department of Electronics, Computer Science and Automatic Control
University of Girona
Girona, Spain

Ithaka

As you set out for Ithaka
hope the voyage is a long one,
full of adventure, full of discovery.
Laistrygonians and Cyclops,
angry Poseidon do not be afraid of them:
you'll never find things like that on your way
as long as you keep your thoughts raised high,
as long as a rare excitement
stirs your spirit and your body.
Laistrygonians and Cyclops,
wild Poseidon you won't encounter them
unless you bring them along inside your soul,
unless your soul sets them up in front of you.
Hope the voyage is a long one.
May there be many a summer morning when,
with what pleasure, what joy,
you come into harbors seen for the first time;
may you stop at Phoenician trading stations
to buy fine things,
mother of pearl and coral, amber and ebony,
sensual perfume of every kind
as many sensual perfumes as you can;
and may you visit many Egyptian cities
to gather stores of knowledge from their scholars.
Keep Ithaka always in your mind.
Arriving there is what you are destined for.
But do not hurry the journey at all.
Better if it lasts for years,
so you are old by the time you reach the island,
ealthy with all you have gained on the way,
not expecting Ithaka to make you rich.
Ithaka gave you the marvelous journey.
Without her you would not have set out.
She has nothing left to give you now.
And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
you will have understood by then what these Ithakas mean.

Konstantínos Kaváfis

To my parents, sisters and to all those
who have offered me their support in this long trip.

A mis padres, hermanas y todos
los que me apoyaron en este largo viaje.

Acknowledgements

I would like to express my gratitude for all those, without whose contribution and support I would not be able to finish my PhD. First of all, I would like to thank my both PhD advisors Josep Lluis de la Rosa and Josefina López Herrera. Without their support I would not be able to finish the PhD. Pep Lluis, thank you for your scientific advice, support, tolerance with my English and substantial suggestions during the writing period of this thesis. Josefina, thank you for your permanent and unconditional predisposition in clarifying my doubts and offering me professional and moral support, enthusiasm and valuable advices, and for your friendship inside and outside the university.

I would like to specially thank Beatriz López Ibañez for giving me a chance to begin PhD and for offering me her unconditional support and collaboration in all these years, thanks, Bea, for everything that you have done for me.

My very special gratitude to Dr. Debbie Zhang, Dr. Simeon Simoff and Dr. John Debenham from University of Technology, Sydney (UTS), for welcoming me in their research group and for their immense help in my work there. I also would like to thanks to all the people of "E-Market Club" for doing my stay in the UTS very pleasant.

I thank the Ministry of Universities, Research and Information Society of the Catalan Government for hiring me with a pre-doctoral fellowship.

I would like to thank reviewers for accepting to review this thesis and give me advices to improve it.

My gratitude also goes to the members of the ARLab group: Mercè, Javier, Edu, Gustavo, Ronald, Christian, Gabriel, Sonia, Silvia, Claudia, Salvador, Araceli, Nico y Maria for all the shared moments in the Lab and for being tolerant with my "deafness".

I would like to thank all members of EIA department, Ana, Marta, professors, PhD students, Ingrid, Tony, Tomas and Marc, each one of you have contributed in someway in the elaboration of this thesis. I also would like to thanks the people of BCDS lab with who I shared my first months in the research world.

Personally, the begining of the PhD has marked "a before" and "a later" in

my life and this change had not been possible without the unconditional help of several people. Perhaps I don't be able to give back everything that they did in that difficult time at the beginning of the 2002. Thanks Professor Alfredo Leiva; Graciela Areche; Mirian; Nancy; Abel, Sr. Franklin Sánchez; Javier Salas; Carlos Benavento; Quartet of Saxophone of UNSJ, especially to Omar Llul; professors and friends from the UNSJ; friends, family and all people from "my dear Jáchal". They did everything possible to make me to begin a new stage in my life, which changed all my expectations. Many thanks! Without your help I would not have been writing these lines.

My very special gratitude goes to Montse, Oscar, Javier, Ela, Juani, Mar, Gema and Monica. You are those, who have suffered all my moments of crisis in the last months and my "small talk". Thanks for your friendship during all these years.

I am amazed by the great atmosphere in the "UNO TEAM". Thanks Arnau, Ricard, Birgit, Javier, Arman, Pio, Marina, François, Olivier, Xavi, Carles, Josep, Robert, Thierry, Nuno, Patricia, John and Johan, their jokes and friendship have made very happy and special the lunch time every day, our Tower of Babel. Thanks to Bianca, Carles, Jordi Pagès, Montse Vila, Sonia, Isela, Wan, Sushu, Fran, Ismael and Gus by all "sopar", conversations in the corridors and moments that we shared. Thanks Toñi, Mari and Maria Angels, the mornings in the UDG had not been the same without your "good morning" and friendship.

Thanks to families, who opened the doors of their home and allowed me to share many things of each culture. Thanks to "la meva familia adotiva d'Olot", for its hospitality and to let me live and enjoy many things and parties as if I was an integrant of its family. Thanks, David and Natalia, for listening to me with patience and for all advices you gave me. Thanks, Sabina and Violeta, for all pretty moments I shared with you. Thanks, Cesar and Giovanna, for accepting me in your house without knowing and together with Mónica for making my stay in Australia a unforgettable and invaluable period. Thanks, Javier, Vicky, Prince and Fanny, for your friendship and for showing me the best things of Mexico. Thanks, Mercè and Quim, for being always disposed to respond to my numerous questions about Catalan and Girona.

Going towards the South, I would like to say thanks to the friends of my whole life: Vero, Naty, Mony, Dany, Alberto, Nancy, Abel, Claudia, Carina, Jose and

Lorena, for their friendship during all these years, for being present in "the before" with "las juntadas", hours of study and "las mateadas y asados", for being present in "the after and the present" with all goodbyes, welcomes, emails and phone calls. Thanks to my uncles, cousins and "nephews" for being always near the telephone for animating me.

Finally, I would like to express my deeper and special gratefulness to my parents Mario and Mirtha and my sisters Gabriela, Laura y Marianela. I specially dedicate this thesis to them. Thanks to give me the values that guide my walk in the life, by their friendship, unconditional support and for being there even if they were far away.

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

Abstract

The work developed in this thesis 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 (genre, 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 and search history. For example, the movie recommender system MovieLens (http://movielens.umn.edu/login) asks users to rate at least 15 movies on a scale of $\star$ to $\star\star\star\star\star$(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 make recommendations according to the site search history. Amazon.com (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 additional available sources.

The solution proposed in this thesis 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 pages where users write their experiences and opinions about a product they buy or possess.

We have found three fundamental problems to achieve this objective:

1. The identification of sources with suitable information for recommender systems.

2. The definition of criteria that allows the comparison and selection of the most suitable sources.

3. Retrieving the information from unstructured sources.

In this sense, the proposed thesis has developed:

- A methodology that allows the identification and selection of the most suitable sources. Criteria based on the characteristics of sources and a trust measure have been used to solve the problem of identifying and selecting sources.

- A mechanism to retrieve unstructured information from users available on the Web. Text mining techniques and ontologies have been used to extract information and structure it appropriately for use by the recommenders.

The contributions of the work developed in this doctoral thesis are:

1. Definition of a set of characteristics to classify relevant sources of information for recommender systems.

2. Development of a measure of relevance of sources according to characteristics defined in previous point.

3. Application of a trust measure to obtain the most reliable sources. Confidence is measured from the perspective of improving the recommendation; a reliable source is one that leads to improved recommendations.

4. Development of an algorithm to select, from a set of possible sources, the most relevant and reliable ones according to measures defined in previous points.

5. Definition of an ontology to structure information about user preferences that are available on the Internet.

6. The creation of a mapping process that automatically extracts information about user preferences available on the web and put in the ontology.

These contributions allow us the achievement of two important objectives:

- Improving recommendations using alternative sources of information that are relevant and trustworthy.
- Obtaining implicit information about user available on the Internet.

Contents

List of Figures

List of Tables

Chapter 1

Introduction, Motivation and Objectives

The search for, analysis and manipulation of information from various sources to solve problems resulting from a lack of information in recommender systems constitutes the main objective of this thesis. This chapter presents the motivation and objectives behind this research. The organization of this document is presented at the end of the chapter.

1.1 Motivation

Every day we receive suggestions and recommendations about places to go, movies to see, products to buy, etc. Recommendations have become important fixtures of our lives, and when we do not get them we ask our friends and other people whose tastes we share for them. In the digital world there are systems that automatically emulate these recommendations or suggestions, sometimes without our realizing it. It no longer surprises us to enter a website like Amazon.com (www.amazon.com) and find or see before anything else the products or types of products we wanted. Even without being registered users of the system we see a sentence that says, "The best selling books are: *The Alchemist*, *Mutant Message Down Under*". When we enter the website of the New York Times (www.nytimes.com) we see the news that interests us most. It is proven that recommendations help increase e-commerce sales [Schafer et al., 1999] and influence decisions made by users when they buy a product

[Schafer et al., 1999] [Wietsma and Ricci, 2005] [Ricci and Wietsma, 2006].

Users are overwhelmed with information about innumerable products, but recommender systems solve this problem providing them with the information and products that are most relevant to them. Recommender systems are computer programmes that, given some information about the user's profile, attempt to predict items that a user may be interested in [wikipedia.org, 2007].

There are application domains that involve the manipulation of information from various sources. For example, in the field of tourism recommendations can indicate cities to travel, places to visit, attractions to see, cultural events to attend, travel routes, hotels to stay in, tourism books and guides to take, typical local products to buy, etc. Manipulating this information is not a trivial task for recommender systems [Stabb et al., 2002]. With this great volume of options, these systems need help to manage all the available information about products or services and users and their preferences to be able to make the recommendations. This information can be found in structured sources like databases or in unstructured sources like Web pages where users introduce information about their experiences with the products.

Many researchers have extensively studied methods to filter the information or products according to user preferences such as Content-based Filtering, Collaborative Filtering, Knowledge-based Filtering or using various filtering methods together [Balabanovic and Shoham, 1997] [Abbattista et al., 2002] [Xu et al., 2005] [Burke, 2002] [Herlocker et al., 2000]. All of these methods require a minimum amount of information from users to be able to predict their preferences. Without enough information recommendations are ineffective [Adomavicius, 2005]. This problem is more serious when it concerns the first recommendation made to a user: if the first recommendations satisfy users, it is very probable that they will come back to interact with the system [Nguyen and Ricci, 2004].

Our proposal is to create a methodology that obtains user information found implicitly in various sources. This proposal analyses purchase databases (structured sources) and review pages with user opinions of products (unstructured sources).

1.2 Objectives

The main objective of this thesis 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 recommendations depends on the information available for filtering methods to predict if a product will please a user or not. In order that this objective can be achieved more specific objectives have been defined:

a) Define a methodology to select the best structured source of information for recommender systems. Criteria that allow sources to be qualified or evaluated to obtain the most relevant and trustworthy must be defined.

b) Propose a mechanism to retrieve information about user preferences available on the Internet.

1.3 Organization of the thesis

Chapter 2 presents a review of existing publications and work related with recommender systems. A study has been carried out to analyse methods employed to make recommendations and, when implementing these methods, the problems stemming from any lack of information and solutions to these problems.

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 recommenders. Two case studies have been carried out to show the utility of the methodology with respect to the improved results of the recommendations made. The methodology and case studies are presented in Chapter 3.

Review websites can be powerful sources of information about user preferences. On these sites consumers post information about their experience with products, their tastes or their interests. A mechanism to acquire this information and structure it in an appropriate way to be used by recommender systems is presented in Chapter 4. This mechanism of information retrieval is especially useful for making recommendations to new users of the system.

Finally, Chapter 5 presents the conclusions of the thesis, including a list of publications and conference contributions. Lines of future research stemming from results obtained in this research work are also discussed.

Chapter 2

State of the art

This chapter presents a general review of recommender system literature regarding problems stemming from insufficiency of user information and also explains various solutions proposed to solve this problem.

2.1 Recommender systems

Users of e-commerce sites are overwhelmed with the vast amount of information and products they are offered. Equipping e-commerce sites with personalisation tools means adapting these sites to the characteristics and preferences of each user [Billsus and Pazzani, 2000] [Billsus et al., 2002] [Boll, 2002] [Wu et al., 2003]. When users have been recognised, the content and presentation of the products is decided according to their preferences. Personalisation increases the usability of these sites by making the interactions easier and faster and increasing their reliability [Kibum et al., 2002] [Tam and Ho, 2003] [Babaguchi et al., 2003] [Evans et al., 2006] [Cosley et al., 2003]. Haul [Haübl and Trifts, 2000] show that interactive decision tools designed to assist users in a purchased decision may have highly desirable properties over it. Recommender systems are tools that make personalised recommendations to users or groups of users in e-commerce sites [Ansari et al., 2000] [Marlin, 2003]. A recommender system is a tool that receives as input information about users and based on this information recommends products that would please users most [Billsus and Pazzani, 1998] [Schafer et al., 1999] (See Figure 2.1).

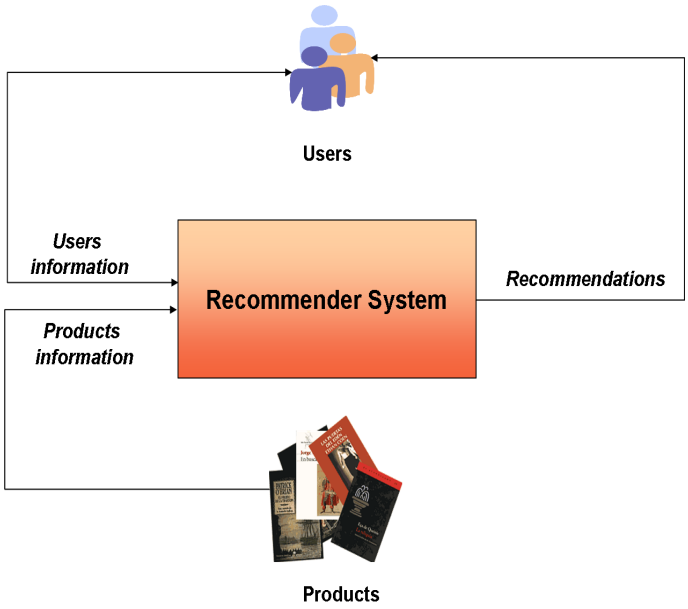


Figure 2.1: Input and outputs in a recommender system

This process involves three types of basic actions that must be taken:

- Obtaining information from users about their preferences and interests.
- Filtering the information or products that are most relevant for the users according to their preferences.
- Integrating information from various databases; for example, recommendations in tourism involve packaging products from different domains.

2.1.1 Obtaining information about users

There are two types of information about users in a recommender system: *Explicit Information* and *Implicit Information*. Many recommender systems use the evaluations users have made of certain products, collected after asking users to rate products using a scale [Yang et al., 2004] [Wietsma and Ricci, 2005] [MovieLens, 2006]. The users introduce their preferences explicitly [Towle and Quinn, 2000]. This way of collecting information requires a certain effort on the part of the users, for example, spending time entering their preferences [Clerkin et al., 2002] and it has

been proven that they do not necessarily provide reliable estimations of user preferences [Farzan and Brusilovsky, 2006]. Another source of information contains preferences that users have implicitly entered into the system. The users are not aware of how the recommender functions; the system monitors their behaviour and automatically infers the preferences [Oard and Kim, 1998] [Shahabi and Chen, 2003] [Middleton et al., 2002] [Middleton et al., 2004] [Farzan and Brusilovsky, 2006]. This can be done analysing the URLs they have accessed [Cho et al., 2002] [Zaýane et al., 2003] or analysing the purchases they have made. Amazon [Amazon, 2006] applies these two techniques to make recommendations.

Many people prefer to express their opinions freely about products, introducing text onto web pages, as they would do in a conversation with a friend about their experience with a product. The digital world offers various popular ways for user to exchange such information [Dellarocas, 2003] [Curien et al., 2006], such as product discussion forums and sites, blog communities, etc. There is growing evidence that these sources inform about and influence customer purchase decisions [Senecal and Nantel, 2004] [Chevalier and Mayzlin, 2003]. This type of information is difficult to collect and the problem remains unresolved. Part of its complexity stems form extracting information from the text and converting it in such a way that it can be used by recommender systems.

Independently of the method used to acquire user evaluations of products (explicit or implicit), recommender systems need a sufficient amount of information to make more precise recommendations [Calypool et al., 2001]. A *chronic lack of information* [Adomavicius, 2005] is one of the key problems in these systems and insufficient information is a problem in the two situations mentioned below.

- Sparsity of evaluations. This problem arises when the number of evaluations obtained is very small compared to the quantity needed to make the predictions for the recommendations [Adomavicius, 2005]. Bad recommendations can be caused by problems related to the density of user evaluations [Carenini et al., 2003].

- Cold start, or lack of initial information. This problem can be caused by three situations: when a new user who has not yet evaluated any products enters the system, when the system wants to recommend a new product that has not yet been evaluated by any users or when a new recommender system that has not

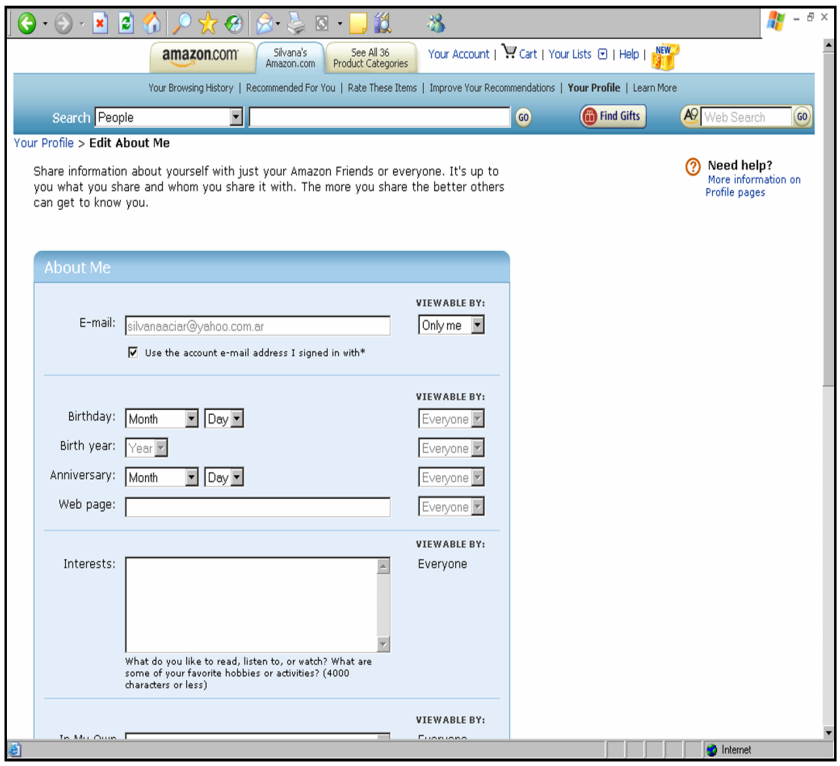


Figure 2.2: Interface used by Amazon to collect user's preferences

yet obtained any user evaluations of products is initiated [Schein et al., 2002].

Solutions to the lack of information problem

Several ways of solving the information scarsity problem behind the poor functioning of the recommenders can be found in the literature. If users are satisfied, especially after the first recommendation is made, it is very likely they will return to interact with the system again [Nguyen and Ricci, 2004]. The traditional way of acquiring these evaluations is through explicit input provided by users in graphic interfaces. Recommenders using this method include Entree [Burke, 2000], NutKing [Ricci et al., 2002], Amazon [Amazon, 2006] and Movielens [MovieLens, 2006]. Figure 2.2 shows the interface used by Amazon for users to enter their interests explicitly.

In [Nguyen and Ricci, 2004] the initial interests of users are an aggregation of historical and current information. This information is obtained from various sources,

such as:

- Information from the context, specifically the location and time of access to the system.
- Information about the preferences that users provide explicitly to the system.
- Information about the behaviour of similar users.

The first two provide current information while the last one offers information obtained from the past behaviour of similar users. Using these three sources of information, initial user preferences for air routes are obtained by the recommender system.

In [Carenini et al., 2003] situations are identified in which users are motivated to "converse" with the system. The objective is to obtain more explicit evaluations and minimise the number of questions posed to users and the time and effort they dedicate to this task. People are motivated to provide information if they perceive that it contributes to obtaining their objectives or those of their friends, if it contributes to solving an error that they discovered and if it helps other people from whom they could obtain things in the future.

In [Rashid et al., 2002] five ways of solving the problem of too little information from users to recommend a product are proposed:

1. Recommend products chosen randomly with uniform probability from the set of products available.
2. Recommend the products that are most popular in the user community. When there is no initial user information, present them the product that other users have evaluated most.
3. Use entropy to measure the diversity of the product evaluations. A product that has been given a variety of evaluations will provide more information than a product evaluated by all with statements like "*I like it*" or "*I don't like it*". Using this measurement the products are presented in descending order of entropy.

Strategy	**User effort**	**Precision**
Random	⋆	⋆⋆
Popularity	⋆⋆⋆⋆⋆	⋆⋆⋆⋆
PopEnt	⋆⋆⋆	⋆⋆⋆⋆⋆
Item-Item	⋆⋆⋆⋆⋆	⋆⋆

Table 2.1: Evaluation of the strategy to recommend a new item [Schein et.al., 2002]

4. Use "PopEnt", a measurement that combines a product's entropy measurement with its popularity measurement, calculated based on the logarithm of the popularity of the product multiplied by its entropy (log Popularity ∗ Entropy) and using the benefits of both forms, popularity and entropy. Products are recommended to users in descending order of this measurement.

5. Recommend products that are similar to those that users have evaluated previously. Use of this strategy requires that users have evaluated a certain number of products.

These strategies have been evaluated from the point of view of the effort required by users to carry them out and with regard to the precision of the recommendations [Schein et al., 2002]. The random strategy requires less effort from users, while the strategy of recommending the most popular product and that of recommending the product to users who have evaluated similar products saying "*I like them*" require greater user effort. Taking into account the precision of the recommendations made using one of these strategies means that the most precise recommendations are obtained recommending the products recommended through the "PopEnt" strategy ((logPopularity) * Entropy) and the least precise recommendations were made with the random item recommendation strategy to users who had evaluated products similar to the product in question. These results are summarised in Table 2.1. Recommendations combining collaborative filtering and content-based filtering have been proposed to solve the problem of new users [Burke, 2002] [Semeraro et al., 2005]. Using only collaborative filtering, new users who have provided few evaluations of products will not receive very precise recommendations because of the difficulty of finding other similar users. Combining this method with the content-based filtering method allows a comparison of similarities between their profiles and other profiles in order to find sets of users similar to them.

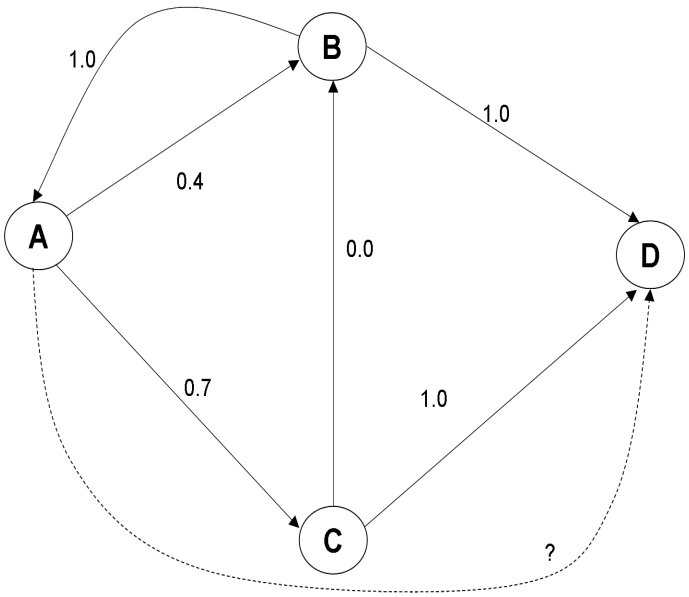


Figure 2.3: Example of web of trust [Massa and Avesani, 2004]

In [Massa and Avesani, 2004] it is argued that the scarsity of information problems can be solved incorporating the "*Confidence*" measure among users into Collaborative Filtering (CF). The community of users is modelled as a social network called "Web of Trust" where the relations between the network nodes (users) is a trust value. For example, looking at Figure 2.3, user C has a lot of confidence in the evaluations made by user D and believes in his opinion of certain products, but does not have confidence in the evaluations made by user B, and to some extent it is possible to know if user A trusts user D by means of the confidence user A has in users B and C, who are in the trust network of A. Predicting trust values based on their propagation through the network finds more similar users than with methods used until now in CF. This proposal solves the problems stemming from a lack of information because at some level of the trust propagation network it is possible to find a reliable friend whose evaluations can be obtained.

2.1.2 Filtering relevant information, products or services

To carry out recommendations the systems use one of the following methods.

- **Content-based methods (CBF)**: Such recommender systems uniquely char-

acterize each user without having to match her/his interests to other users [Balabanovic and Shoham, 1997] [Adomavicius, 2005] [Xu et al., 2005]. They can provide a list of content features that explain why an item has been recommended. Such a list can strengthen user confidence in the recommendation and provide reflection of user own preferences. For example one of the system that applies CBF is LIBRA [Mooney and Roy, 2000] that uses naive Bayes text classifier to make content-based book recommendations exploiting the product descriptions found in Amazon.com.

- **Collaborative filtering(CF)** : In the literature are several approaches that use this method to make recommendations about the preferences of a user based on collected taste information of other users [Yang et al., 2004] [Herlocker et al., 2000] [Linden et al., 2003] [Melville et al., 2002].

  The underlying assumption of CF methods is that those who agreed in the past also tend to agree again in the future [Billsus and Pazzani, 1998]. In other words, such recommender system emulates the behaviour of a user recommending a product to her friend, because some other users that she knows and believes that have similar tastes to her friend, like the product. Technically such recommender system operates similar to a case-based reasoning system, without the adaptation step. It maintains a case base of the preferences of individual users, for a given patron finds other users whose known preferences correlate significantly with the patron, and recommends to a person other items enjoyed by her/his matched patrons. The system can provide the list of some of these patrons relating them to their other purchases in order to provide user with some explanation and confidence in the recommendation. Another view of CF is the Item-Item CF. For a particular item, find other similar items. In other words is estimating the rating for the item based on ratings for similar items. Can use same similarity metrics and prediction functions as in user-user model. In practice, it has been observed that item-item often works better than user-user CF [Linden et al., 2003][Deshpande and Karypis, 2004].

- **Knowledge-base recommender** this method asks a user about the requirement of wanted products and reasons about what products meet the user's requirements based on the answers. Infer a match between the items and the user's needs [Burke, 2000][Felfernig et al., 2007]. Knowledge based recommender do not need an initial database of users' preference or data about

particular rated items. It has the product domain knowledge and the knowledge should be stored and organized in a inferable way.

- **Hybrid recommender** [Burke, 2002] combine two o more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one. Most commonly, CF is combined with some other technique in an attempt to avoid the ramp-up problem. For example PTV system [Smyth and Cotter, 2000] uses this approach to assemble a recommended program of television viewing. It uses content-based techniques based on textual descriptions of TV shows and collaborative information about the preferences of other users. Recommendations from both techniques are combined together in the final suggested program.

2.1.3 Integrating information from various sources

Information integration has become very important in many fields due to the accessibility of a greater and greater number of information sources. For instance in Customer Relationship Management (CRM), integrated information on individual customers, business environment trends, and current sales can be used to improve customer services. In the area of e-commerce and e-business, integrated information enables and facilitates business transactions and services over computer networks. In recommender systems exists application domains, such as tourism, where is necessary to access source from different domains. These sources are semantical and structural heterogenous. Information integration systems [Calvanese et al., 1998] [Levy, 2000] [Castillo, 2002] attempt to provide users with flexible access to information from multiple autonomous, distributed and heterogeneous data sources through a unified interface without worrying about the underlying syntactical details.

There are two main types of issues in information integration [Koeller, 2001]:

- **Physical integration:** includes the study of different network protocols and access to different networks.

- **Logical integration:** includes the integration of the information between the heterogenous systems. It can be descomposed in two subtypes: Schema Integration and Data Integration.

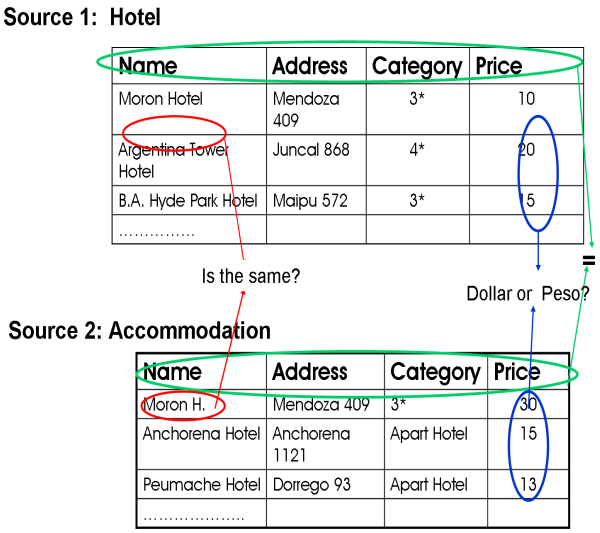


Figure 2.4: Heterogeneity problems in two information sources

Example 1:

Figure 2.4 shows two sources used in a tourism recommender system. It search for information about hotels in Argentina. Source 1 produces hotel information, Source 2 produces accommodation information. Each of these sources contain information about hotels in Buenos Aires. We can see that the "Moron Hotel" and "Moron H." can be the same hotel because has the same address, but the label is different and the price is different, we can assume that the currency is different. In Source 1 the price can be expressed in pesos argentinos and in Source 2 in dollars.

In order to be able to use these different sources of information, the meaning of the information exchanged has to be understood throughout the system. Semantic conflicts occur in any context where the same interpretation of information is not used. Goh [Goh, 1997] identifies three main causes of semantic heterogeneity.

- Confounding conflicts occur when information items seem to have the same meaning, but differ in reality, e.g. due to different temporal contexts.

- Scaling conflicts occur when different reference systems are used to measure a value. Examples are different currencies or marks.
- Naming conflicts occurs when naming schemes of information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms.

All of them are the reasons that cause the next specific heterogeneity problems:

- System heterogeneity problem: includes incompatible hardware and operating systems.
- Syntactic heterogeneity problem: refers to different languages and data representations.
- Structural heterogeneity problem: includes different data models.
- Semantic heterogeneity problem: refers to the meaning of terms using in the interchange, different terms can be used to refer to the same concept.

Many technologies have been appearing to deal with these problems. The first three categories have been addressed using technologies such as CORBA and DCOM (Distributed Component Object Model). Recently XML has gained acceptance as a way of providing a common syntax for exchanging heterogenous information. A number of schema-level specifications (usually as a DTD (Document Type Definition) or an XML Schema)have recently been proposed as standards for their usage in e-commerce. However, these standards do not solve the problems of semantic heterogeneity. A solution to the problems of semantic heterogeneity should equip heterogenous and autonomous software systems with the ability to share and exchange information in a semantically consistent way. This can be achieve in many ways, each of which might be the most appropriate given some set of circumstances. One solution is for developers to write code which translates between the terminologies of pairs of systems. This may be a useful solution where the requirement is for a small number of systems to interoperate. These codes are denominated wrappers. However, this solution does not scale as the development costs increase as more systems are added and the degree of semantic heterogeneity increases.

In the literature there are many applications that uses these approaches to integrate information from differente sources such as:

- **Phoebus:** Phoebus search and query unstructured information sources such as auction site listings. The system can even support aggregate queries over the data sources, which is difficult with keyword search. This allows data to be sorted, searched and linked to other data sources where standard values for the attributes are required to link the sources together [Michelson and Knoblock, 2006].

- **Prometheus:** provides uniform access to the sources [Michalowski et al., 2004]. Prometheus provides an infrastructure that can be used to:

  - Quickly build applications that integrate data from various data sources, and
  - Can be used as a test bed for information integration researchers to build and test new information integration techniques.

- **Information Manifold:** it has been a pioneering project in combining database approach and Artificial Intelligence approach in information integration. This project allow users specify the description and any constraint on each sources, this allow users to integrate the sources in the fastest possible way. The user queries the system based on a mediated schema. The query is them broken down into sources using a Language called "Carin" which combines the expressive power of Datalog a database query language and Description Logic [Kirk et al., 1995].

- **Ariadne:** deals with the integration of web sources. Use Query Centric Approach and also rely on source description [Knoblock et al., 2001]. Use mapping table to align the Ontology across semantically disparate sources. Ariadne breaks down query processing into 2 phases:

  - Phase 1: Query Preprocessing where it goes through finding ways to combine possible sources for answering the query.
  - Phase 2: It takes a sub optimal plan and tries to improve it by applying several rewriting rules.

- **IBM-Websphere:** WebSphere helped define the middleware software category and is designed to set up, operate and integrate e-business applications across multiple computing platforms using Web technologies. It solves typical

pains associated with business integration, such as integrating customer data, speeding the development of portal applications and aggregating information across and beyond the enterprise [Group, 2006].

- **ISU-INDUS:** INDUS (Intelligent Data Understanding System) is a federated, query-centric system for knowledge acquisition from distributed semantically heterogenous data sources that employs ontologies (controlled vocabularies of domain specific terms, and relationships among terms) and inter-ontologies mappings, to enable a user to view a collection of such data sources (regardless of location, internal structure and query interfaces) as though they were a collection of tables structured according to an ontology supplied by the user [Castillo et al., 2003].

In short all of these approaches can provide a uniform query interface for user to query multiples and heterogenous information sources. These systems model data sources in the form of relations. These systems also contain a set of virtual domain relations that the user utilizes to specify the queries to the mediator system.

Ontologies to resolve structural and semantic heterogeneity problems

In any realistic scenario involving interoperability between systems, semantic heterogeneity is a significant problem. Ontologies have emerged as a solution of this problem. It is typically used as a form of knowledge representation and sharing. An ontology is a collection of concepts and their relationships that can collectively provide an abstract view of an application domain [Gruber, 1993] [Guarino et al., 1999]. Ontologies have been well studied in many aspect such as:

- **Ontologies uses:** which include ontology architectures and applications [Ciocoiu et al., 2001].
- **Ontology mappings:** matching of the information from the resources into the ontologies and the matching of different ontologies [Noy, 2004] [Ehrig and Sure, 2004].
- **Ontology representation:** different capacity of representation, languages and tools [Ilebrekke, 2002].

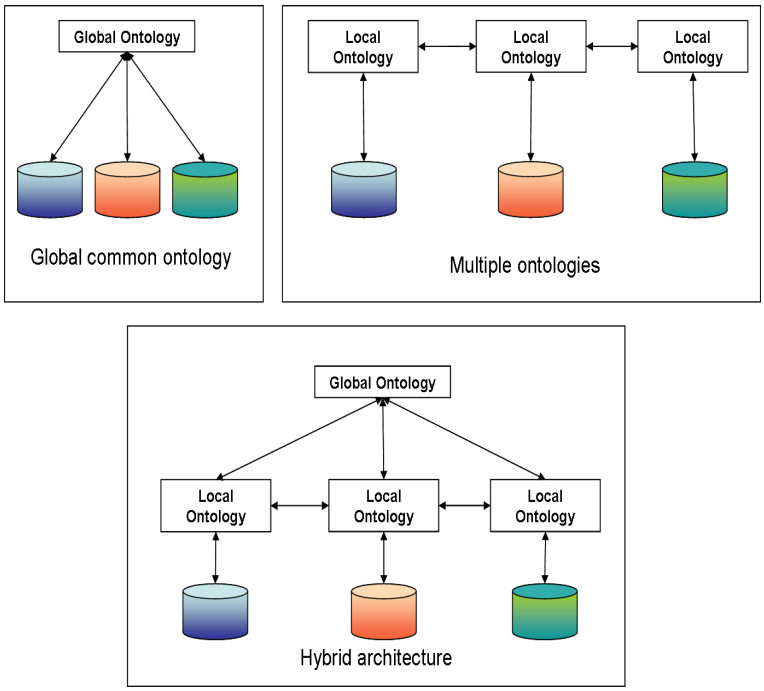


Figure 2.5: Ontology architectures to integrate the information from different sources

- **Ontology engineering:** acquisition, support and reusability of the ontologies [Jarrar, 2005].

We will focus on mapping ontologies, which will allow us to obtain semantic interoperability between heterogeneous sources of information. One solution to the problem of semantic heterogeneity using ontologies will need to accommodate various ontologies from different domains. Figure 2.5 shows three types of ontology-driven architectures used to integrate information.

- **Global common ontology:** in this case is necessary the existence of a minim common vocabulary between the sources to integrate, in many case this is hard to found. Specially between the sources from different domains.
- **Multiple ontologies:** each source have their own ontology. In this case a mapping between the ontologies is make establishing the relationship between them. It is not necessary a minim common ontology.
- **Hybrid architecture:** each source have their own ontology and also, a global common ontology is defined. In this case, the advantage of both previous

architectures are combined.

As can be seen in the figure, there are three basic actions that must be carried out:

1. **Ontology definition**: Creation of an ontology describing the information contained in the information sources. *Protegè* − 2000, OntoEdit, OilED, WebODE and Ontoligua are tools that allow ontologies to be edited [Noy and M.A.Musen, 2002].

2. **Source-Ontology mapping**: Once the ontology has been defined, the information from the sources must be put inside the ontology. This process can be carried out manually, copying the information about each concept or automatically searching in the domain for information necessary to complete the ontology like FCA-Merge does [Kalfoglou and Schorlemmer, 2005].

3. **Ontology-Ontology mapping** : Mapping two ontologies consists of finding, for each concept of the ontology, a corresponding concept in another ontology which means the same semantically [Ehrig and Sure, 2004]. Automatic mapping of two ontologies is the desire of all systems but until now it has not been entirely possible. Acquiring complete meaning about the real world is a difficult task. As a result, human intervention is necessary in the process of identifying the correspondences between the various ontologies. The tools developed to carry out mapping in a semi-automatic way are: PROMPT, MAFRA, ONION, Chimaera, FCA-Merge, GLUE and Observer [Kalfoglou and Schorlemmer, 2005].

2.2 Our contribution

All recommender systems use one or many of the recommendation methods described in this chapter, for example content-based recommendation, collaborative recommendation 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). 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 preferences from 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 of 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 recommendations that have been 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 reviews constitute unstructured sources. Taking into account these clarifications, this thesis proposes:

1. **A methodology to select the best source of structured information, called *ACQUAINT*.** It makes known characteristics of sources that offer information and their relevance to make recommendations. As in daily life when a person is introduced to us, at first sight, and depending on 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 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 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 product reviews.

Both contribution made during the elaboration of this doctoral thesis will be present in next chapters.

Chapter 3

ACQUAINT Methodology

In this chapter we present a new solution to the problem of obtaining information about users from other sources of information to have more knowledge about them and improve recommendations to them. This methodology has been applied to two case studies using real data. The results discussed at the end of the chapter show that recommendations are more effective with this methodology.

3.1 Introduction

According to Vézina and Militaru [Vézina and Militaru, 2003], in a universe, where there is a lot of imperfect information and a large supply available for users, it is extremely difficult to identify their own needs and preferences and the way to satisfy them. However, the emergence of new intermediaries on the Internet has facilitated the interaction between supply and demand. Actually, the main mission of such intermediaries is to ease transactions by facilitating the collection, organisation, and evaluation of dispersed information. In this context, the functions of a recommender system appear crucial to these activities.

Recommender systems [Adomavicius, 2005] are an example of the tools used by many e-commerce sites to help users decide which product to buy [Rashid et al., 2002] [Wietsma and Ricci, 2005] [Ricci and Wietsma, 2006]. Many users have limited patience for locating what they need when there is a lot of information and no effective guidance is provided to look for it [Palmer, 2002]. The recommender system is like a sales clerk in the corner store who, when lifetime clients enter, surprises them with

new products he knows they will like. On the other hand, if the clients are new, he begins to ask them questions to know what products they want. Like a seller, these systems model the behaviour of the users based on redundant information with the objective to develop, improve and retain relations with the user and to offer them customised products. One of the most important challenges both in marketing and recommender systems is to find out useful information about potential customers or users of a certain product or service. Usually, the first step while looking for information is to scan any available source, either internal or external to the firm. However, the main problem that comes up is the usefulness of such information. Thus, a priori, we need to develop a mechanism that would provide evidence on the relevance of such information for recommendations. To avoid making an endless search in the huge pile of available data, which would employ too many resources, the information should be previously classified or indexed to be easily found.

Currently there are no methods to automatically indicate which sources of information are the most appropriate for recommendations. This paper 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. This methodology obtains information about users from various sources where information about their interactions has been stored. As has been seen until now in the chapter on the state of the art, 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. Despite the interaction, the time comes when new information about users or preferences about other aspects are not known. This methodology lets one finish or attempt to know users better through a search for information in other sources.

3.2 ACQUAINT Methodology

This methodology has been defined specifically for sources of structured information. The steps that comprise it are listed below.

1. Obtain a set of characteristics representative of the information contained in the sources. These characteristics must allow the most relevant source

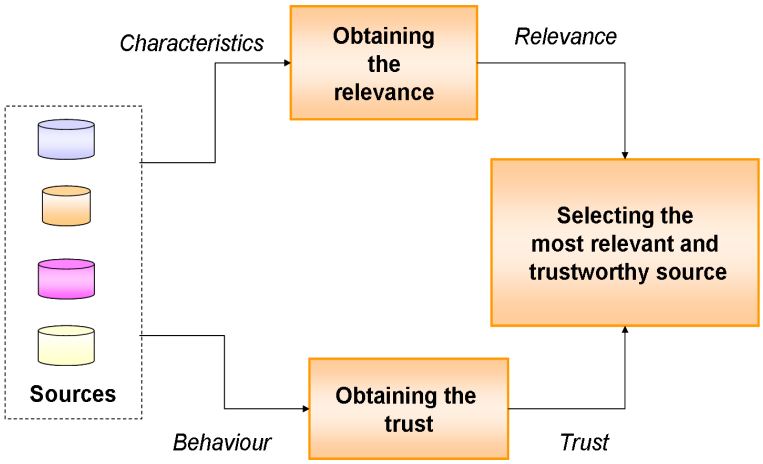


Figure 3.1: Methodology to select relevant and trustworthy sources for a recommender system

to be compared with others before being chosen, and must be **intrinsic** to the sources. According to the Dictionary of the Spanish Royal Academy [Española, 2006], 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 of a source is obtained from the results of the recommendations made previously with this source.

3. Choose the most suitable source. A measure of suitability is obtained for each of the available sources. It serves to decide, in a justified way, which sources are the most adequate for the recommendations. The suitability measure uses the relevance measure from step 1 and the trust measure from step 2.

The steps of the methodology are illustrated in Figure 3.1. In the following sections each one is explained in greater detail.

3.2.1 Obtaining the set of source characteristics

An information source can simply be defined as a repository where information or knowledge is stored. There are costs associated with the use of information sources

those costs are an important determinant in the choice of which information sources are finally used. For example the accuracy of a marketing campaign could be increased using information from the appropriate information sources. The increase of the accuracy produces more sales that result in more profits for a company. Recommendation can be better if the recommender knows where is the suitable information to predict user's preferences to offer products. Recommender systems are similar to data base marketing, but also different from it. Marketing systems support the make decisions about how to market products to consumers, usually by grouping the consumers according to marketing segments and grouping the products in categories that can be aligned with the marketing segments. By contrast, recommender systems directly interact with consumers, helping them find products they will like to purchase. But both are similar on the basis that same information can be used, a recommender system offers a product for each consumer and in the opposite way database marketing systems try to search for possible consumers for a product. Both could use the same information sources. These sources contain specially information about users and products Carenini [Carenini et al., 2003] identifies three types of information used by recommender systems:

a) Demographic information about the users (age, genre, place where they live, etc.)

b) User preferences for product characteristics (type of movie, actor, director, etc.)

c) Previous experience (evaluations of purchased products)

Based on this information the affinity that the user has is calculated for some products and the products with most affinity are recommended for the user.

Sources that provide information that is timely, accurate and relevant are expected to be used more often than sources that provide irrelevant information. It is an intuitive idea but how to know which of the sources have to be selected to obtain better results? Factors indicating if the sources can be used for some purpose have to be created.

A review of the literature concerning the evaluation of information sources to obtain a set of features that evaluate the sources with respect to their suitability for recommendations leads to research in which the analysed sources are web

pages. Judith Edwards [Edwards, 1998] presents three aspects on the evaluation of Web resources such as: Access, Quality and Easy to Use. Access includes the reliability of the site and the hardware and software specification that have been used to access the site. Quality includes both the authority of the site, for example the reputation of the organization, and the quality of the content. Easy to use is determinate by the difficulty to navigate a Web site. Boklaschuk and Caisse [Boklaschuk and Caisse, 2001] they define nine criteria to evaluate educational web sources. These criteria are: Audience, Credibility, Accuracy, Objectivity, Coverage, Currency, Accessibility, Aesthetic and Navigation. Rieh [Rieh, 2002] defines general characteristics such as reputation and credibility to classify these sources. Among these studies, we highlight the one conducted by Naumann [Naumann et al., 2004] that defines the completeness measurement of the sources to determine which ones are most appropriate to answer queries.

As defined by the authors, these features cannot be used for recommender systems. The sources of information used in the systems are those that allow information to be obtained from users to offer them products or services that interest them. This information includes demographic information, product evaluations, products they have purchased in the past, information about the context or information about similar users [Pazzani, 1999] [Carenini et al., 2003] [Endo and Noto, 2003] [Adomavicius, 2005] and is generally contained in a database. Preferences are obtained from this information. Taking into consideration which information is necessary, to supply recommenders sources must:

- Contain information about specific users the system needs to know to make recommendations. This means knowing if the source contains information about, for example, users A, B, C and D. If the source contains information about all those users (A, B, C and D), it will be more complete than a source only containing information about users A and C.

- Have demographic information about the users, which is a type of information needed to make recommendations [Burke, 2000].

- Contain a lot of information because while more information there is about the behaviour, purchases or evaluations of users, more precise will be recommendations [Adomavicius, 2005].

- Have updated information. Tastes and preferences of users can change over time. Movies they liked ten years ago might not be the same as the ones they like now. But movies they liked two months ago can be, with more probability, the same ones they like now.

- Contain attributes needed to make recommendations. For example, given the request "Recommend for each user from Paris the five best-rated restaurants that have a menu price less than 30 euros per person", in response, searches must be performed in databases that contain the attributes: city where the user lives, the menu prices of restaurants and the evaluation given to them by other users.

Example 1:

Let's consider a set of sources used to look for information to recommend a perfume to women. Those sources that have the most **complete** and **updated** information about women (**demographic information**) and their street and e-mail addresses (**relevant attributes**) are able to send information about the new product and could be required to recommend the perfume.

Sources should be measured in some way to know if they meet these requirements. Their features should have certain properties such as:

- Being representative of the information needed to make recommendations.

- Allowing the comparison and selection of sources.

The previously mentioned features used to evaluate the quality of the sources are not representative of the information needed to make recommendations. That is, they do not indicate the number of users, demographic information or information about relevant attributes which are necessary to make recommendations. However, they do function as criteria for buying and selecting sources, although not applicable to recommenders. For these reasons we have redefined some of the measures mentioned previously, adapting them for recommender systems and we have defined

new measures that allow us to know if a source meets the information requirements for these systems.

What follows is a list of each one of the characteristics that a source must have to be used in recommender systems.

- **Completeness**. The objective of this work is to obtain information about users in a range of available sources related to them. While greater number of users found in the source, more complete it will be.

- **Diversity**. This characteristic allows the demographic information contained in the sources to be known. Recommender systems use this information to group users according to their genre, place where they live, their age, etc, and in that way are able to make recommendations to entire groups of users. While more diverse the source is, more user groups it will have.

- **Frequency**. This characteristic aims to know the quantity of information contained in the sources. One of the requirements is "having a lot of information". That information about users is obtained from their interactions with the system. Users interact to carry out a search, to make a purchase, to leave an opinion about a product, etc. If we consider that each time users interact with the system information can be obtained from them, then we can measure the quantity of information as the quantity of interactions. The frequency of the interactions is therefore an indicator of the quantity of information contained in the sources. For example, given sources A and B with 10 and 20 users respectively, if the 10 users in source A interact approximately 100 times and the users in source B only interact twice, source A will have more information about users than source B has.

- **Timeliness**. This characteristic indicates the degree of source updating. The information dealt with concerns user preferences that change over time for various reasons. For example, it is more likely that users who would have been recommended an action movie 10 years are recommended children's movies now because they are parents who buy children's movies for their children. For that reason, a source with updated information will be more valuable than a source with less recent information.

- **Number of relevant attributes**. Although the source is complete, has a lot of information about users and is timeliness, it might not have the exact information to make the recommendations. For example, a source might be timeliness, might have many user interactions, might be very diverse, but not contain the type of movies that the user has bought and making recommendations requires knowing the type of movies the users buy. The quantity of relevant attributes is a characteristic that expresses if that information is in the source or not.

These are the characteristics of an information source that should be evaluated before being used by a recommender system. There should be a way of obtaining a value for each of them from the data stored in the sources. Then, based on these characteristics, know if a source offers the information that recommenders need and be used by them. Equations have been defined to obtain a numerical value that represents to which degree a given characteristic is present in the source. Drawing an analogy between sources and people, the intrinsic characteristics of people are, for example, their genre, age, height and name, and the values given to these characteristics are: genre = man, age = 28, height = 1.67 metres and name = Mario. So each one of these intrinsic characteristics of the information sources has a value and the equations defined to obtain it are presented in the following section.

Completeness

The completeness of the sources has been studied by Motro [Motro and Rakov, 1998] who formalised the concept of partial completeness in relational databases by restricting the completeness of the information that must be part of a complete database. More recently, in [Naumann et al., 2004] the concepts of "Coverage" and "Density" were introduced to measure the completeness of databases on the Web.

"Coverage" describes how many objects of the real world can provide a source of information S and is obtained with the following equation 3.1.

$$c(S) = \frac{|S|}{|W|} \tag{3.1}$$

where $|S|$is the quantity of different objects in source S and $|W|$ is the quantity of objects in the real world.

"Density" measures how much data (not null values) source S can provide for each one of these objects and is calculated with equation 3.2.

$$d(S) = \frac{1}{|A|} \sum_{a \epsilon A} d_S(a) \tag{3.2}$$

Given the set A of attributes, $|A|$ is the quantity of attributes of A and $d_S(a)$ is the density of an attribute $a\,\epsilon\,A$ (quantity of non null values of this attribute).

The equation used to measure the completeness of a source S using these two measures is:

$$Completeness(S) = \frac{\sum_{a \epsilon A} d_S(a)}{|W| * |A|} \tag{3.3}$$

Where $|W|$ is the quantity of objects of the real world and $|A|$ is the quantity of attributes of S.

The aim of this work is to know how complete a source S is with respect to the quantity of users for whom it can provide information. This information can be obtained using only the "Coverage" measure. So we reformulate and redefine this measure to adapt it to our problem:

Definition 1. 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 $|U|$.

$$Completeness(S) = \frac{|C|}{|U|} \tag{3.4}$$

Diversity

In ecology, a diversity index is a statistic used to measure the biodiversity of an ecosystem [wikipedia.org, 2007]. Diversity indexes provide information about the composition of the community (for example, the quantity of species), and also take into account the relative abundance of the various species. In this work, the diversity measure is used to represent, in a single value, the quantity of species (groups of users) in a source. Knowing information about the diversity of the source, rec-

ommender systems can:

- Differentiate users one from the other.
- Follow a given criterion to group users according to a relation or degree of similarity between them.

The diversity of information sources is measured using the "index of diversity" defined by Shannon and Weber in biology [Hilderman and Hamilton, 2000].

Definition 2. The diversity of a source S is equal to the entropy H.

$$Diversity(S) = H \tag{3.5}$$

H represents the entropy and is calculated as:

$$H = -\sum(p_i \log_2 p_i) \tag{3.6}$$

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

$$p_i = \frac{n_i}{N} \tag{3.7}$$

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 genre, age, etc.

Frequency of interactions

Analysis of the frequency of user interactions with the sources of information is the technique we use to determine the quantity of information about the users in the sources examined and with what frequency they interact.

We use the RFM (Recency, Frequency and Monetary) technique to measure the frequency of the interactions. RFM has been used in Direct Marketing for more than 30 years [Hughes, 2000]. Frequency is defined as the number of times that a client has made a purchase. The frequency of client interactions is an integral part of the RFM trilogy, which is used in marketing campaigns to predict the response

of clients. In the RFM technique clients are grouped according to the frequency of their interactions, which can be measured in several ways. For example, a seller can count the number of purchases a client has made in a year, while in a bank frequency can be measured as the quantity of checks a client has written and the number of deposits he or she has made. A hotel can measure the number of nights a client has occupied a room. A telephone company can measure frequency as the number of calls made or the number of minutes talked. Each company has a way of measuring frequency [Hughes, 2000]. In our case, we take the concept of frequency as defined in the RFM and we adapt it to our problem. The RFM measures the frequency of interactions of each user, but we need to know the frequency of interactions of a set of users in a source of information that could be used by a recommender. For example, if the source is a database of purchases, while more purchase records for individual users found in the database, better their tastes for certain products can be known by analysing the products they purchased in the past. If the source is a database storing information about visits made by users to a web site, while more stored visits there are, more information there will be about users.

Categories have been defined to obtain this measure. Each category represents a certain number of user interactions. For example, if we have a database containing information about the movies rented by the users of a video club, each record in the database represents a rental. Each of these records contains the user identifier, the movie rented and other data about the movie. Let's suppose that the defined categories are:

Category f_1 : 1 - 10 interactions

Category f_2 : 11 - 25 interactions

Category f_3 : 26 - 50 interactions

Category f_4 : 51 - 100 interactions

Category f_5 : 101 - 200 interactions

Category f_6 : + 201 interactions

Category f_1 includes the users who have rented a film from 1 to 10 times, category f_2 includes the users who have rented a film from 11 to 25 times, and so on. It

is clear that if the database contains more users in category f_6 there will be more information about them to obtain their preferences. This measurement is important because recommender systems choose sources from among those that can provide the most information to solve the problems of a lack of information.

Definition 3. The frequency of interactions of a source S is the sum of 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.

$$Frequency(S) = \frac{\sum w_i * |f_i|}{N} \tag{3.8}$$

Timeliness

Generally the updating of the information is defined as the time that has passed since the last update of the data [Naumann, 2004]. This way of measuring the "age" of the information is applied to resources on the Web, but cannot be applied to our problem. Our sources of information are databases containing information about users and their interactions. Application of this definition would provide the date of the client's last interaction but that does not mean that the database has been updated. Let's assume there are two databases A and B. In A three purchases by clients made in December of 2006 have been stored, while in B two purchases by clients have been stored, one made in February of 2006 and the other in January of 2007. Applying the most recent date criterion, B would be the most updated database, but for our purposes the database that would have the most updated information is A because all the purchases in it were made in the last three months. The updating measure defined by Phillip Cykana [Cykana et al., 1996] is more adapted to what we need to know about whether or not a source is more timeliness for our purpose. It measures update as the percentage of the data available within an interval of specified time (for example, day, hours or minutes).

In order to apply this measure to sources containing information about users, the date of their interactions is analysed. If more users have interacted recently, the information used to obtain their preferences and make recommendations to them will be more updated.

Following the example mentioned in the measure of frequency, and using the

database containing user purchase information, while more purchases made in the last months, more updated will be the database and their preferences will be those of now and not those of 10 years ago.

In order to obtain this measure time categories have been defined:

Category p_1 : 01/01/2001 - 31/12/2001

Category p_2 : 01/01/2002 - 31/12/2002

Category p_3 : 01/01/2003 - 31/12/2003

Category p_4 : 01/01/2004 - 31/12/2004

Category p_5 : 01/01/2005 - 31/12/2005

Category p_6 : 01/01/2006 - 31/12/2006

Category p_1 includes users who rented movies from 01/01/2001 to 31/12/2001, category p_2 includes users who rented movies from 01/01/2002 to 31/12/2002 and so on continuously for each of the categories. Clearly if the database contains more users in category p_6, that information will be more updated. As the preferences and tastes of users change over time, it is important to take into account a measure of information updating when using this source.

Definition 4. 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, divided by the quantity of records of S, which is N.

$$Timeliness(S) = \frac{\sum w_i * |p_i|}{N} \tag{3.9}$$

Number of relevant attributes

This measurement is used to determine the existence of relevant information to make the recommendations. Continuing with the example of the video club, if a new movie is available for rent, it can be recommended to users who like movies of the same type (action, drama, etc). In order to recommend the new movie it is

necessary to know information about the users, movie and type of the movie.

Definition 5. 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 $|D|$.

$$RelevantAtributes(S) = \frac{|B|}{|D|} \tag{3.10}$$

Example 2:

The objective is to recommend a book to each of the customers of a retail chain and it is important to search for information about them in the various sources available. Access is given to the following databases:

Source F1: database with information about book purchases
Source F2: database with information about music purchases
Source F3: database with information about video purchases
Source F4: database with information about game purchases

Source F1 is established as the main source and information from the most relevant sources of the set of available sources is added to it. In this case they are F2, F3 and F4. To know if these sources contain relevant information with which to make more precise recommendations, the characteristics of each of them are calculated.

Figure 3.2 presents information about the purchase of books stored in the database that makes up source F1 in our example.

Source F1 has information about 50 purchases made by 12 customers between 2001 and 2006. As can be seen in the figure, the table contains the following data:

IdCompra	Usuario	Sexo	Zona Vivienda	Libro	Genero	Precio	Unidades	Fecha
122255502	u256941	M	Z1	Buenas noches, Luna	Infantil	5,69 €	1	24/12/2006
122255501	u125896	M	Z2	Angeles y Demonios / Angels a	Misterio	9,95 €	1	24/12/2006
122255503	u124589	H	Z1	El Codigo Da Vinci	Misterio	17,90 €	1	21/12/2006
122255504	u125896	M	Z2	La ciudad letrada	Historia	12,00 €	1	18/12/2006
122255505	u125896	M	Z2	Sin remordimientos	Misterio	9,95 €	1	15/12/2006
122255506	u845713	M	Z3	Soldados De Salamina	Historia	12,00 €	1	15/12/2006
122255507	u256941	M	Z1	Huevos verdes con jamón	Infantil	6,95 €	1	02/10/2006
122255508	u845713	M	Z3	Como casarse con el hombre d	Romance	9,95 €	1	28/09/2006
122255509	u584236	M	Z3	Sin city: la gran masacre	Ficcion	18,00 €	1	23/08/2006
122255510	u125896	M	Z2	No Se Lo Digas a Nadie	Novela	11,00 €	1	18/08/2006
122255511	u418756	H	Z4	Once Minutos	Romance	21,95 €	1	05/05/2006
122255512	u245787	M	Z3	Fidel Castro: Biografía a dos vo	Historia	19,00 €	1	01/04/2006
122255513	u245787	M	Z3	Jerusalen: Caballo De Troya 1 (	Historia	9,95 €	1	18/02/2006
122255514	u124589	H	Z1	Las Cinco Personas Que Encor	Novela	11,16 €	1	12/02/2006
122255515	u845713	M	Z3	Mujeres de negro	Romance	10,00 €	1	03/02/2006
122255516	u145286	M	Z3	Irresistible	Romance	15,00 €	1	15/01/2006
122255517	u364125	M	Z4	Eragon (Spanish Language Edit	Ficcion	15,00 €	1	13/01/2006
122255518	u125896	M	Z2	Condicion Fisica para	Deporte	10,00 €	1	07/01/2006
122255519	u458932	H	Z3	La silla de plata (Narnia)	Ficcion	9,95 €	1	07/01/2006
122255520	u458932	H	Z3	La travesía del viajero del alba (	Ficcion	9,95 €	1	24/12/2005
122255521	u458932	H	Z3	El caballo y el muchacho (Narni	Ficcion	9,95 €	1	24/12/2005
122255522	u845713	M	Z3	Eragon (Spanish Language Edit	Ficcion	15,00 €	1	17/12/2005
122255523	u458932	H	Z3	Como casarse con el hombre d	Romance	9,95 €	1	15/12/2005
122255524	u256941	M	Z1	Mi familia y yo	Infantil	5,65 €	1	04/11/2005
122255525	u458932	H	Z3	El Senor De Los Anillos: el reto	Ficcion	10,00 €	1	17/10/2005
122255526	u458932	H	Z3	Autobiografía de un esclavo	Historia	8,96 €	1	13/09/2005
122255527	u125896	M	Z2	La frontera / Borderlands	Historia	8,65 €	1	12/08/2005
122255528	u584236	M	Z3	Asesinos	Ficcion	15,00 €	1	21/06/2005
122255529	u458932	H	Z3	La milla verde	Ficcion	6,95 €	1	30/04/2005
122255530	u458932	H	Z3	Relato de un naufrago	Ficcion	17,00 €	1	02/04/2005

Figure 3.2: Data base containing information about books purchases

identifier of the purchase, identifier of the customer, genre, zone where he/she lived in, book, type of book (novel, science fiction, history, etc.), units bought, price and date of the purchase. Based on this information, the characteristics are calculated to know the degree that this source could be relevant for the recommendations.

Calculation of the characteristics of source F1 from *Example 2* is presented in detail. Then the values of these characteristics for sources F2, F3 and F4, obtained applying the same process, are shown.

Completeness of source F1

The completeness of source F1 is obtained applying equation 3.4. The users about whom information is needed for recommendations are listed below:

Users: u124589; u125896; u145286; u157893; u245787; u256941; u364125; u418756; u425814; u458932; u584236; u845713.

These 12 users are found in source F1, so their completeness is:

$$Completeness(F1) = \frac{12}{12} = 1$$

Diversity of source F1

Diversity is calculated using data about the genre of the users and the zone where they live in. As can be seen in Figure 3.2 the users can be Men (M) or Women (W) and can live in the zones Z1, Z2, Z3 or Z4. Applying equation 3.7 to each of the genre groups leads to:

$$p_M = \frac{4}{12} = 0.33$$
$$p_W = \frac{8}{12} = 0.66$$

With these results, equation 3.6 is applied:

$$H = -((0.33 * (-1.58)) + (0.66 * (-0.58)))/12 = 0.91$$

To normalise this result it is divided by the number of groups (2 groups: M and W):

$$Diversity(F1_{sexo}) = H = 0.91/2 = 0.45$$

Diversity with respect to the zone where the customers live in is calculated following the same process:

$$p_{Z1} = \frac{2}{12} = 0.16$$
$$p_{Z2} = \frac{1}{12} = 0.08$$
$$p_{Z3} = \frac{7}{12} = 0.58$$
$$p_{Z4} = \frac{2}{12} = 0.16$$

With these results, equation 3.6 is applied:

$$H = -((0.16 * (-2.58)) + (0.08 * (-3.58)) + (0.58 * (-0.77)) + (0.16 * (-2.58)))/12 = 1,61$$

To normalise this result, it is divided by the number of groups (4 groups: Z1, Z2, Z3 and Z4):

$$Diversidad(F1_{zone}) = H = 1.61/4 = 0.40$$

Frequency of the source F1 interactions

The frequency of the interactions is calculated using equation 3.8. The following categories are defined:

Category f_1: 1 - 5 interactions
Category f_2: 6 - 10 interactions
Category f_3: 11 - 15 interactions

If we assign the following weights to each of these categories:

$$w_{f1} = 0.3$$
$$w_{f2} = 0.7$$
$$w_{f3} = 0.9$$

And suppose that quantity of users per category ($|f_i|$) is:

$$|f_1| = 10 \text{ users}$$
$$|f_2| = 1 \text{ users}$$
$$|f_3| = 1 \text{ users}$$

With all of this information the frequency of the interactions is calculated as:

$$Frequency(F1) = ((0.3 * 10) + (0.7 * 1) + (0.9 * 1))/12 = 4.6$$

To normalise this result it is divided by the number of categories (3 categories: f_1, f_2 y f_3):

$$Frequency(F1) = 4.6/3 = 0.38$$

Timeliness source F1

The timeliness of the source is calculated using equation 3.9. The categories defined for this purpose are:

Category p_1: 01/01/2001 - 31/12/2001
Category p_2: 01/01/2002 - 31/12/2002

Category p_3: 01/01/2003 - 31/12/2003
Category p_4: 01/01/2004 - 31/12/2004
Category p_5: 01/01/2005 - 31/12/2005
Category p_6: 01/01/2006 - 31/12/2006

The weight assigned to each of these categories is:

$$w_{p1} = 0.3$$
$$w_{p2} = 0.3$$
$$w_{p3} = 0.5$$
$$w_{p4} = 0.7$$
$$w_{p5} = 0.9$$
$$w_{p6} = 0.9$$

The quantity of users who bought a product during the dates indicated for each category are:

$$|p_1| = 2 \text{ users}$$
$$|p_2| = 10 \text{ users}$$
$$|p_3| = 4 \text{ users}$$
$$|p_4| = 4 \text{ users}$$
$$|p_5| = 10 \text{ users}$$
$$|p_6| = 19 \text{ users}$$

The timeliness of source F1 is:

$$Timeliness(F1) = (0.3 * 2) + (0.3 * 10) + (0.5 * 4) + (0.7 * 4)$$
$$+(0.9 * 10) + (0.9 * 19)/50 = 0.70$$

Number of relevant attributes of source F1

The quantity of relevant attributes of source F1 is obtained with equation 3.10. There are 3 attributes needed to make the recommendations: customer, product and type.The three attributes are available in source F1, as can be seen in Figure 3.2. The number of relevant attributes of F1 is:

$$RelevantAttributes(F1) = \tfrac{3}{3} = 1$$

Characteristics	**F1**
Completeness	1.00
Diversity	
Genre	0.45
Zone	0.40
Frequency	0.38
Timeliness	0.70
Relevant Attributes	1.00

Table 3.1: Characteristics of the source F1

Characteristics	**F1**	**F2**	**F3**	**F4**
Completeness	1.00	0.83	0.91	0.58
Diversity				
Genre	0.45	0.44	0.42	0.43
Zone	0.40	0.28	0.37	0.36
Frequency	0.38	0.28	0.30	0.20
Timeliness	0.70	0.49	0.60	0.76
Relevant Attributes	1.00	0.66	1.00	0.66

Table 3.2: Characteristics of sources from Example 2

Table 3.1 presents the values obtained from each of the characteristics for source F1.

The values of the characteristics of sources F2 (music), F3 (videos) and F4 (games) have been obtained in the same way as they were for source F1. These values are shown in Table 3.2.

These values indicate that source **F1** is the most *complete*, containing all the customers from whom information is needed. The *diversity* values are the highest, with not all customers being of the same genre or from the same zone. It is more *timeliness* so recent information about preferences and tastes can be obtained. It has a *lot of information* about book purchases because users bought them more often and it contains all the *attributes* necessary to make recommendations: customer, product and genre. On the other hand, source **F4** is the least *complete* because it does not contain information about all customers about whom information is needed. Although the *diversity* values are not the lowest, the *frequency* value indicates that customers interacted very little in this source. A value of 0.76 in the *timeliness* indicates it is an updated source, that means all games were purchased recently, but **F4** does not contain all the attributes needed to make the recommendations. Source **F3** is more *complete* than **F4**, more *diverse* in relation to zones

where users live and contains all the necessary *attributes*, but it is less *diverse* with respect to the genre of the users and less *timeliness* than **F3**. Source **F2** is more *diverse* in relation to genre, but less in relation to the zone lived in. It is also less *timeliness*, contains less information about the *interactions* and does not contain all the *attributes* needed to make recommendations. 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 unlikely. 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 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.

Definition 6. The relevance (**R**) of a source S is the sum of the values c_j of each of the characteristics j multiplied by the weight w_j assigned to each of these characteristics divided by the quantity of characteristics $|N|$.

$$R(S) = \frac{\sum w_j * c_j}{|N|} \tag{3.11}$$

Example 3:

In this example the relevance of each one of the sources from Example 2 (F1, F2, F3 and F4) is calculated using the values of their characteristics and their weights are listed in Table 3.3:

R(F1)=((0.90*1.00)+(0.50*0.45)+(0.50*0.40)+(0.90*0.38)
+(0.70*0.70)+(0.90*1.00))/6 = **0.51**
R(F2)=((0.90*0.83)+(0.50*0.44)+(0.50*0.28)+(0.90*0.28)
+(0.70*0.49)+(0.90*0.66))/6 = **0.38**
R(F3)=((0.90*0.91)+(0.50*0.42)+(0.50*0.37)+(0.90*0.30)
+(0.70*0.60)+(0.90*1.00))/6 = **0.47**
R(F4)=((0.90*0.58)+(0.50*0.43)+(0.50*0.36)+(0.90*0.20)
+(0.70*0.76)+(0.90*0.66))/6 = **0.37**

Characteristics	w_j
Completeness	0.90
Diversity(Genre)	0.50
Diversity(Zone)	0.50
Frequency	0.90
Timeliness	0.70
Relevant Attributes	0.90

Table 3.3: Weight of characteristics representing the importance that they have in a recommendation

As indicated by the relevance value, the most relevant source is source **F1**, which is the main source. Among all the sources available, F2, F3 and F4, the most relevant is source **F3**.

In order to evaluate if adding information to source F1 from the most relevant source F3 really improves the recommendations, recommendation with only information from F1 and recommendations with information from F1 and F3($F1 \leftarrow F3$) were made and evaluated.

To evaluate the recommendations the measure of precision defined by Salton [Salton and Buckley, 1988] and widely used in the field of Information Retrieval (IR) was used. In recommender systems the measure of precision represents the probability that a recommendation will be successful and is obtained using the following equation.

$$Precision = \frac{Pr}{P} \tag{3.12}$$

Where Pr is the quantity of successful recommendations. In our problem this parameter is calculated as the quantity of recommended products purchased by custommers. P is the total quantity of recommendations made; in this study it is obtained as the quantity of recommended products. The results shown in Table 3.4 demonstrate that the precision obtained when making recommendations with information from both sources **F1**←**F3** is higher than that obtained when making recommendations with information only from **F1**. Also, this table presents the precision obtained from recommendations made adding information to source F1 from the other sources:

$(F1 \leftarrow F2)$, $(F1 \leftarrow F4)$, $(F1 \leftarrow (F2, F3))$, $(F1 \leftarrow (F2, F4))$,$(F1 \leftarrow (F3, F4))$,$(F1 \leftarrow (F2, F3, F4))$.

Sources	Precision
$F1$	0.41
$F1 \leftarrow F2$	0.64
$F1 \leftarrow F3$	**0.73**
$F1 \leftarrow F4$	0.27
$F1 \leftarrow (F2, F3)$	0.86
$F1 \leftarrow (F2, F4)$	0.34
$F1 \leftarrow (F3, F4)$	0.45
$F1 \leftarrow (F2, F3, F4)$	0.52

Table 3.4: Precision of recommendations of Example 3

These results are illustrated in the graph in Figure 3.3. As can be seen, the recommendations made with information from source F1 plus information from source F3 have a high precision value, even though better precision was obtained with information from source F1 plus the information from sources F2 and F3.

3.2.2 Obtaining the trust measure of 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 et al., 1998], recommender systems [Montaner et al., 2002] [O'Donovan and Smyth, 2005] [Massa and Avesani, 2004] and social networks [Yu and Singh, 2002] [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 1 a reliable source. The trust of a source S is computed as the expected

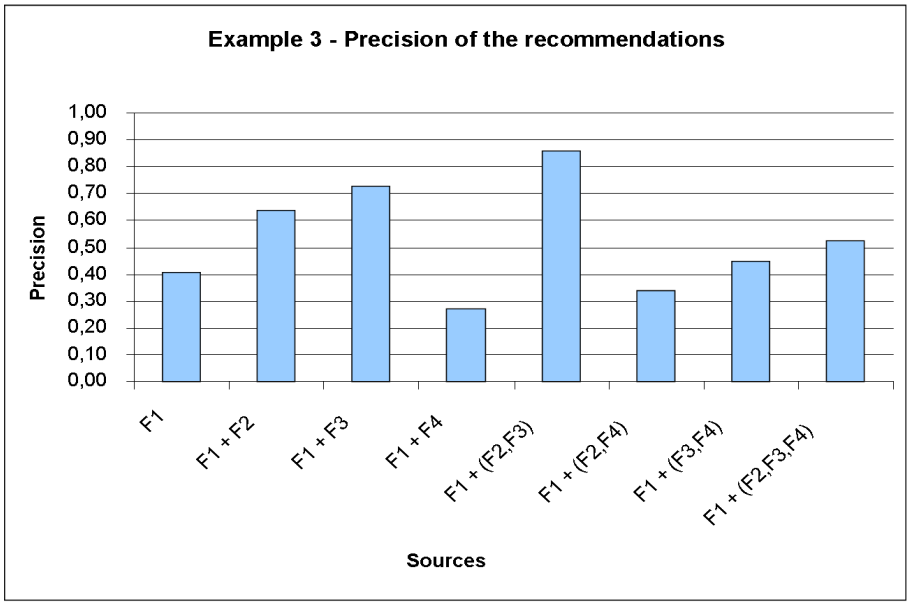


Figure 3.3: Precision of recommendations using sources from Example 2

value of a variable B_s given the parameters α and β. B_s. B_s is the expected value that S has relevant information. This value is obtained using equation 3.13.

$$T(S) = E[B_s/\alpha, \beta] \tag{3.13}$$

E is computed as follows:

$$E[B_s/\alpha, \beta] = \frac{\alpha}{\alpha + \beta} \tag{3.14}$$

The parameters α and β are defined as:

$$\alpha = m_S^{1:t} + 1 \tag{3.15}$$

$$\beta = n_S^{1:t} + 1 \tag{3.16}$$

where $m_S^{1:t}$ is the number of successful recommendations using source S, $n_S^{1:t}$ is the number of unsuccessful recommendations.

Example 4:

In this example the measure of trust of each of the sources from Example 2 is calculated. As it is the first time that those sources are used to make recommendations, there is no information to know whether or not they are reliable, so the trust value is:

Source F1: $\alpha = 0 + 1 = 1$
$\beta = 0 + 1 = 1$
$T(F1) = \frac{1}{1+2} = 0.5$

Source F2: $\alpha = 0 + 1 = 1$
$\beta = 0 + 1 = 1$
$T(F2) = \frac{1}{1+2} = 0.5$

Source F3: $\alpha = 0 + 1 = 1$
$\beta = 0 + 1 = 1$
$T(F3) = \frac{1}{1+2} = 0.5$

Source F4: $\alpha = 0 + 1 = 1$
$\beta = 0 + 1 = 1$
$T(F4) = \frac{1}{1+2} = 0.5$

The trust value is 0.5 when sources are used for the first time. As the sources are used in recommendations and can be evaluated, the trust value will change. For example, Table 3.5 shows the precision values of the recommendations obtained using information from source F2. If the precision is $>= 0.5$, the recommendation is considered "successful" and is indicated as $Result = 1$; otherwise the recommendation is considered "unsuccessful" and is indicated as $Result = 0$. Using these data, the trust in source *F2*, which has been used 12 times to make recommendations, is:

$$\alpha = 10 + 1 = 11$$
$$\beta = 2 + 1 = 3$$
$$T(F2) = \frac{11}{11+3} = 0.78$$

A trust value equal to 0.78 indicates that the source is reliable to make the recommendations. That means, the recommendations were successful most of the time that this source was used.

Source	Precision	Result
F2	0.73	1
F2	0.58	1
F2	0.70	1
F2	0.50	1
F2	0.25	0
F2	0.52	1
F2	0.69	1
F2	0.61	1
F2	0.36	0
F2	0.65	1
F2	0.56	1
F2	0.62	1

Table 3.5: Result of recommendations using information from the source F2

3.2.3 Selecting the most suitable source of information

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 composed of 3 elements:

1. A set (S) of candidate sources.
2. A selection function *Selection(R(s),T(s))* 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 chosen ($F \subset S$).

In every step the algorithm chooses a source of **S**, let us call it **s**. Next it checks if the $s \cup F$ 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.

The parameters of the selection function are the relevance (**R(s)**) and trust (**T(s)**) of the sources. This function returns a value between 0 and 1, and is obtained through equation 3.17.

$$selection(R(s), T(s)) = R(s) * T(s) \tag{3.17}$$

Algorithm to select relevant and trustworthy source

Algorithm (S: Set of candidates sources)
$F := \emptyset$;
while ($S <> \emptyset$) **do**
  if Selection(R(s),T(s))> threshold **then**
    $F := F \cup s$;
    $Eliminate(s, S)$;
  end if
end while
return F;

If a source s has a lower R(s), the characteristics are not good for the recommendation and the value of T(s) is lower, meaning that the source was not useful in previous recommendations, and in this case the selection function will return a lower value. By contrast, if it has a higher R(s) and T(s), the source is useful and contains good information for the recommendation.

Example 5:

Continuing with the sources from Example 2 and assuming that at a given moment these sources have the relevance and trust values shown in Table 3.6, the selection function when applying the selection algorithm is:

$$selection(R(F1), T(F1)) = 0.51 * 0.89 = 0.45$$
$$selection(R(F2), T(F2)) = 0.38 * 0.67 = 0.25$$
$$selection(R(F3), T(F3)) = 0.47 * 0.67 = 0.31$$
$$selection(R(F4), T(F4)) = 0.37 * 0.25 = 0.09$$

If we use the sentence:

$$IF(Selection(R(s), T(s)) > 0.10)$$

The sources selected are the sources F1, F2 and F3. With information from the selected sources and using the algorithm, 20 recommendations were made. Recommendations with the selected sources based on R were also made. The precision value was obtained for one of the recommendations shown in Figure 3.4. These values indicate that recommendations made based on relevance and trust are more stable and their results improve. In other words, applying the algorithm tends to find the optimal combination of sources for more precise recommendations.

s	R(s)	T(s)
F1	0.51	0.89
F2	0.38	0.67
F3	0.47	0.67
F4	0.37	0.25

Table 3.6: Relevance and Trust of the source F2 from Example 2

3.3 Experimental results

This section presents two case studies carried out to show the relevance of the proposals made in this thesis. The experiments were performed using information from databases in Consumer Package Goods Domain (retail) and databases from various domains obtained from the information available on Amazon.com. The first case study was carried out to test the characteristics and their application in any recommendation method. The sources were selected taking into account the relevance value. In the second case study, more complete than the first, the selection was made based on the relevance value and the trust measure of the sources. These results are compared with those obtained when making recommendations with sources of information selected according to different criteria. In order to show that ACQUAINT methodology can be used by any recommender system; three methods have been implemented to make recommendations such as CBF (Content-Based Filtering), RFM (Recency, Frequency and Monetary) and CF (Collaborative Filtering). CBF uses a deep knowledge of a user. It needs to know the user's preferences about attributes of a product. RFM which is widely used in Marketing to segment customers' base on purchased history. CF makes recommendation to a user according the preferences of similar users. The three methods use different user knowledge. If we obtain good results applying the methodology with the three methods, so we can say that the methodology is general.

The next sections present each one of the case studies in detail.

3.3.1 Case study 1: Recommendation in retail domain

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 chapter. The databases contain real information about 4137 customers, products and their purchases made in the period 2002-2003. All these purchases were made

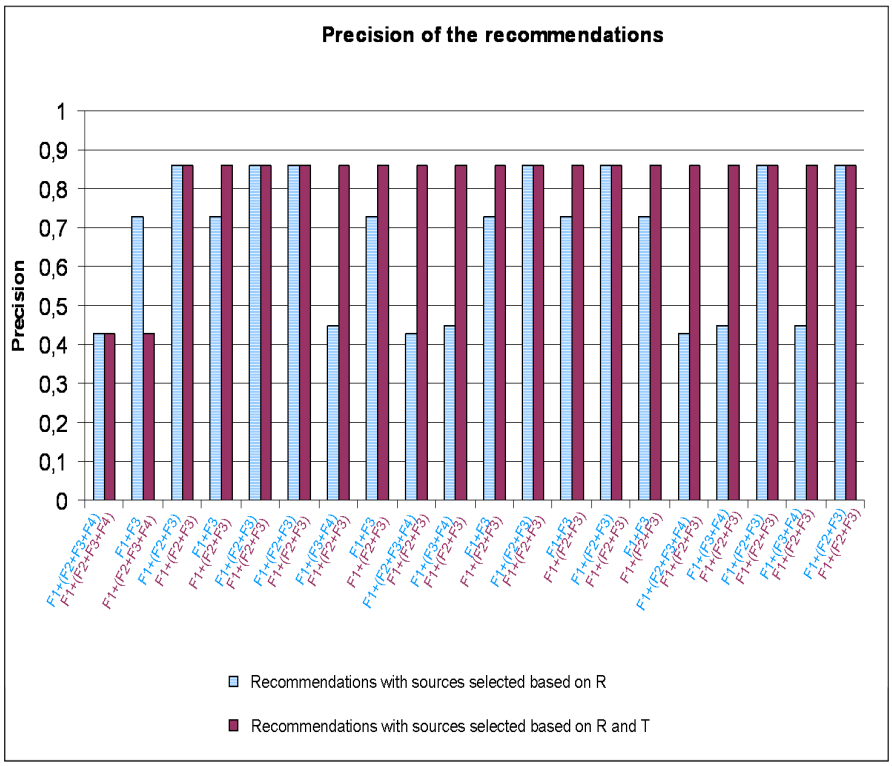
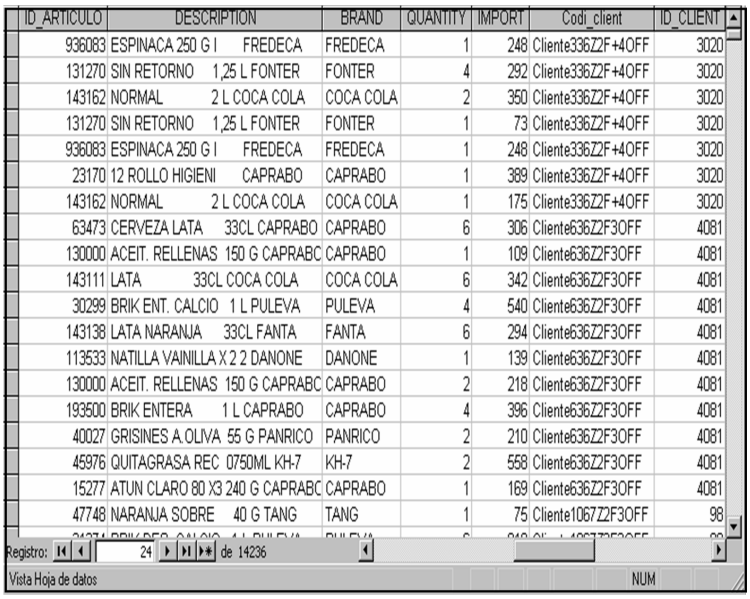


Figure 3.4: Comparison of the precision of recommendations made based on the Relevance and Trust of the sources

either on the Internet (online) or in the supermarket (offline). The common customersus of the various databases are identifiable because each one has the same identifier in all database. Figure 3.5 shows a sample of a database, specifically a transaction database, employed in this research.

Two sources of information were used in this case study: off-line sources contained information about the purchases made in the supermarket and on-line sources contained information about the purchases made on the Internet. The table shown in Figure 3.5 contains information from one of the databases about the transactions conducted by customers in the supermarket (offline). As shown in this figure, it contains the information gathered during the purchase process: receipt number, type of product, number of purchased products and amount spent on each item and other data.

To carry out our experiments we randomly divided both transaction databases into eight sub-databases named: S1, S2, S3, S4, S5, S6, S7 and S8. These sub-databases constituted our main sources of information. We were able to evaluate their characteristics and determine their relevance and choose the most relevant to make recommendations. Let



ID_ARTICULO	DESCRIPTION	BRAND	QUANTITY	IMPORT	Codi_client	ID_CLIENT
936083	ESPINACA 250 G I FREDECA	FREDECA	1	248	Cliente336Z2F+4OFF	3020
131270	SIN RETORNO 1,25 L FONTER	FONTER	4	292	Cliente336Z2F+4OFF	3020
143162	NORMAL 2 L COCA COLA	COCA COLA	2	350	Cliente336Z2F+4OFF	3020
131270	SIN RETORNO 1,25 L FONTER	FONTER	1	73	Cliente336Z2F+4OFF	3020
936083	ESPINACA 250 G I FREDECA	FREDECA	1	248	Cliente336Z2F+4OFF	3020
23170	12 ROLLO HIGIENI CAPRABO	CAPRABO	1	389	Cliente336Z2F+4OFF	3020
143162	NORMAL 2 L COCA COLA	COCA COLA	1	175	Cliente336Z2F+4OFF	3020
63473	CERVEZA LATA 33CL CAPRABO	CAPRABO	6	306	Cliente636Z2F3OFF	4081
130000	ACEIT. RELLENAS 150 G CAPRABC	CAPRABO	1	109	Cliente636Z2F3OFF	4081
143111	LATA 33CL COCA COLA	COCA COLA	6	342	Cliente636Z2F3OFF	4081
30299	BRIK ENT. CALCIO 1 L PULEVA	PULEVA	4	540	Cliente636Z2F3OFF	4081
143138	LATA NARANJA 33CL FANTA	FANTA	6	294	Cliente636Z2F3OFF	4081
113533	NATILLA VAINILLA X 2 2 DANONE	DANONE	1	139	Cliente636Z2F3OFF	4081
130000	ACEIT. RELLENAS 150 G CAPRABC	CAPRABO	2	218	Cliente636Z2F3OFF	4081
193500	BRIK ENTERA 1 L CAPRABO	CAPRABO	4	396	Cliente636Z2F3OFF	4081
40027	GRISINES A.OLIVA 55 G PANRICO	PANRICO	2	210	Cliente636Z2F3OFF	4081
45976	QUITAGRASA REC 0750ML KH-7	KH-7	2	558	Cliente636Z2F3OFF	4081
15277	ATUN CLARO 80 X3 240 G CAPRABC	CAPRABO	1	169	Cliente636Z2F3OFF	4081
47748	NARANJA SOBRE 40 G TANG	TANG	1	75	Cliente1067Z2F3OFF	98

Figure 3.5: An example of the database used in the study: a transaction database

us apply the ACQUAINT methodology; first of all, we have to obtain the characteristics of each source. In fact, these characteristics and their value help to answer some very important questions related to the use of information for recommendations:

- How we can measure the information contained in the sources to acquire more knowledge about users?
- To what extent do recommendations improve the use of more information sources?

Table 3.7 shows the main values of each characteristic from the eight databases used in this study case. These values have been obtained using the formulas from Section 3.2.1.

The quantity of relevant attributes needed to make the recommendation (represented by parameter D in Equation 3.10) is 8. They include date of purchase, type of purchased product, number of units, price, total amount, genre of the consumer, area consumers live in and number of members in the family.

With the purpose of measuring the diversity, the customers were clustered according to the:

- Area where they live (Z): Zone 1, Zone 2, Zone 3 and Zone 4. These zones have been established by an expert in the supermarket.

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 3.7: Characteristic of databases from Caprabo, S1-S8

- Number of members in the family (F): family composed by 1 member, family with 2 members, family with 3 members, family with 4 members and family with more than 4 members.

- Genre (G): 2 groups, men and women.

The frequency of each source has been calculated using 4 periods of time: Period 1 (P1) from January of 2002 until June of 2002, Period 2 (P2) from July of 2002 until December of 2002, Period 3 (P3) from January of 2003 until June of 2003 and Period 4 (P4) from July of 2003 until December of 2003. The oldest period (P1) has the lowest weight (w_i) and the most recent period (P4) has the highest weight. The timeliness measure has been obtained using the same period of times: P1, P2, P3 and P4.

Once the value of the characteristics has been obtained, it is necessary to know if the recommendation could be improved by selecting the most relevant information sources. Therefore, we need to know the relevance of each source based on its characteristics.

Using equation 3.11 and the values of w_i shown in Table 3.8, the relevance value of source S1 is:

$$R(S1) = (0.9 * 0.8) + (0.5 * 0.1) + (0.3 * 0.13) + (0.3 * 0.33)$$
$$+(0.3 * 0.20) + (0.7 * 0.23) + (0.5 * 0.25)/7 = 0.39$$

Thus we have calculated the relevance of each source and the results are shown in Table 3.9.

As can be seen in the Table 3.9, the sources S3 and S7 have lower relevance coefficients than other sources. In order to select the source to integrate the information we apply the next rule:

Characteristics	w_i
Relevant attributes	0.9
Completeness	0.5
Diversity (Z)	0.3
Diversity (F)	0.3
Diversity (G)	0.3
Frequency	0.7
Timeliness	0.5

Table 3.8: w_i weight using to calculate the relevance of each source

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
R(S)	0.17	0.21	0.14	0.23	0.24	0.25	0.13	0.26

Table 3.9: Relevance of sources of information

Given S1 where $R(S1) = R1$
Given S2 where $R(S2) = R2$
if $(R(S1) <= R(S2))$ **then**
  S2 is selected
  else S2 is discarded
end if

Once the source has been selected, we need to make recommendations and evaluate them to know the effectiveness of the characteristics. We apply three recommender methods to make recommendations, the RFM (Recency, Frequency and Monetary)algorithm, which is used in marketing to segment clients, and two of the most widely used methods in recommender systems, Content Based Filtering (CBF) and Collaborative Filtering (CF). In the next sections we detail the implementation of these methods in our problem.

Recommendation applying the RFM algorithm

Recommendations are made based on the analysis of the customer purchasing behaviour. We use the following variables to make a recommendation: the purchase frequency, the value of the purchase, the date of the last purchase and the type of purchased products.

Example 6:

Given a User X with the following information about his purchases:

Last purchase = 10/08/2006

Amount= $ 1000

Frequency = *low*

Product = *computer*

For this customer, the recommender will not recommend a computer because the frequency to buy such a product is low and the date of the last purchase is very recent.

In order to obtain this kind of knowledge, the RFM (Recency, Frequency, Monetary) algorithm was used [Hughes, 2000]. This is an algorithm that segments the customers according to their purchase behaviour. In fact, an RFM algorithm is a segmentation technique that allows more specific predictions of the buying behaviour of target groups of customers. It is used when there is a former purchase experience and it is based on segmenting groups of customers according to each variable, R, F and M. It improves the probability that a customer will buy a product based on the recommendation made according to his previous buying behaviour. Furthermore, it also allows the behaviour of segments of customers who show decreasing interest in certain products to be evaluated.

Overall, this analysis of behaviour provides the key to obtaining favourable answers to the recommendations that are made to potential customers. With the relevance of each source, the RFM algorithm was executed to obtain the buying behaviour of the customers in the selected database. The products were recommended according to this behaviour and the analysis of purchase receipt.

In order to evaluate the results of the recommendations we employed the precision measurement given by equation 3.12. We made the recommendations using all the sources independently of their relevance and we also made recommendations using the most relevant sources to demonstrate the effectiveness of our proposal. The results are shown in Figure 3.6 and Figure 3.7 respectively.

The first analysis of precision was done including all the sources without considering their relevance, S1-S8. In this case (see Figure 3.6), the results show that the precision of the recommendations aggregating information from all these sources is limited. For

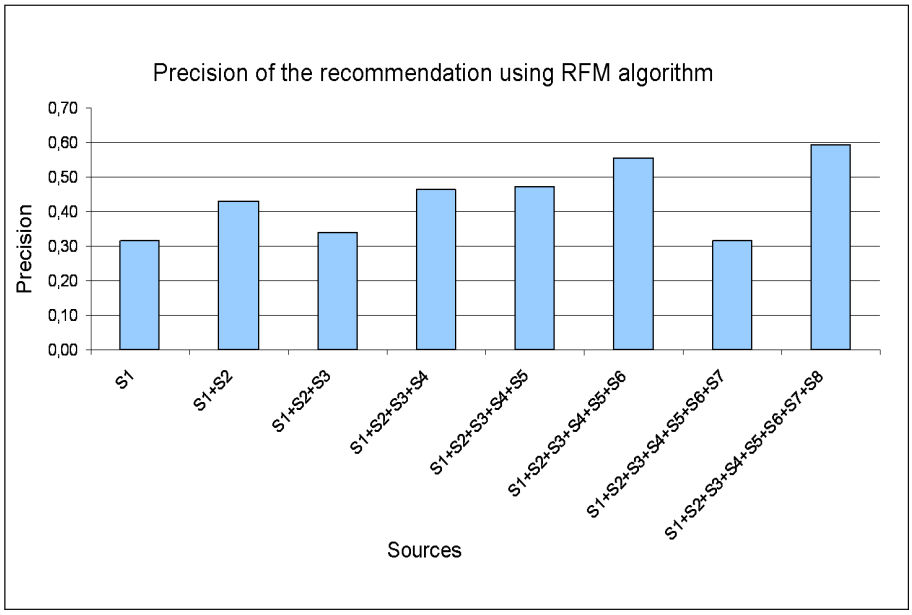


Figure 3.6: Precision of recommendations including all sources

example, using the first six databases, the precision is 5.50 but if we add S7 (the least relevant source), it falls to approximately 3. On the contrary, in Figure 3.7 we observe that if we only aggregate information sources with a relevance over 0.15, the precision shows a steadily increasing trend. Overall, the results show that the measurement of relevance is an important criterion to select sources of information to improve the precision of the recommendations.

Recommendations using CBF

In this method the attributes of products are extracted and compared with a user profile (preferences and tastes), in this case, a user is called customer, because is analysed the profile of a customer in the supermarket . Vectors are used to represent customer profiles and products. The cosine function based on the vector space proposed by Salton [Salton and Buckley, 1988] has been used to establish the relevance that a product has for a customer.

$$Cos(P,U) = \frac{\sum_{i=1}^{n}(p_i * u_i)}{\sqrt{\sum_{i=1}^{n} p_i^2} * \sqrt{\sum_{i=1}^{n} u_i^2}} \tag{3.18}$$

The products that have a higher value of similarity are recommended to customer.

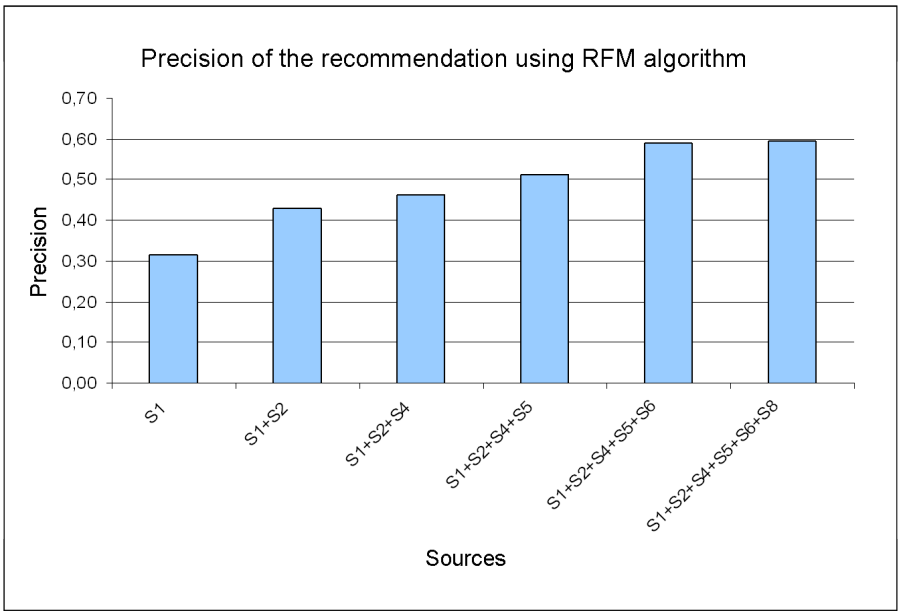


Figure 3.7: Precision of recommendations including relevant sources

Figure 3.8 shows the relevant attributes of the product defined by an expert from a supermarket. The customer preferences have been established based on these attributes and are represented by a vector:

$$U =< u_1, u_2, ..., u_i >$$

The weight u_i has been obtained using the TF-IDF method (Term Frequency times Inverse Document Frequency) [Salton and Buckley, 1988], which has been calculated based on previous purchases of the customer.

$$u_i = t_i * log_2(\frac{N}{n_i}) \tag{3.19}$$

t_i is the frequency of attribute i in the purchases, n_i is the number of customer who have bought a product with attribute i and N is the total number of customers.

The product vector is composed of the weight p_i which has been assigned by an expert in the supermarket where a brand of the product has higher weight because the brand is considered a very influential criterion that customers use to decide which product to buy..

$$P =< p_1, p_2, ..., p_i >$$

codi	nom
2	Marca
3	Tipo de compra
4	Genero sujeto consumidor
5	Perfil consumidor
6	Edad consumidor
7	Implicacion
8	Transportable
9	Frecuencia uso
10	Tipo producto
11	Ciclo de venta
12	Complementariedad
13	Caducidad
14	Madurez
15	Fresco
16	Salud
17	Precio
18	Origen
19	Practico, almacenaje, conservacior
20	Sensibilidad_Precio_PF
21	Sensibilidad_Precio_PFS
22	Sensibilidad_Precio_SIO
23	Sensibilidad_Precio_C
24	Sensibilidad_Precio_VL
25	Sensibilidad_Precio_CIO
26	Sensibilidad_Precio_HD
27	Sensibilidad_Precio_PB

Figure 3.8: Relevant attributes in the consumer package goods domain (retail)

The relevance of each product for the users has been established with equation 1 using the vectors representing the users and the products (See Figure 3.9).

The products with a value of relevance > 0.6 have been recommended to the customers.

The experiments have been done implementing the CBF with the information from all the sources (8 databases) without the methodology. The precision of recommendations has been evaluated using equation 3.12 and the results obtained are shown in Figure 3.10.

In Figure 3.11 the precision of recommendations made using the CBF with information of the selected sources in the methodology can be observed.

The graphs show how the precision of the recommendations made with these methods and the selected sources increase. Figure 3.10 shows the precision of the recommendations using all information sources and demonstrates that the aggregation of the information in the recommendations causes the precision to decrease. Figure 3.11 shows the precision of the recommendation using information only from the selected sources. This result shows that the smart integration of the data sources increases the precision. The selection of the sources is established based on characteristics of each source.

cliente	producto	Relevance
91850	32318	0,3633179926
91850	34971	0,3633179926
91850	34864	0,3166948446
91850	34781	0,6488153614
91850	34699	0,3969170281
91850	34533	0,3166948446
91850	32998	0,3166948446
91850	32993	0,3166948446
91850	32865	0,3969170281
91850	32492	0,3633179926
91850	32491	0,3633179926
91850	32490	0,3633179926
91850	32378	0,3633179926
91850	33535	0,3969170281
921330	44683	0,4541998345
921330	48587	0,8585791209
921330	48497	0,5599386555
921330	48472	0,8585791209
921330	48194	0,4541998345
921330	47748	0,4541998345
921330	47343	0,5599386555
921330	46127	0,5599386555
921330	46087	0,5599386555
921330	45976	0,8585791209
921330	43890	0,8585791209
921330	45479	0,4541998345
921330	44452	0,4541998345
921330	49326	0,5599386555
921330	45606	0,5599386555
921330	55178	0,8585791209
921330	65745	0,4541998345
921330	63473	0,4541998345
921330	60339	0,4541998345

Figure 3.9: Relevant product for customers

Recommendations using the CF

Information provided by customers with similar interests is used to determine the relevance that the products have for the customer. Similarity between customers is calculated for this purpose and the recommendations are made based only on this similarity; the purchased products are not analyzed as is done in the FBC. Also, the cosine vector similarity [Salton and Buckley, 1988] is used to compute the distance between the representation of the present customer and the other customers. All customers are represented by vectors.

$$Cos(U,V) = \frac{\sum_{i=1}^{n}(u_i * v_i)}{\sqrt{\sum_{i=1}^{n} u_i^2} * \sqrt{\sum_{i=1}^{n} v_i^2}} \tag{3.20}$$

Where U and V are the customer vectors.

The same attributes shown in Figure 3.8 have been used in the CF method to obtain a vector representation of the customer's preferences. The vector for each customer is:

$$U =< u_1, u_2, ..., u_n >$$

The weight u_i has been obtained from previous purchases of the customer using the TF-

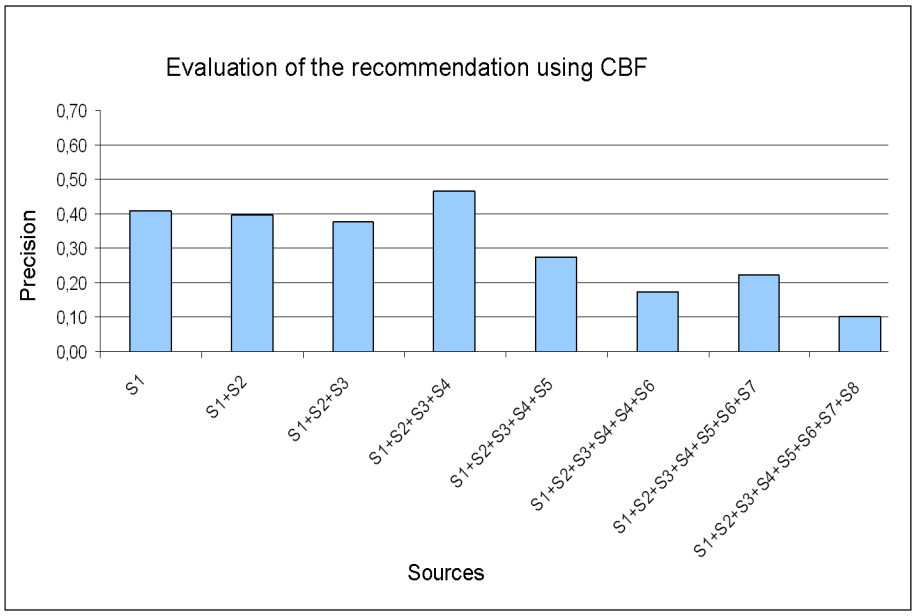


Figure 3.10: Precision of recommendations using CBF with all sources

IDF method (Term Frequency times Inverse Document Frequency) [Salton and Buckley, 1988] as in the CBF

$$u_i = t_i * log_2(\frac{N}{n_i}) \tag{3.21}$$

Where t_i is the frequency of attribute i in the purchases, n_i is the number of customer who have bought a product with attribute i and N is the total number of customers.

The similarity between customers has been established with equation 3.20 using the vectors representing the customers (see Figure 3.12).

The products bought by other customers with a value of similarity > 0.6 have been recommended to the customer.

The experiments have been done implementing the CF with the information from all the sources (8 databases) without the methodology. The precision of recommendations has been evaluated using equation 3.12 and the results obtained are shown in Figure 3.13.

Figure 3.14 presents the precision of the recommendations made using the CF with information from the selected sources with the methodology.

The graphs show how the precision of the recommendations made with these methods and the selected sources increases. Figure 3.13 shows the precision of the recommendations

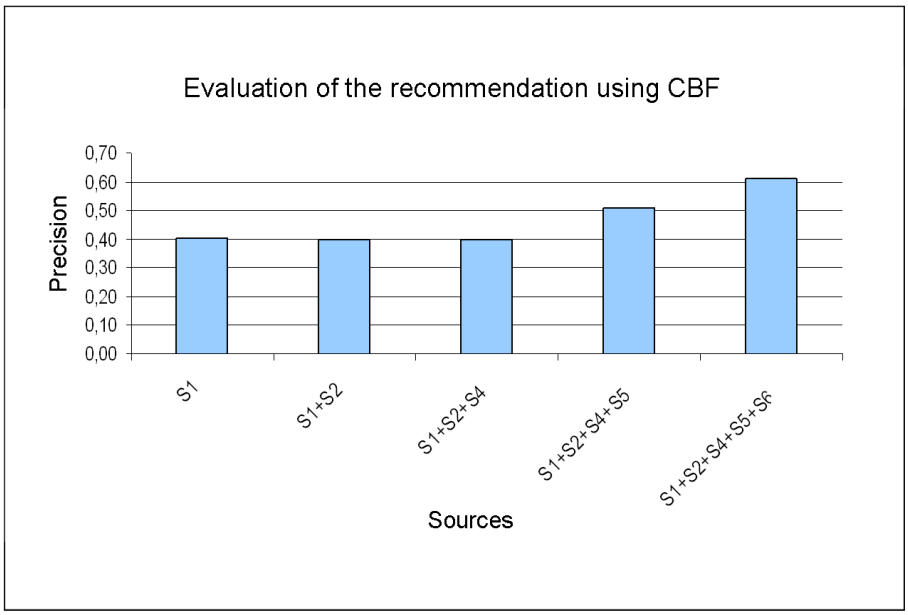


Figure 3.11: Precision of recommendations using CBF only with the selected sources with the methodology

using all information sources and it can be observed that the aggregation of the information in the recommendations causes the precision to decrease.

Figure 3.14 shows the precision of the recommendation using information only from the selected sources. This result shows that the integration of the data sources increases the precision. The selection of the sources is established based on the characteristics of each source.

3.3.2 Case study 2: Selecting relevant information using data from different domains

In this section, we will describe the recommendation domains choose to carry out our experiments. Suppose an scenario composed by four recommendation domains such as: Book recommendations, Compact Disk (CD) recommendations, Magazines recommendations and DVD recommendations. Each domain contain information about users, items and ratings. We have collected data from Amazon.com to obtain such information to build a data set for each domain. A popular feature of Amazon is the ability for users to submit reviews to the web page of each product. As part of their review, users must rate the product on a rating scale from one to five stars. Consumer's reviews about a product

VectorClienteU1_cliente	VectorClienteU2_cliente	SIM
1000285	299840	0,7426971197
1000285	2997940	0,7426971197
1000285	2998776	0,7426971197
1000285	2998773	0,7426971197
1000285	2998766	0,7112738823
1000285	2998515	0,7112738823
1000285	2998386	0,6798506450
1000285	2998342	0,6798506450
1000285	2998339	0,7112738823
1000285	2997973	0,7426971197
1000285	2998743	0,7112738823
1000285	2997953	0,7112738823
1000285	2998291	0,6484274077
1000285	2998022	0,7112738823
1000285	2998031	0,7426971197
1000285	2998045	0,7426971197
1000285	299807	0,6798506450
1000285	299819	0,7112738823
1000285	2998196	0,6798506450

Figure 3.12: Similarity between customers - CF

have been used to obtain information about users, the user's knowledge and experience with the product and a rating as valuation about the product. A review in Amazon.com include the following sections:

- User section contained the name of the review writer that could be real name or a fictitious name. The real name is identified by the target called "Real Name". A Real Name attribution is a signature based on the name entered by the author as the cardholder name on his or her credit card, i.e. the author represents this name as his/her identity in the "real world." [Amazon, 2006]. Also, this section could include the place where he/she lives.

- Product section refers to the product which the users give the reviews.

- Rating section reflects the summarized user's opinion about an experienced product and represents the overall user satisfaction for a given product. A rating is an "stars scale" ranging from 1 star ($\star$) to 5 star ($\star\star\star\star\star$), where 1 reflects "worst" and 5 reflects "best".

- Date reflects the time when the reviews were made.

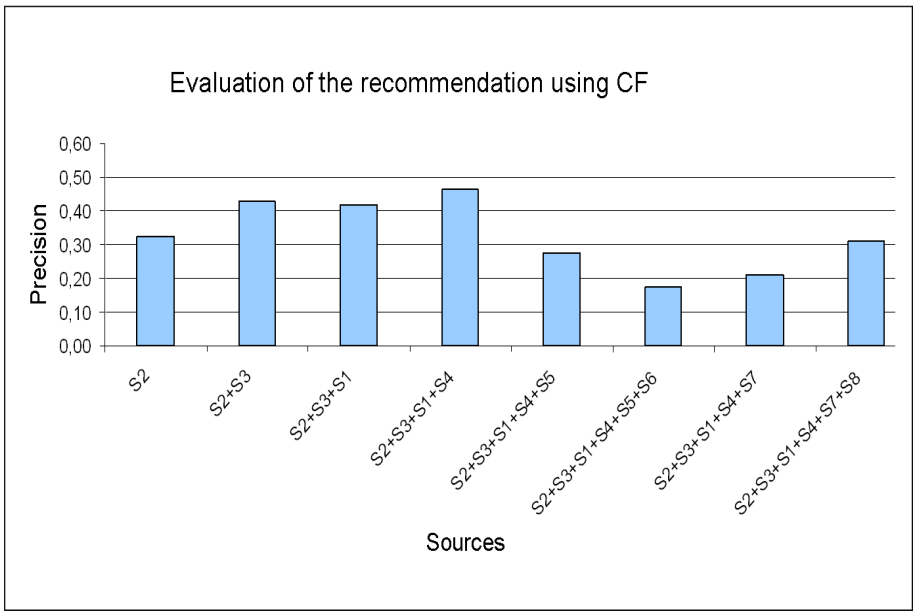


Figure 3.13: Precision of recommendations using CF with all sources

- Content section contained a full description of the user's opinion and experiences with a product. This part is text.

These sections are illustrated in Figure 3.15. Reviewers can submit reviews for a wide range of product such as CDs, DVDs, software, consumer electronics, kitchen items, tools, lawn and garden items, toys and games, baby products, apparel, sporting goods, gourmet food, jewellery, watches, health and personal-care items, beauty products, musical instruments, industrial and scientific supplies, groceries and more.

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 source. The information collected is resumed below:

Data set:

Domains: Book, CD, DVD, Magazine
Source S1: Information from Books domain
Rating = 732
Users = 124
Books =699

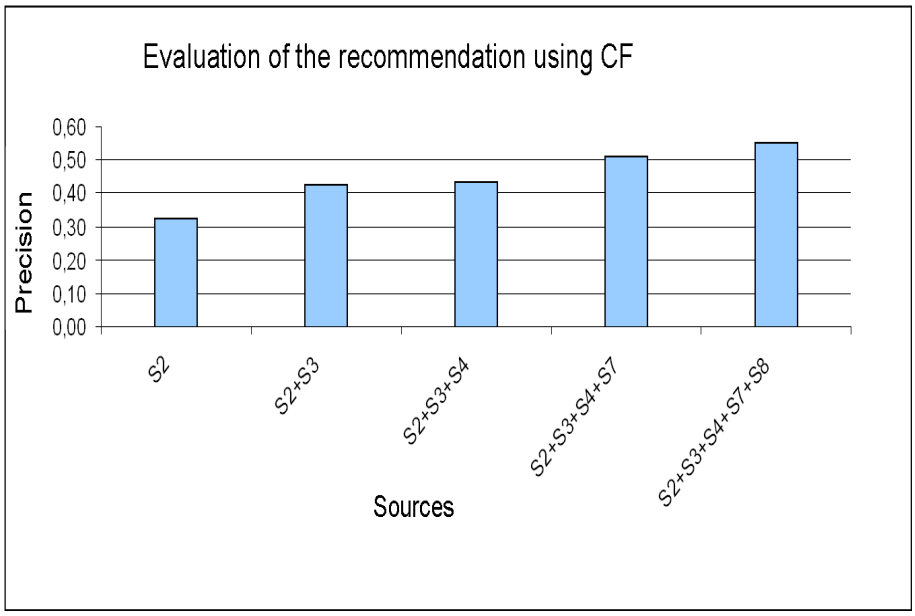


Figure 3.14: Precision of recommendations using CF only with the selected sources with the methodology

---

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

The objective of this case study is to evaluate the methodology used to make recommendations with the sources selected by the **ACQUAINT** methodology. In order to evaluate the precision of the recommendations, we divide the data set into a Training set and a Test set. 80% of the data were used to obtain the preferences of the users and the

Figure 3.15: Information from Amazon.com

remaining 20% for Testing.

The case study has been divided into three experiments using the same sources with different characteristics because the information contained in them had changed.

Experiment 1

The first step in selecting sources of information is obtaining the characteristics of each of them. Table 3.10 shows the values of each one of the characteristics obtained from applying the equations defined in Section 3.2.1. In this table the relevance of each one of the sources is also shown.

Completeness was obtained taking into account the quantity of users in source F1 present in other sources. With the purpose of measuring **diversity**, the users were grouped according to the area they lived in (this information exists on Amazon.com). To calculate **frequency** four categories were defined:

Category f_1 : 1 - 50 interactions

	F1	**F2**	**F3**	**F4**
Completeness	1,00	0,64	0,68	0,24
Diversity	0,50	0,3	0,31	0,22
Frequency	0,63	0,37	0,27	0,32
Timeliness	0,79	0,33	0,26	0,24
Relevant Attributes	1,00	0,9	0,4	0,7
R(s)	0,39	0,25	0,19	0,17

Table 3.10: Experiment 1 - Characteristics and relevance value (R) of sources

Category f_2 : 51 - 100 interactions

Category f_3 : 101 - 150 interactions

Category f_4 : + 151 interactions

To calculate **timeliness** three categories of time were defined:

Category p_1 : 01/01/2000 - 31/12/2001

Category p_2 : 01/01/2002 - 31/12/2003

Category p_3 : 01/01/2004 - 31/12/2006

Fifty recommendations were made. Each one was made with information from sources selected by the *ACQUAINT* methodology according to their relevance value **R** and trust value **T**. The R values are shown in Table 3.10. These values indicate that F1 and F2 are the most relevant. Next the result of three of the 50 recommendations made with information from the sources selected by *ACQUAINT* is shown.

Three interactions performed to make recommendations			
		Iteration: 1	
	F2	**F3**	**F4**
Relevance	0.50	0.38	0.34
Trust	0.50	0.50	0.50
Selected sources:	F2, F3		
Precision of the recommendations:	0.58		
		Iteration: 2	
	F2	**F3**	**F4**
Relevance	0.50	0.38	0.34
Trust	0.66	0.66	0.50
Selected sources:	F2, F3		
Precision of the recommendations:	0.58		
		Iteration: 3	
	F2	**F3**	**F4**
Relevance	0.50	0.38	0.34
Trust	0.75	0.75	0.50
Selected sources:	F2, F3, F4		
Precision of the recommendations:	0.52		

The sources selected in the first recommendation were **F2** and **F3**. The trust of each of them is 0.5. As this was the first time these sources had been used, there was no information about whether they are reliable or not. The recommendations were made with information from F1 plus information from F2 and F3.

$$Recommendations(F1 \longleftarrow (F2 + F3))$$

Their precision was evaluated using equation 3.12 and the value obtained was **0.58**. For this case study recommendations with a value > 0.50 were considered successful, so this the firs recommendation made using information from sources F1$\longleftarrow$ (F2 + F3) was successful, this information is taking account in the next iteration of the algorithm to calculate the Trust of these sources. This means, with this result the trust value of F2 and F3 will be higher and therefore selected in the next recommendation.

The results of the 50 recommendations made with information from the selected sources

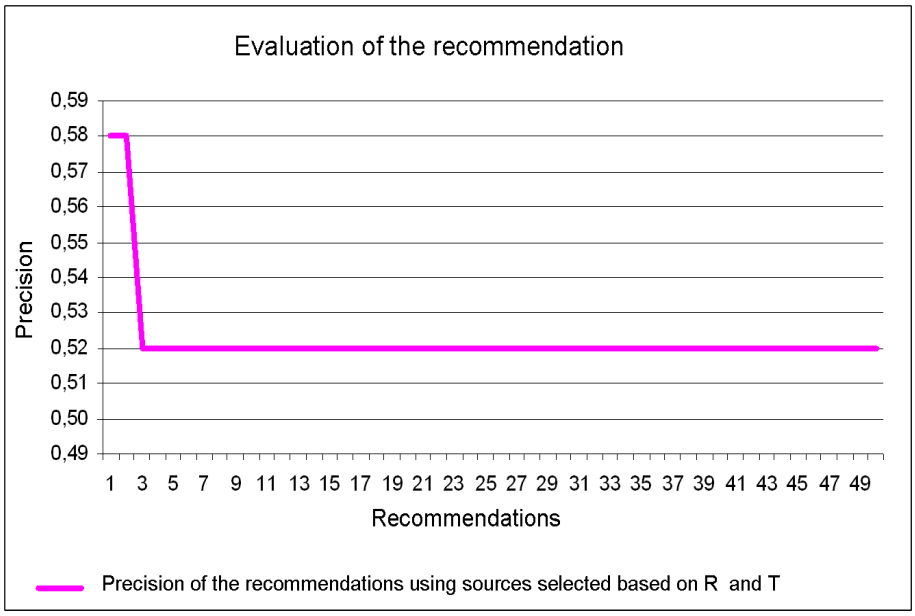


Figure 3.16: Recommendation results using the selected sources based on their relevance and trust

based on their relevance and trust are presented in Figure 3.16. As can be seen, the recommendations made in all cases except the first two recommendations have resulted in a precision value of 0.52.

To evaluate the effectiveness of the *ACQUAINT* methodology, recommendations were made with the same sources of information selected according to the following criteria:

1. Recommendations with the sources selected based on the relevance value (R) obtained from the intrinsic characteristics.

2. Recommendations with sources selected based on the measure of trust (T).

3. Recommendations only with source F1.

4. Recommendations with all the sources of information, F1, F2, F3 and F4.

5. Recommendations made with the optimal combination of sources.

The precision of recommendations made with sources selected based on their relevance value (R) are shown in Figure 3.17. Many of the recommendations made resulted in a precision value < 0.5.

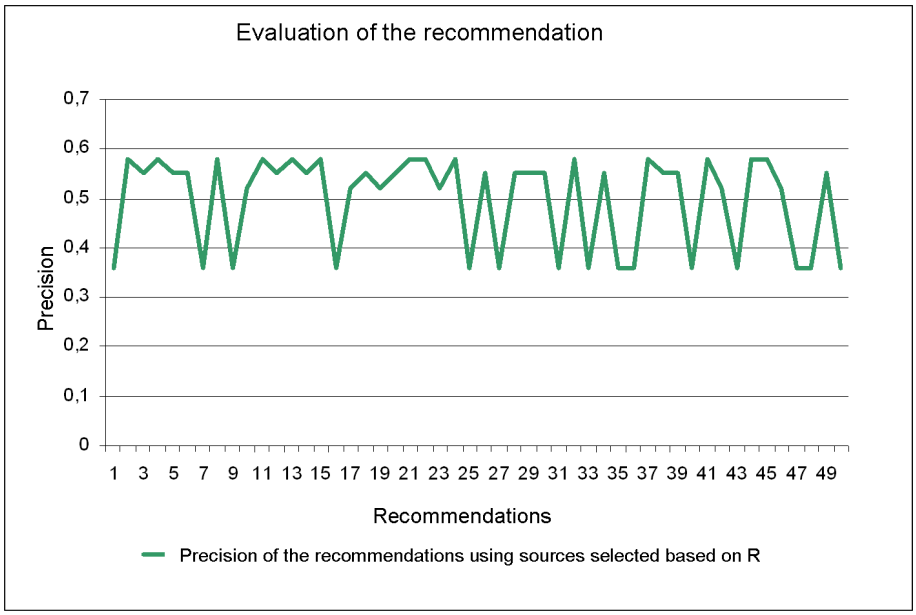


Figure 3.17: Recommendation results using the selected sources based on their relevance

Figure 3.18 shows the precision of recommendations made with sources selected based on their trust value (T). The precision values indicate that more precise recommendations were made with sources selected according to T than with recommendations based on R.

Figure 3.19 includes three other criteria: the precision obtained when making recommendations with information only from source F1; the precision of the recommendations made with information from all the sources of information; and the precision achieved with the optimal combination of sources . The optimal combination of sources in each of the iterations was found making recommendations, adding information from the sources and taking the combination with the highest precision.

As can be observed in the graph, recommendations made only with information from source F1 have a lower precision value than those made with the addition of information from other sources. However, adding information from all the sources is not optimal; the ideal would be to find the optimal combination of sources to make recommendations with better precision. In environments where the number of sources available is excessive, performing an exhaustive search for the optimal combination adds more complexity to the system. Using the source selection algorithm based on its relevance and trust, recommendations whose precision approaches that obtained making recommendations with the optimal combination of sources can be made.

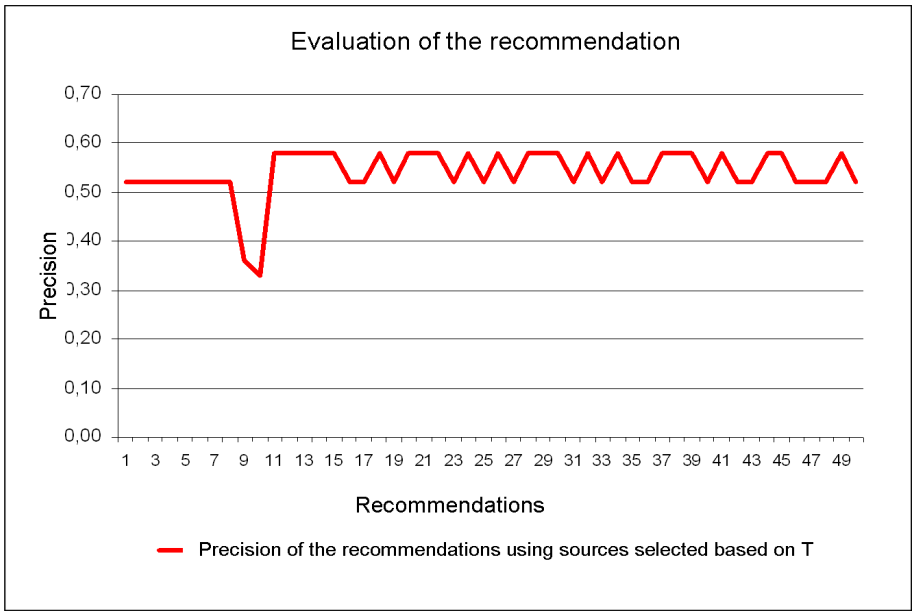


Figure 3.18: Experiment 1 - Recommendation results using the selected sources based on their trust

Experiment 2

Fifty more recommendations were made with the same sources but with different characteristics. As the information contained in them has changed, the values of the characteristics in each one of them have also changed. Table 3.12 shows the values of each of the characteristics obtained after applying the equations defined in Section 3.2.1. This table also shows the relevance of each of the sources.

Fifty recommendations were made. Each of them was made with information from sources selected by the *ACQUAINT* methodology. The sources were selected according to

	F1	**F2**	**F3**	**F4**
Completeness	1,00	0,56	0,79	0,35
Diversity	0,50	0,30	0,65	0,56
Frequency	0,63	0,90	0,40	0,70
Timeliness	0,79	0,33	0,65	0,45
Relevant Attributes	1,00	0,40	0,30	0,30
R(s)	0,39	0,25	0,28	0,24

Table 3.11: Experiment 2 - Characteristics and relevance value (R) of sources

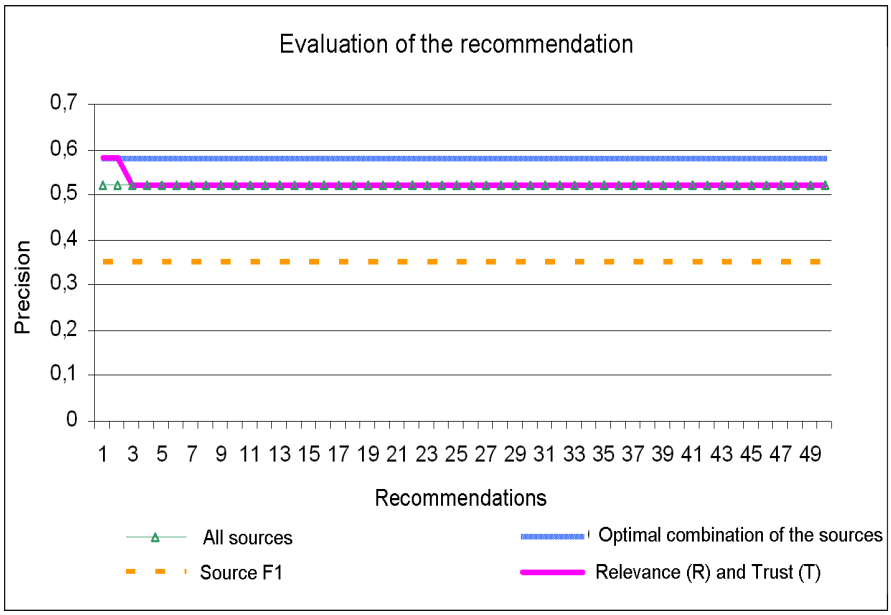


Figure 3.19: Experiment 1 - Evaluation of the results

their relevance value **R** and their trust value **T**. The results of these 50 recommendations are presented in Figure 3.20.

As can be observed in the graph, this set of values of characteristics produced better results in terms of the precision of the recommendations than with the characteristics of Experiment 1. This is also demonstrated in Figure 3.21 which shows the precision of the recommendations made with the sources selected based on their relevance value (R). Many of the recommendations made resulted in a precision value > 0.5.

Figure 3.22 shows the precision of recommendations made with sources selected based on their trust value (T). The precision of most of the recommendations is above the value of $0,5$.

The graph of Figure 3.23 is used to compare the precision of recommendations made with sources selected by ACQUAINT and the precision of recommendations made only with 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 recommendations made only with information from source F1 produce less precise results than if the recommendations are made with the addition of information from various 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 sources selected based on

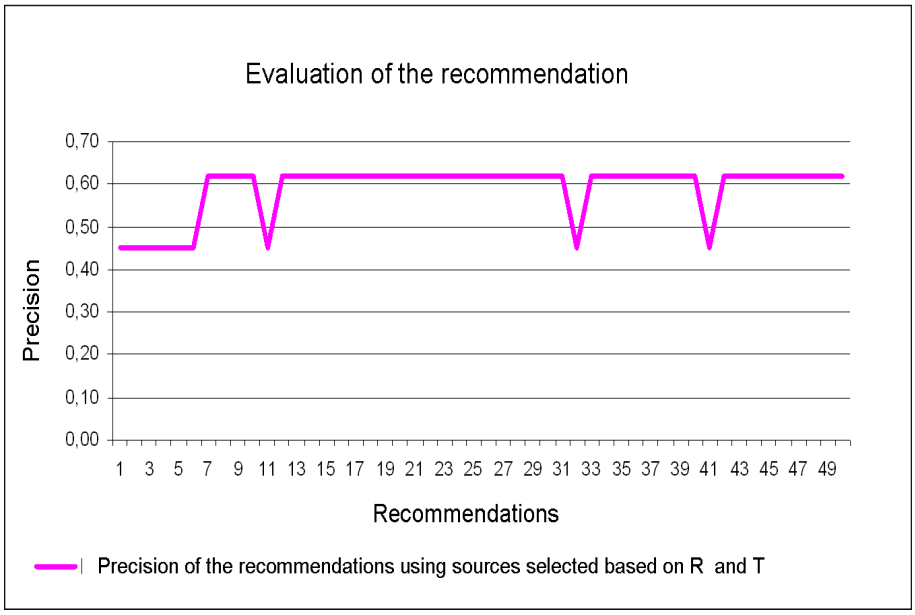


Figure 3.20: Experiment 2- Recommendation results using the selected sources based on their relevance and trust

R and T closely approaches or is equal to that obtained with the optimal combination of sources.

Experiment 3

In experiment 3 another set of characteristics from the same sources has been tested. Table 3.12 shows the values for each of the characteristics obtained when applying the equations defined in Section 3.2.1. The table also shows the relevance of each of these sources.

Fifty recommendations were made. Each one of them was made with information from

	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 3.12: Experiment 3 - Characteristics and relevance value (R) of sources

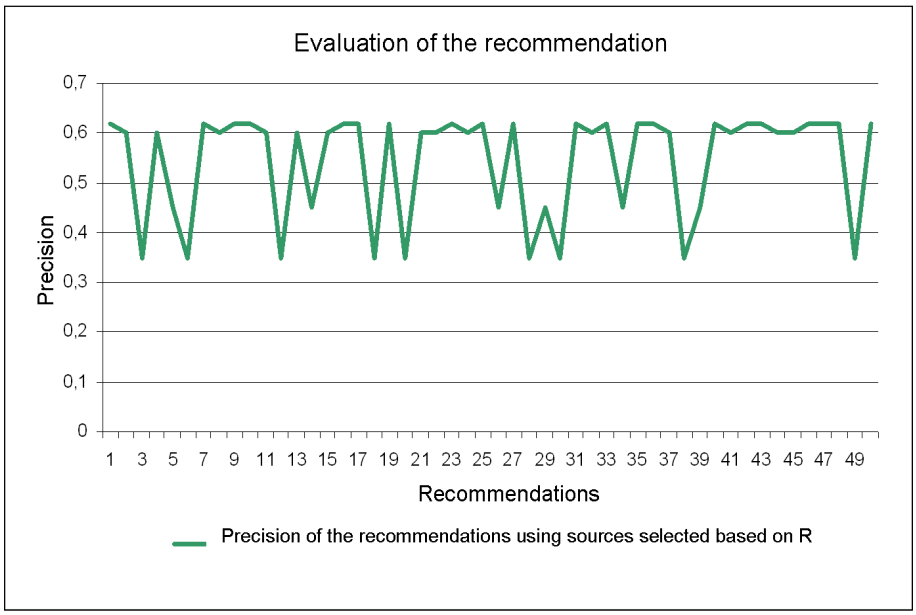


Figure 3.21: Experiment 2- Recommendation results using the selected sources based on their relevance

sources selected by the *ACQUAINT* methodology based on their relevance value **R** and their trust value **T**. The results of these 50 recommendations are presented in Figure 3.24.

The precision of the recommendations made with the sources selected based on their relevance value (R) is shown in Figure 3.25. With this set of characteristics all the recommendations obtained a very high precision value ($> 0,5$), although the curve shows unstable behaviour. This behaviour is also evident in Figure 3.26 where the precision of the recommendations made with the sources selected based on their trust value (T) is presented.

The graph in Figure 3.27 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.

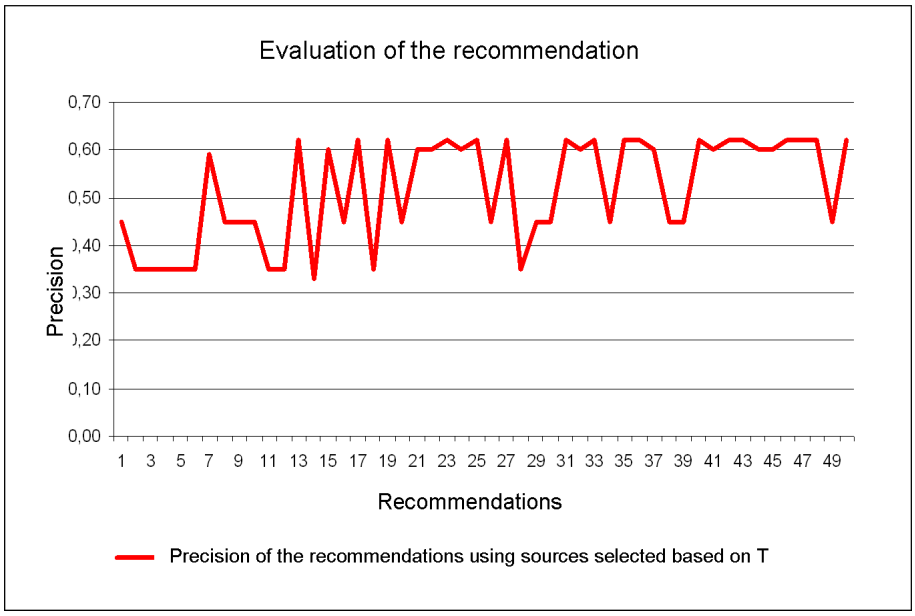


Figure 3.22: Experiment 2- Recommendation results using the selected sources based on their trust

In the three experiments and with different values of measurement, it has been proven that the precision of recommendations made with a combination of information from sources based on R and T are equal to or come close enough to the maximal precision obtained with the information obtained from the optimal combination of available sources.

3.3.3 Conclusion

In Recommender field has 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 [Stuart et al., 2002].

The search for and selection of relevant and trustworthy sources that allow us to get more user information to make better recommendations is one of the subjects analyzed in this chapter. A new methodology has been proposed for that purpose. The new methodology denominated "ACQUAINT" aims to improve the precision of recommendations and

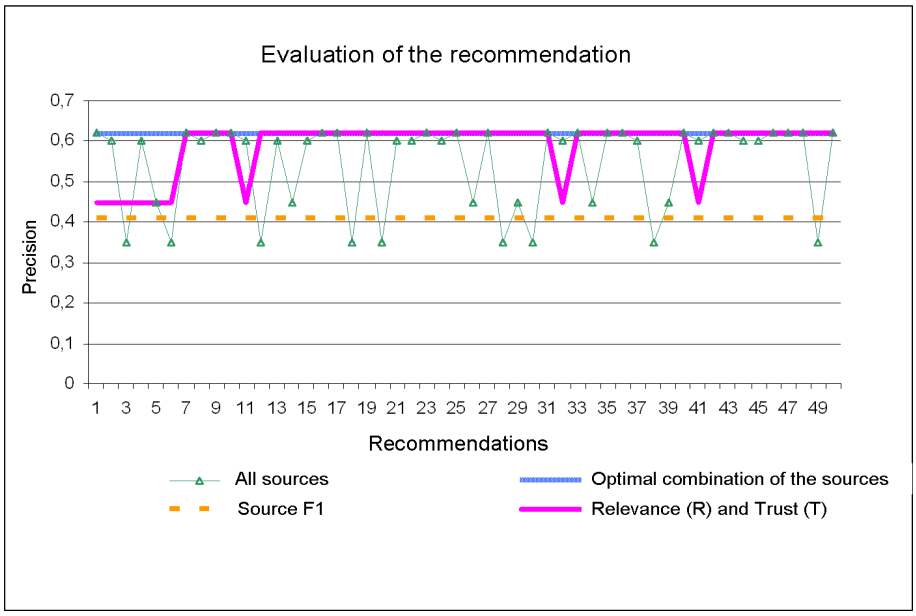


Figure 3.23: Experiment 2 - Evaluation of the results

to solve the problem of lack user's information. Many existing approaches use explicit information that the user gives about his/her preferences and interests. ACQUAINT shows a new way to obtain this information which uses characteristics of the source to know whether it can provide with such information, or not. The characteristics represent relevant information contained in the sources such as the quantity of users in the source, demographic information, frequency of the interaction, timeliness of the information and number of relevant attributes. A relevance value has been defined based on these characteristics to select the most relevant source. The user preferences are established from the selected sources to make the recommendation.

First, the methodology has been used with three recommender methods: RFM, CBF and FC obtaining good results. Data from retail domain was used in this case study. The results of the recommendation made by using the three methods have been analysed. The precision of the recommendations based on selected sources by ACQUAINT is better than when used all available information sources. This confirms that the characteristics are representative of the information contained in sources and are good criteria to choose a source. However, when we have tested the methodology in the second study case, we have noticed that we need other criteria to select sources because the precision of recommendations in this case is no so good. Some time the sources selected based on their relevance result in lower precision. So, the selection process is needs to know the result of the past recommendation made by using a source, this information allow to know

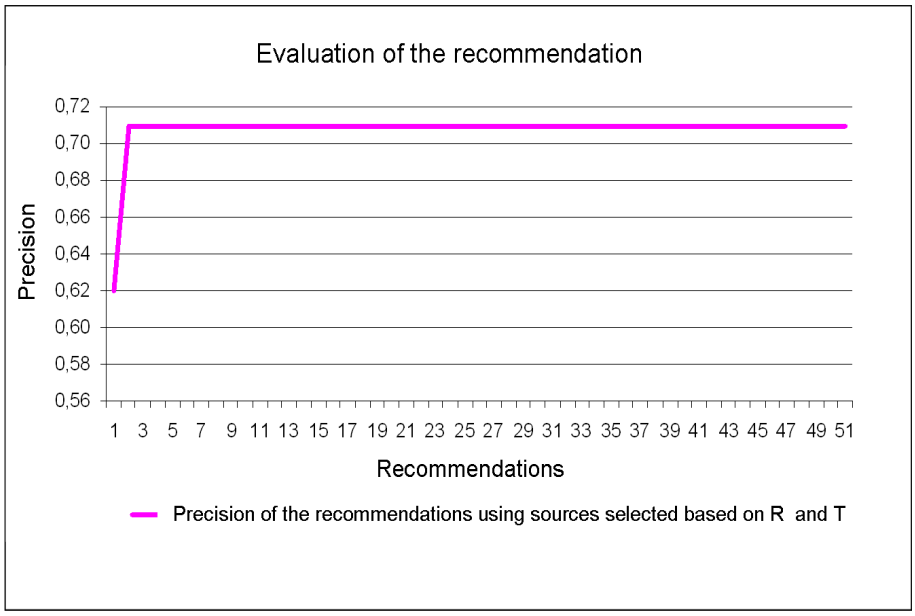


Figure 3.24: Experiment 3- Recommendation results using the selected sources based on their relevance and trust

if a source is trustworthy or not to be used in a recommendation. A trust measure has been applied to obtain the reliability of sources. The results obtained in this case study confirm that when using information from relevant and trustworthy sources, the obtained precision is equal or become closer to the highest precision.

In short, two case studies showing the performance of the methodology when using relevant and trustworthy information sources in recommender systems have been presented, obtaining good results. However, this chapter has left some questions that the following chapter covers. Our proposed methodology has been applied to obtain user information stored in structured sources. Next chapter provides a technique to retrieve unstructured user's information available on web pages. Once this information is, the ACQUAINT methodology can be applied.

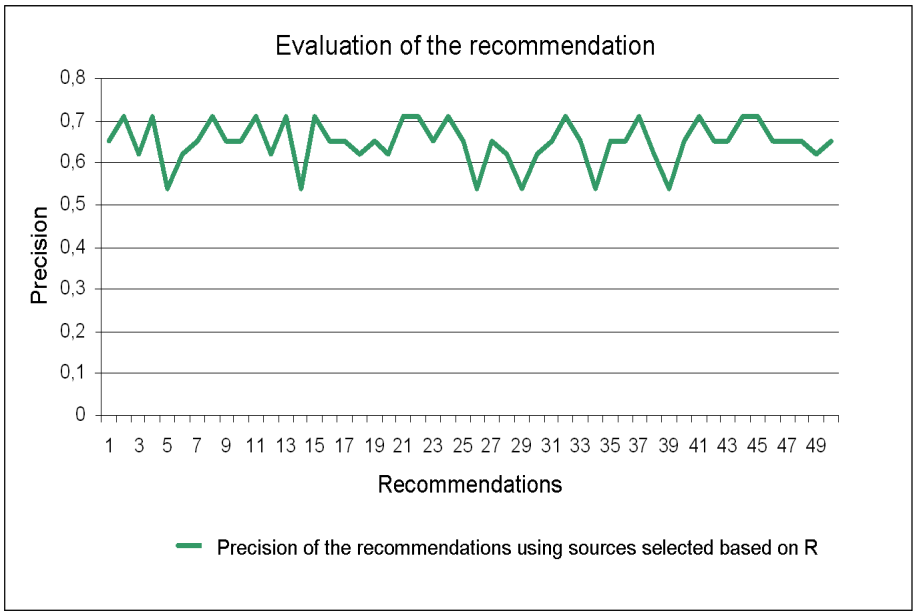


Figure 3.25: Experiment 3- Recommendation results using the selected sources based on their relevance

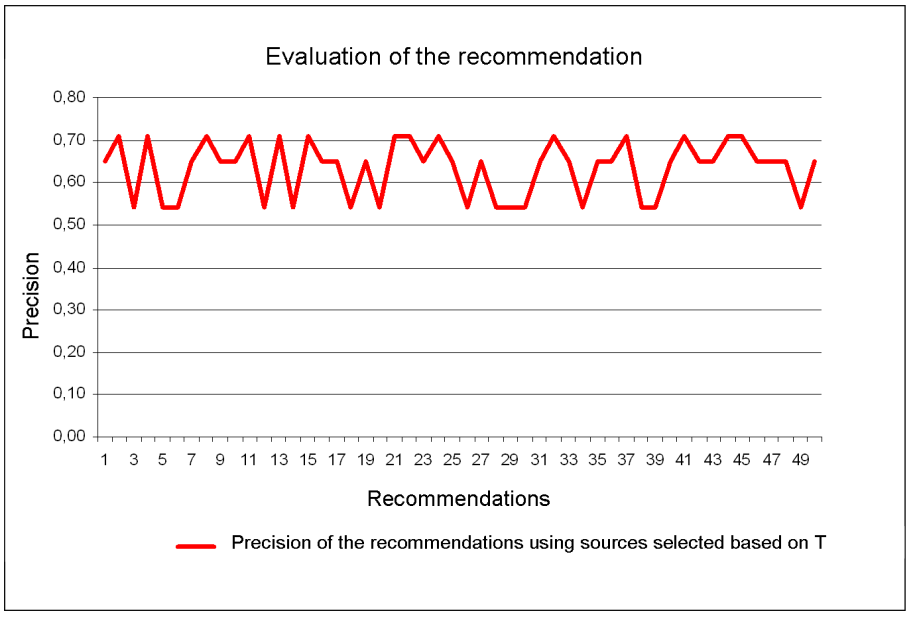


Figure 3.26: Experiment 3- Recommendation results using the selected sources based on their trust

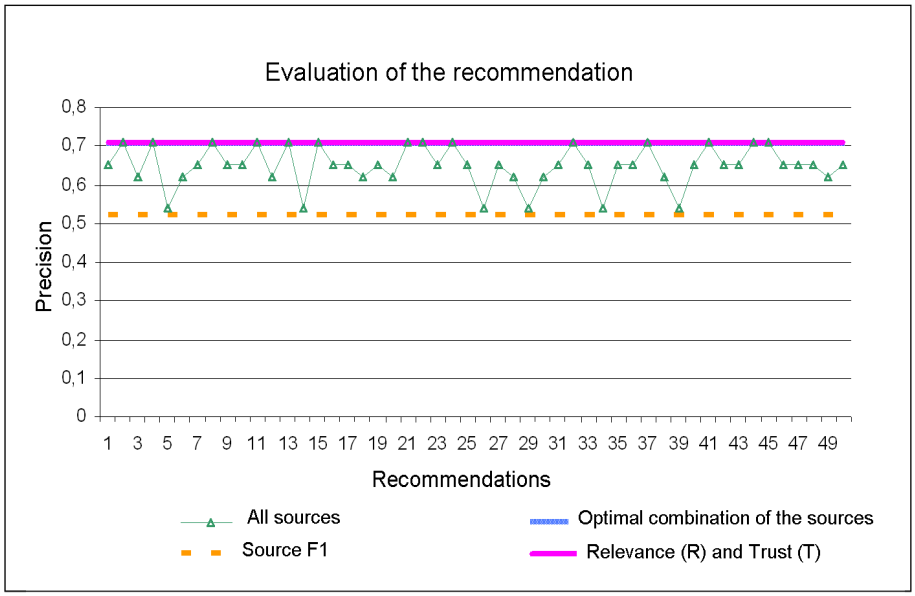


Figure 3.27: Experiment 3 - Evaluation of the results

Chapter 4

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 chapter, 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 (i.e. such data is available on the Internet) 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 webpages

4.1 Introduction

Rather than completing forms with rating values, many consumers prefer to use natural language and express their opinions about the product in a free text form, similar to a conversation with a friend. In the online world, there are several popular ways for consumers to exchange their experiences with a product [Dellarocas, 2003] [Curien et al., 2006] - product review forums, virtual community logs, product discussion boards and e-commerce sites. There is growing evidence that such forums inform and influence consumers' purchase decisions [Chevalier and Mayzlin, 2003] [Senecal and Nantel, 2004]. Decision-makers use advice from an expert either to increase their decision accuracy or to reduce their effort expenditure [Schrah et al., 2006]. Despite the importance and value of such information, there is no comprehensive mechanism that formalizes:

- The process of selection and retrieval of opinions, and
- The utilization of retrieved opinions.

Part of the problem resides in the complexity of extracting information from text data and converting it into product recommendation. Adomavicius provided an overview of recent developments in recommender systems [Adomavicius, 2005]. According to this review, the recommender systems that utilize review comments using text mining techniques are yet to be developed. Ricci [Ricci and Wietsma, 2006] [Wietsma and Ricci, 2005] proposed to utilize review comments for product description and user behaviour study. He believed the review comments could be widely used in recommender systems and result in better recommendations. Ricci and Wietsma [Ricci and Wietsma, 2006] [Wietsma and Ricci, 2005] so far seem to be the only recommender system that integrates reviews in the recommendation process. The authors use product reviews in the product selection decision process for a mobile recommender system. They employ social-filtering algorithms to extract knowledge from the reviews. The main aim of their system is to improve the explanation of the recommendation providing the relevant reviews of users with similar tastes. The reviews are used to give explanations of the recommendations, but they are not used to make recommendations. This chapter addresses the problem of utilisation of consumer opinion about products, expressed online in a free text form in order to generate product recommendations. Figure 4.1 shows the overall process structure of the proposed recommender system. The realisation of this process structure requires the completion of several tasks, including:

- The development of an information representation structure using the quality/feature ontology.
- The implementation of a text mining algorithm for mapping automatically the information from the reviews into the information structure of the ontology.
- The development of a ranking mechanism that computes the rating of a product using the information from the consumer reviews stored in the ontology.
- The development of a recommender mechanism, which computes recommendations in response to a user request.

The process collects relevant product consumer reviews and builds a collection of relevant reviews. Technically, the procedure for collection of product reviews follows the algorithms for automated news extraction from news sites developed in [Zhang and Simoff, 2006].

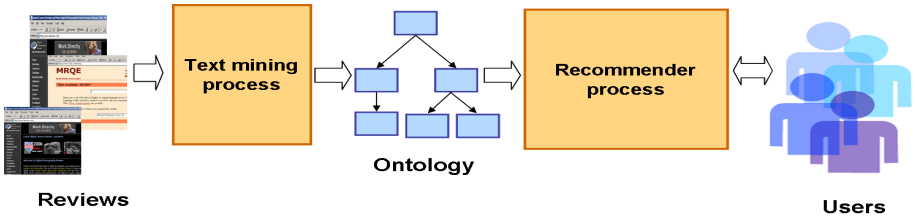


Figure 4.1: Process to integrate the information from different sources using ontology mappings

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. The text mining process maps the review comments into the ontology. 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. The major contribution of this work is the overall framework for automating the utilisation of consumer reviews, and its individual components. Where possible it utilises existing algorithms (for example, in the text mining process), as the goal of the reported work is to demonstrate the strengths of the overall approach.

4.2 Representation of the consumer reviews

The goal of this step is to find a suitable tool for extracting the information contained in the text and converting it into structured data. Identifying an appropriate representation of consumer opinions that can be used in the system is a key problem. One way to convert these opinions to a structured form is to use translation ontology, which is typically used as a form of knowledge representation and sharing. Review comments are firstly mapped into ontologies to allow the ranking calculations become possible. In this application, the

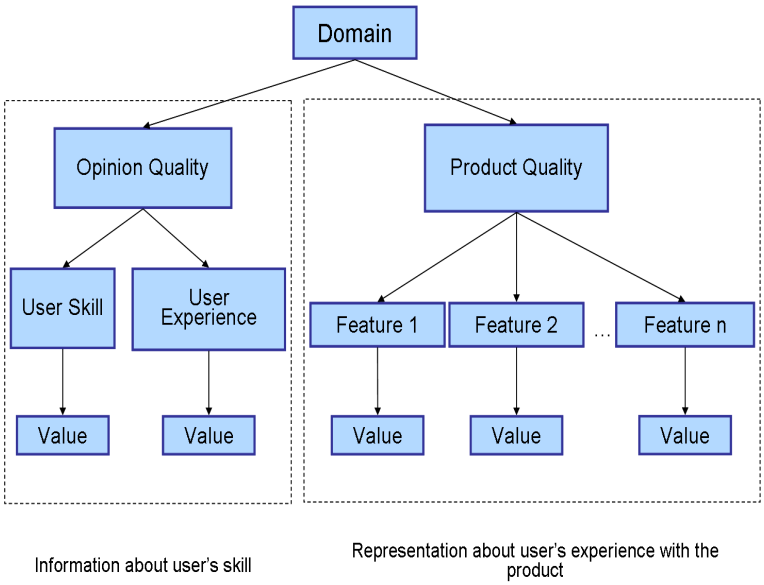


Figure 4.2: Structure of the ontology used in the recommendation from consumer opinions applications

ontology contains two main parts: Opinion Quality and Product Quality, which summarize the consumer skill level and the consumer experience with the product in the review, respectively. Figure 4.2 shows the general structure of the ontology. The Opinion Quality includes several variables to measure the opinion provider's expertise in the product. The Product Quality represents the opinion provider's valuation of the product features, which is highly domain specific.

4.3 Mapping review comments into ontology instances

Ontology provides a controlled vocabulary and relationship to describe the consumer skill level and the consumer experience with the product in the review comment in the system. The classes and relationships in the ontology are only required to be defined once and can be used until the products have new features. Each review comment is represented as an ontology instance. The mapping of the ontology instances in manual way is a tedious and time consuming job. This section describes a methodology to create ontology instances automatically using text mining techniques. As the ontology has been defined, the mapping process includes the identification of both the classes involved in the instance

and their attributes. The mapping process is composed by two steps:

1. Sentence selection and classification: This step identifies the class attributes. In the user valuation from the text data, each feature from the comment is assigned either "Good" or "Bad". Therefore, the sentences in the review are selected and classified into three categories: "Good" comments, "Bad" comments and "Quality". "Quality" category contains the sentences that indicate the opinion quality.

2. Concept identification: Once the relevant sentences are selected, this step identifies the classes that the selected sentences belong to. The concepts which implicated in the sentences determine the classes in the ontology, are identified by related words used as synonymous of the concept.

The next sections detail both steps of the mapping process.

4.3.1 Sentence selection and classification

Under the text mining paradigm, each sentence is treated as a document in this application. To group review sentences into "Good", "Bad" and "Quality", shallow parser was firstly considered as an analyzer tool. However, most of parsers give complicate and incorrect results. Furthermore, each document is very short. Classification algorithms based on term frequencies do not provide satisfaction results either. Therefore, rule based classification techniques are employed. As described in previous section, three categories have been defined to classify the sentences: "Quality", "Good" and "Bad". "Quality" category groups those sentences that contain information about the skill of the consumer. "Good" category groups those sentences that contain information about some features that consumer has valuated as the strengths of the product. "Bad" category groups those sentences that contain information about some features that the consumer considers as weaknesses of the product.

At this stage, the work has been focused on providing the overall concept of utilizing text mining for automatic mapping of review comments into ontology instances. Hence, we employed an off-the-shelf text mining kits. The Text-Miner Software Kit (TMSK) and the Rule Induction Kit for Text (RIKTEXT) have been used to obtain the classification rule sets [Weiss et al., 2004]. TMSK generates a dictionary from a set of documents (sentences in our case) and converts a set of sentences into sparse vectors based on the dictionary. The dictionary and the vectors representing each category are used by RIKTEXT for learning a classifier. RIKTEXT is a complete software package for learning decision rules

from document collections. The rules are induced automatically from data. The output is a rule set of classification of "Good" "Bad" and "Quality" category from training data. Opinions of 68 reviews about the digital camera: Canon PowerShot SD500 (Digital IXUS 700) from www.dpreview.com web have been used to create the training data set. Each sentence of each review is treated as a document. 195 sentences have been obtained for the "Good" category, 127 sentences for "Bad" category and 47 sentences for "Quality" category. The available data have been spited into training and tests portions. Test cases are selected randomly in RIKTEXT and we specified how many cases should be used for testing. We choose two-thirds of the available cases for training and the rest for testing. The results are presented in Table 4.1

As can be seen, it displays a number of rule sets. Each rule set is numbered under the column "RSet". A "*" delineates the rule set with the minimum error rate. A "**" indicates the best rule set according to the error rate and simplicity. "Rules" is the number of rules in the rule set. "Vars" indicates the total number of conjuncts in the left-hand-side of the rules. The column "Train Err" gives the error-rate of the rule sets on the training data. "Test Err" is an error-rate estimate and Test SD is the standard deviation of the estimate. "Mean Var" is the average number of variables of the resampled rule set that approximates in size the rule set for the full data. "Err/Var" gives an indication of the quality of the solution. The chosen rules are those that have minimum error rate or are very close to the minimum but may be simpler than the minimum (**). Precision, recall and f-measure obtained from training and test cases are shown at the end of the table. Tables 4.2 and 4.3 show the rule sets obtained from classification of "Bad" and "Quality" categories, respectively.

4.3.2 Concept identification

Once the sentences have been classified into one of the categories, the concept (class) in the ontology implicated in the sentence is needed to be identified. Each concept in the ontology contains a label name and a related word list. A related word list of a concept contains vocabulary (a set of keywords) through which the concept can be matched with one sentence in the comments. Table 4.4 shows the related word list defined for this problem. For example related word for the concept "Comparison" found in reviews can be "compare, compared, equal, same, etc".

Table of pruned rule sets
(* = minimum error; ** = within 0-SE of minimum error)

RSet	Rules	Vars	Train Err	Test Err	Test SD	MeanVar	Err/Var
1	32	34	0.1696	0.3171	0.0420	0.0	0.00
2**	30	31	0.1736	0.3171	0.0420	0.0	0.67
3	29	29	0.1795	0.3333	0.0425	0.0	2.00
4	24	24	0.2091	0.3659	0.0434	0.0	3.00
5	22	22	0.2249	0.3496	0.0430	0.0	4.00
6	1	1	0.2308	0.4715	0.0450	0.0	0.14

Random test cases: 123 (33.3 %)

Selected rule set

1. fast >= 1 − − > gd
2. results −− > gd
3. good >= 1 − − > gd
4. nice −− > gd
5. overall −− > gd
6. pocketable −− > gd
7. great −− > gd
8. use >= 1 − − > gd
9. underwater >= 2 − − > gd
10. problems AND no −− > gd
11. function −− > *gd*
12. like >= 1 − − > *gd*
13. better >= 1 − − > *gd*
14. speed −− > *gd*
15. compact −− > *gd*
16. pocket −− > *gd*
17. ps >= 1 − − > *gd*
18. love −− > *gd*
19. user −− > *gd*
20. sd500 −− > *gd*
21. quality >= 1 − − > gd
22. small −− > *gd*
23. able −− > *gd*
24. sd −− > *gd*
25. far −− > *gd*
26. photos −− > *gd*
27. shots −− > *gd*
28. really >= 1 − − > *gd*
29. mode >= 1 − − > *gd*
30. [TRUE] −− > *gd*

Statistics results

Training Cases: precision:71.6049 recall:89.2308 f-measure:79.4521
Test Cases: precision:67.5676 recall:76.9231 f-measure:71.9424

Table 4.1: Table of pruned rule sets for "Good Category"

Table of pruned rule sets
(* = minimum error; ** = within 0-SE of minimum error)

RSet	Rules	Vars	Train Err	Test Err	Test SD	MeanVar	Err/Var
1	25	28	0.1731	0.3710	0.0434	0.0	0.00
2	25	27	0.1755	0.3710	0.0434	0.0	1.00
3	24	25	0.1779	0.3468	0.0427	0.0	1.50
4	20	20	0.2043	0.3065	0.0414	0.0	2.20
5	14	14	0.2500	0.2742	0.0401	0.0	3.17
6	12	12	0.2740	0.2661	0.0397	0.0	5.00
7	10	10	0.3029	0.2661	0.0397	0.0	6.00
8**	9	9	0.3221	0.2581	0.0393	0.0	8.00
9	8	8	0.3438	0.2742	0.0401	0.0	9.00
10	7	7	0.3678	0.2823	0.0404	0.0	10.00
11	1	1	0.3870	0.6532	0.0427	0.0	1.33

Random test cases: 124 (33.3 %)

Selected rule set

1. purple >= 1 − − > bd
2. iso >= 1 − − > bd
3. manual −− > *bd*
4. problem >= 1 − − > bd
5. battery >= 1 − − > bd
6. not −− > bd
7. lcd −− > bd
8. no −− > bd
9. [TRUE] −− > *bd*

Statistics results

Training Cases: precision:73.0159 recall:54.1176 f-measure:62.1622
Test Cases: precision:70.3704 recall:44.1860 f-measure:54.2857

Table 4.2: Table of pruned rule sets for "Bad Category"

Table of pruned rule sets
(* = minimum error; ** = within 0-SE of minimum error)

RSet	Rules	Vars	Train Err	Test Err	Test SD	MeanVar	Err/Var
1	20	34	0.0422	0.1371	0.0309	0.0	0.00
2	17	28	0.0552	0.1290	0.0301	0.0	0.67
3	13	18	0.0779	0.1774	0.0343	0.0	0.70
4	9	9	0.1104	0.1371	0.0309	0.0	1.22
5**	7	7	0.1299	0.1129	0.0284	0.0	3.00
6	6	6	0.1396	0.1210	0.0293	0.0	3.00
7	5	5	0.1558	0.1371	0.0309	0.0	5.00
8	4	4	0.1818	0.1371	0.0309	0.0	8.00
9	3	3	0.2143	0.1210	0.0293	0.0	10.00
10	2	2	0.2500	0.1290	0.0301	0.0	11.00
11	1	1	0.3019	0.1290	0.0301	0.0	16.00

Random test cases: 124 (33.3 %)

Selected rule set

1. own $--> ql$
2. bought $--> ql$
3. digital $--> ql$
4. powershot $--> ql$
5. cameras $>= 1 --> ql$
6. sony $>= 1 --> ql$
7. [TRUE] $-->$ ql

Statistics results

Training Cases: precision:74.0741 recall:64.5161 f-measure:68.9655
Test Cases: precision:60.0000 recall:37.5000 f-measure:46.1538

Table 4.3: Table of pruned rule sets for "Quality Category"

Features	**Synonymous**
autosetting	auto setting
battery	battery
button	butoons
button	button
camera	camera
camera	feature
camera	features
camera	functions
card	card
card	cf-card
card	sd
card	sd-card
color balance	color balance
color balance	colour balance
color balance	consistent balance
color balance	white balance
comparison	compare
comparison	compared
comparison	comparison
comparison	equal
comparison	same
.....	.....

Table 4.4: Related word for each concept in the ontology

4.3.3 Notes on implementation

The ontology was created manually to ensure that it is complete and well-defined. However, mapping the ontology instances from review comments is fully automatic after the training process. Similar to other classification applications, collection and labelling training examples for sentence classification are manual processes. Once the system has been trained, it automatically classifies a sentence into either "Good" or "Bad" category. In the concept identification step, the synonym database was created manually. The concept is identified automatically if a keyword in the database is identified in the sentence.

4.4 Recommendation using consumer's reviews

The review comments are firstly mapped into an ontology to make the ranking calculations possible. Since it has been explained in the Section 2, the ontology contains two main parts: Opinion Quality, which summarise the consumer skill level and the consumer experience with the product in the review, respectively. A set of measures: Opinion Quality (OQ), Feature Quality (FQ), Overall Feature Quality (OFQ) and Overall Assessment (OA) are computed based on the data in the ontology. Opinion Quality (OQ) is defined to evaluate the weighting value of opinions according to the opinion provider's expertise. Overall Feature Quality (OFQ) is the global valuation of the feature from all reviews, which is calculated from the Feature Quality (FQ) value of individual comment. Overall Assessment (OA) provides a final score of the product based on the valuation of each feature. The recommendation in response to a user request is given based on these measurements. The recommendation is made based on the review comments that are summarised by an Overall Feature Quality (OFQ) value for each feature. In the next sections are detailed the calculation of these measures.

4.4.1 Rating the consumer skill level

The review comments were given by people with diverse experience and skill levels. In general, people who have longer history of using the product can provide more professional opinions. Therefore, these diverse opinions should not be treated equally. The opinions from more experienced people should be taken in account to a greater extent than those from people with little knowledge of the product. Opinion Quality (OQ) is defined to evaluate the weighting value of opinions according to the opinion providers' expertise.

Definition 1. Opinion Quality (OQ) is the sum of the weight w_j, given for each

variable j representing the skills and experiences of consumer i divided by the number of variables representing the information about consumer's skill and expertise provided in the ontology.

$$OQ_i = \frac{\sum_j^n w_{ij}}{n} \tag{4.1}$$

The Opinion Quality is calculated by the values stored in the corresponding part of the ontology. An Opinion Quality value is calculated for each piece of comment.

4.4.2 Product quality ranking

The product is ranked according to the consumer comments for each feature. Due to the difficulties of quantification of user valuation from texture data, each feature from the comment can only be assigned either "Good" or "Bad", which is calculated as "1" or "-1" respectively. For each feature, a Feature Quality is calculated, which is a function of consumer valuation and Opinion Quality.

Definition 2. Feature Quality (FQ): The quality value for each feature f of the product in a review is the rating multiply by the Opinion Quality value of the consumer

$$FQ_f = r * OQ_i \tag{4.2}$$

4.4.3 Selection of the relevant opinion and making recommendations in response to a user request

When a user requests the evaluation of a particular product based on certain features, the Overall Feature Quality is calculated from the reviews that contain the valuation of this feature.

Definition 3. Overall Feature Quality (OFQ) is the global valuation of the feature from all reviews, which is calculated by the average value of Feature Quality.

$$OFQ_f = \frac{\sum(Scaling factor * FQ)}{Number of Opinions} \tag{4.3}$$

Here Scaling Factor is used to do the minor adjustment of the user valuation, which can be set to:

$$Scalingfactor = \frac{1}{n} \tag{4.4}$$

n is the number of all the features rated by the consumer. Each review rated different number of features so n could be different. To provide the user with a comprehensive valuation of the product quality in related to the requested features, an Overall Assessment score is defined.

Definition 4. Overall Assessment (OA) provides a final score of the product based on the valuation of each feature. It is calculated as the sum of all OFQ (calculated by equation 4.3) multiplied by the Importance Index.

$$OA = \sum OFQ * ImportanceIndex \tag{4.5}$$

The Importance Index measures the different influence of the features to consumer's decision making, which can be assigned in two ways: according to the importance of the feature expressed in the user request or by the frequency that the features have been rated in the consumer reviews.

4.5 Study Case

In this section, we present the different steps that must be considered in order to offer a recommendation about digital camera in response to a user request. Example was conducted using digital cameras. Data from the Digital Photography Review (www.dpreview.com) were chosen.

Each day consumers visit this page to rate and add opinions about different digital cameras. Firstly, we explain how the ontology has been defined and the reviews is mapping into the ontology.

4.5.1 Representation of the consumer reviews - Digital camera ontology

First of all, we define an ontology. In this research, an ontology has been developed for digital camera domain (See Figure 4.3). Each concept in the ontology was obtained analyzing the reviews from the consumers of different digital cameras from www.dpreview.com. Consumers can choose any digital camera and rate it on a scale of half start to four starts.

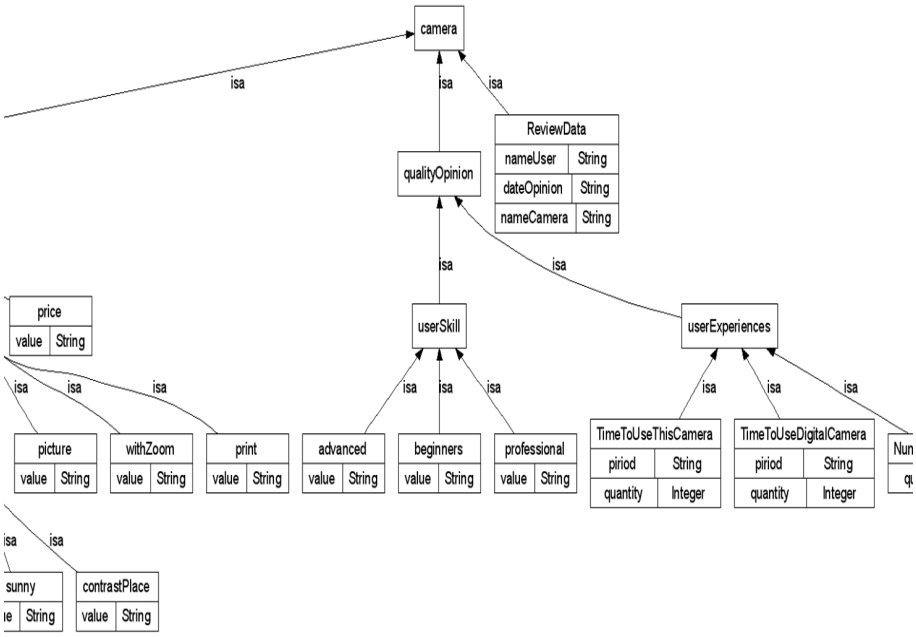


Figure 4.3: Digital camera ontology

They can also write free form text reviews about the camera. For the construction of digital camera reviews ontology, first was made a list of all possible objects necessary to cover given cameras reviews. This possible list should include different digital cameras such as Canon, Sony, etc. Furthermore, different cameras can be qualified by features such as size, zoom, lens, quality picture, etc. This information is represented by the concept "Features". And the different consumer's reviews can be qualified by opinions from beginners, professionals and by the level of expertise using digital cameras.

4.5.2 Mapping a review comment into an ontology

Once the ontology has been defined, it is necessary to match the information of the review with the ontology. We now show a new mapping to map the information into the ontology. Example was conducted using the review shows in Figure 4.4. The next sections describe the classification process applied to the new review.

Canon PowerShot SD630 (nzjss, 16 Apr 06):

You cant take a photo unless you are carrying a camera and this is a good one to carry all the time! I have owned most IXUS cameras since the first - Just traded my Ixus 700 on this one (an Ixus 65 but still have an Ixus 40).The huge screen is excellent for framing shots and showing them off. I like the new low shutter speed showing on the screen and the grid lines. The screen is quite scratch resistant (unlike the 40 or 700 which scratch easily). Images are "good" (not stunning) better handheld in low light than the 700 but still blurry for distant features in landscapes. Gives excellent 6x4 prints on Canon CP330. The smaller 5-way selector button is touch sensitive and shows fancy icons on the screen that get bigger when you touch (not push) that side of the button - nice! The Flash is undersized but works well with closer subjects and has good balance for backlit shots. The bigger strap eye on this camera allows you to fit the strap of your choice. I use a black early Ixus strap that works better with my Jenova black leather Ixus belt case. Interesting to see other camera brands copying the Canon zoom button around the shutter release – works well on the Ixus 65/SD630 with the new low profile buttons. The slim mode switch on top is now needs the left hand. This is my "main" camera - and while there are times when I want a longer or wider lens I would not be draging along the big camera that they are attached to. Recommended - if there was a better camera this size I would have bought it!The Images are not that sharp but better than the 700. No manual ASA select and still no Battery power meter. Not so much room for the right thumb to grip the back but practicing a different finger arrangement holding the camera in the fore-finger and thumb with other fingers clamped on the strap. The Auto-Rotate is such crap - who needs to see a smaller image rather than turn the camera? Mode switch in this position changes easilly when pulling the camera out of it's case.

Figure 4.4: Review from Digital Photography Review (www.dpreview.com)used in the example

Classify each Sentence into One of "Good"," Bad" and" Quality" Category

The set of rules obtained in the section 3 is applied to each sentence of the new review to classify it into one category. For example the first sentence has been classified into the Good category. The second sentence has been classified into the Quality category as it is illustrated next:

Sentence 1: You cant take a photo unless you are carrying a camera and this is a good one to carry all the time!
Goods rules: rule 26
Bad rules: none
Quality rules: none
Classification:GOOD

Sentence 2: I have owned most IXUS cameras since the first - Just traded my Ixus 700 on this one (an Ixus 65 but still have an Ixus 40)
Goods rules: none
Bad rules: none
Quality rules: rule 5
Classification:QUALITY

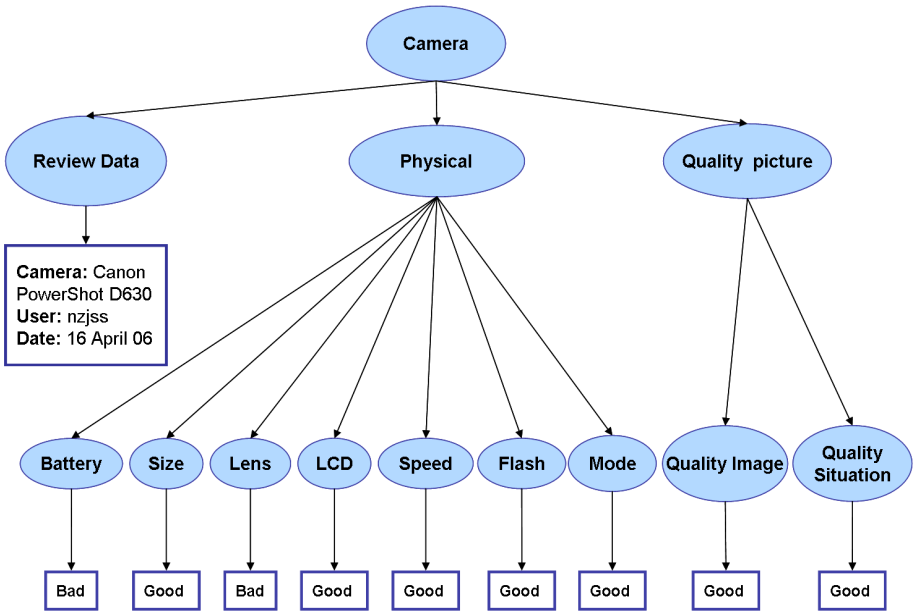


Figure 4.5: Mapped ontology from consumer review comment

Finding the Concept Represented in the Sentence

For each sentence that is in the "Good" or "Bad" categories, the mapping feature of this sentence is found by searching the keywords that are in the related word list. For example in the first sentence, that has been classified as a good opinion. Also, a word has been found, which is related to the concept "size". We suppose that the size of the camera is good for a customer so is assigned the value good for the feature size in the ontology. Follow are show some cases of the concept identification made in the case study:

Sentence 1: You cant take a photo unless you are carrying a camera and this is a good one to carry all the time!
Classification:GOOD
Concepts: size (related word: carry)

Sentence 14: This is my "main" camera - and while there are times when I want a longer or wider lens I would not be draging along the big camera that they are attached to.
Classification:BAD
Concepts: lens (related word: lens)

Figure 4.5 shows the mapping of the new review into the predefined ontology.

Consumer Information					
Consumer	John	Karem	James	Laura	Andy
Camera	Sony361	sony361	Olympus810	OlympusZ25	CanonTW45
Opinion Quality					
Consumer Skill	Beginner	Professional	Beginner	Beginner	Professional
Time to Use this Camera	2 months	1 year	2 weeks	3 months	2 months
Time to Use Digital Camera	4 months	1 year	3 weeks	5 months	2 years
Number of Different Cameras	2	1	1	2	3
Product Quality					
Features	**Size:** good **Interface:** bad **Doc.** good **Zoom:** good	**Zoom:** good **Interface:** bad **Price:** bad **battery:** good	**weight:** good **Doc.:** bad **material:** bad **sof.:** bad **wifi:** bad **Start Up**: bad	**Size:** good **Interface:** bad	**Size:** good **Doc.:** good
Consumer Information					
Consumer	Bernat	Amadeus	Micky	Carmen	Bern
Camera	OlympusZ25	CanonTW45	Olympus810	Olympus700	Olympus700
Opinion Quality					
Consumer Skill	Beginner	Beginner	Professional		Beginner
Time to Use this Camera	2 moths	1 months	2 weeks	3 moths	2 moths
Time to Use Digital Camera	4 moths	moths	3 weeks		8 moths
Number of Different Cameras	2	1	1	2	3
Product Quality					
Features	**Size:** good **Interface:** good **Doc:** good **Zoom:** good **Start up**: good **Software**: good	**Zoom:** good **Interface:** good **Price:** bad **battery:** good	**weight:** good **Doc.** good **material:** good **sof.:** good **wifi:** bad **StartUp**: good	**Size:** good **Interface:** good **Material:** good **Software:** good	**Size:** good **Doc.:** good

Figure 4.6: Ontology instances mapped from consumer reviews

4.5.3 Computing the recommendation

In this section, detail calculations of the recommendation in response to user request are given. Figure 4.6 shows 10 ontology instances mapped from review comments, which are used for the recommendation calculations in the example.

Obtaining the Opinion Quality (OQ)

The Opinion Quality is calculated by equation 4.1 in Section 4.2. Table 4.5 presents the weighting value of each variable defined in the equation.

The OQ values for each consumer in Table 4.5 are:

$OQ_{john} = \frac{0.5+0.7+0.7+0.5}{4} = 0.6$

$OQ_{karen} = \frac{0.9+0.9+0.9+0.3}{4} = 0.75$

$OQ_{James} = \frac{0.5+0.5+0.3}{3} = 0.43$

$OQ_{Laura} = \frac{0.5+0.7+0.7+0.5}{4} = 0.6$

	Value	**Weight(w_i)**
Consumer skill	Beginner	0.5
	Advanced	0.7
	Professional	0.9
Consumer experience		
Time to use this camera	Day	0.3
	Week	0.5
	Month	0.7
	Year	0.9
Time to use digital Camera	Day	0.3
	Week	0.5
	Month	0.7
	Year	0.9
Number of different cameras	One	0.3
	Two	0.5
	Three	0.7
	(+)Three	0.9

Table 4.5: Variables representing the consumer level expertise in using a digital camera

$$OQ_{Andy} = \frac{0.9+0.7+0.9+0.7}{4} = 0.8$$
$$OQ_{Bernat} = \frac{0.5+0.7+0.7+0.5}{4} = 0.6$$
$$OQ_{Amadeus} = \frac{0.5+0.7+0.7+0.3}{4} = 0.55$$
$$OQ_{Micky} = \frac{0.9+0.5+0.5+0.3}{4} = 0.55$$
$$OQ_{Carmen} = \frac{0.7+0.5}{2} = 0.6$$
$$OQ_{Benn} = \frac{0.5+0.7+0.7+0.7}{4} = 0.65$$

With the calculated values, the best opinion came from Andy: note that Andy is a professional photographer; he has used digital cameras for longer period of time than the other consumers in the sample and he has used three different cameras. This information leads to the assumption that Andy's opinion is the most valuable opinion within the sample.

As it is shown in Equation 4.1, the Opinion Quality is calculated as the average of the four variables. In the case of missing information on one or more variables, the Opinion Quality is calculated as the average of the remaining variables. In the worst case that

Consumer	John
Camera	SonyW70
OQ_{John}	0.6
Size	good
Interface	bad
Documentation	good
Zoom	good

Table 4.6: Information about John's opinion

Consumer	John
Camera	SonyW70
OQ_{John}	0.6
FQ_{Size}	0.6
$FQ_{Interface}$	-0.6
$FQ_{Documentation}$	0.6
FQ_{Zoom}	0.6

Table 4.7: Feature Quality (FQ) for each feature rated by John

none of the information is available, the Opinion Quality is set to be the average of the lowest possible values of all variables, which is $0.35(= (0.3 + 0.3 + 0.3 + 0.4)/4)$ according to table 4.5.

Obtaining the Feature Quality (FQ)

Feature Quality (FQ) value for each feature rated by the consumers is also calculated. For example as shown in Table 4.6, John gave the value "good" or "bad" for each feature of the digital camera SonnyW70 and his OQ value is 0.6.

As described in previous sections, by assigning the value 1 for "good" and -1 for "bad" in equation 2, the Feature Quality for each feature in John's opinion are calculated, as shown in Table 4.7.

The same process has been applied to all consumers. The OQ and FQ for each review comment are calculated offline to achieve quick response to the user requests. The recommender system requires from the user to input the model of the camera he (she) is interested and selects the features that he (she) is most concern. The features in the selection panel are the same set of features that is covered by the ontology.

4.5.4 Making a recommendation

A user request "***I would like to know if SonyW70 is a good camera, specifically its interface and battery consumption***" is presented. Three keywords (SonyW70, interface and battery) can be identified. Firstly, only the opinions for SonyW70 are selected. Observing in Figure 4.6, there are three opinions about SonyW70's cameras: John's opinion, Karen's opinion and James's opinion. Then the OFQ of each feature is calculated using equation 4.3.

$OFQ_{interface} = \frac{1}{2}(\frac{1}{4} * (-0.6) + \frac{1}{4} * (-0.75)) = -0.165$

$OFQ_{battery} = \frac{1}{4} * 0.75 = 0.18$

The Overall Assessment for the digital camera SonyW70 based on the two features requested is obtained using equation 4.5. The Importance Index was calculated in two ways. For the case of using the importance index from the user request where the user has expressed that the interface is more important than the battery, so the value of 1 is assigned for interface and 0.5 for battery. Using these values the OA for SonyW70 camera is:

$OA = -0.165 * 1 + 0.18 * 0.5 = -0.075$

In the case of no user preference is given, the importance index are calculated based on the frequency of the feature being reviewed:

$$ImportanceIndex = \frac{n}{N} \tag{4.6}$$

Where n is the number of times that the feature appears in the reviews and N is the total number of reviews. Using equation 4.6, the OA for SonyW70 camera is:

$ImportanceIndex_{Interface} = \frac{6}{10} = 0.6$

$ImportanceIndex_{Battery} = \frac{2}{10} = 0.2$

$OA = -0.165 * 0.6 + 0.18 * 0.2 = -0.063$

Assigned the value "Good" for OA and OFQ > 0 and "Bad" for OA and OFQ < 0 the SonyW70 camera is "Bad" according to consumers' opinions. The response for the user request is shown in Figure 4.7. The best camera with the features the user concern is also recommended. The same process is applied to all other cameras review. CanonA630 is recommended considering this information. The complete recommendation is show in Figure 4.8).

The analysis of the experimental results was carried out with 195 "Good" comments,

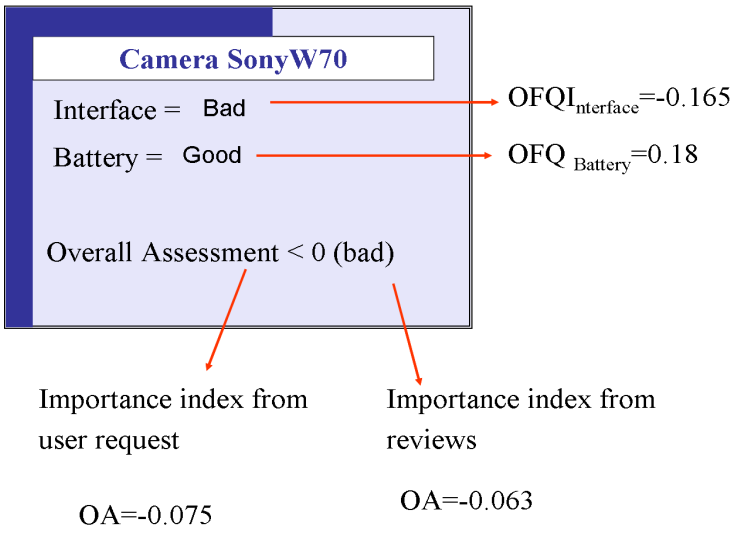


Figure 4.7: Final recommendation answer to the user request generated from consumer's opinios about digital cameras

127 "Bad" comments and 47 "Quality" comments from 68 user reviews of digital cameras. The following conclusions about the mapping process can be drawn.

- The comments used in the example are all for one model of camera (Canon PowerShot SD500). The recall and precision measures can be further improved in the classification process by using multiple models.
- "Good" category contains more training data than other categories so it achieved the best results amongst.
- There are some long and complicated sentences that cannot be classified into any category. These sentences should be broken into several short sentences before the classification.

Despite of these issues, the results obtained are considered good as we can accurately map a large portion of a review into the predefined ontology.

4.6 Experiments using a real case study

In this section, we describe the results of a real case with 33 people from various different countries such as Spain, Colombia, Mexico, Argentina, Russia, Germany, Belgium and

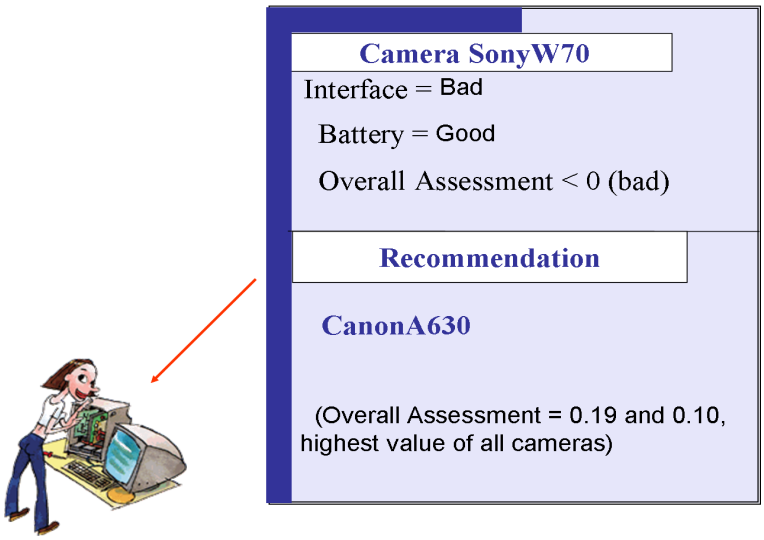


Figure 4.8: Recommendation in response for a user request from consumers' opinions

Rumania. There were 13 women and 20 men with ages ranging from 23 to 44 whose are working or studying in the University of Girona. The objective of this experiment was to test the potential of recommendations made using information based on other web users' opinions about digital cameras.

There were three parts to the study: (i) obtaining information from opinions, (ii) calculating the recommendation in response to the user's requests and (iii) evaluating the recommendation. To obtain the data, user opinions on digital cameras were collected from the website www.dpreview.com, and then text mining techniques were applied, as described in previous sections, resulting in ontologies representing the opinion of each one of the users about a digital camera.

Using this data, and having analysed the requests made by the 33 users, the recommendation was calculated using measures of Opinion Quality, Feature Quality, Overall Feature Quality and Overall Assessment. The third part consisted of the user filling in a questionnaire designed to reveal how satisfied they were with recommendation they were given.

We will discuss each one of these parts in more detail in the following sections. First we will clarify two things. We will call the users of the system to which a digital camera was recommended, the "user-system", while the users who provided opinions and from whom we obtained digital camera evaluations will be called "user-reviewers".

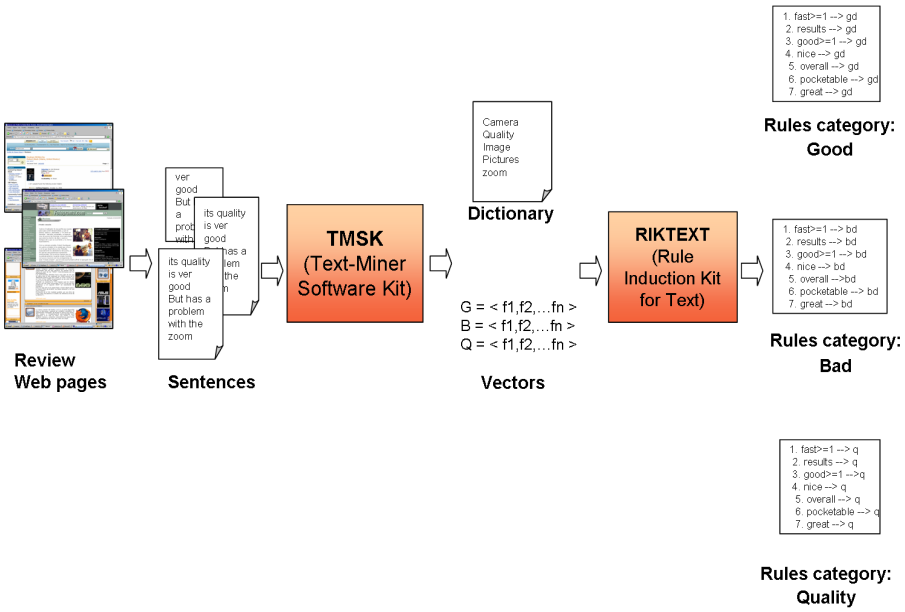


Figure 4.9: Inputs and outputs for classifier consumer's reviews process

4.6.1 Obtaining the opinion of the users

The goal of this study is to investigate the potential of our system in making recommendations when nothing is known about the "user-system" (very cold start), i.e., when the user is a new user of the system and this is the first time that a digital camera has been recommended to him. The camera recommended in this situation is the one which is most highly evaluated by other users and which most closely responds to the request made.

To get user opinions on digital cameras, text mining was carried out using opinions on 200 cameras of 19 different brands: Canon, Casio, Epson, Fujifilm, HP, Kodak, Konica, Kyocera, Leica, Minolta, Nikon, Olympus, Panasonic, Pentax, Ricoh, Samsung, Sanyo, Sony and Toshiba.

Text mining was carried out for each one of the cameras, obtaining a set of rules for each one of the categories: Good, Bad and Quality for each camera. Figure 4.9 shows the software input used to obtain this set of rules: TMSK and RIKTEXT. For example, the following rules were obtained for classifying the opinions on the Kodak EasyShare V610 camera.

This set of rules was obtained for each one of the cameras. Once the rules were obtained, each one of the opinions was read and a search made in every sentence to see if some concept of the ontology exists, using the same method carried out in the

Selected rule set
1. good >= 1 − − > gd
2. nice −− > gd
3. great >= 1 − − > gd
4. pocket −− > gd
5. easy AND use −− > gd
6. pleased −− > gd
7. best −− > gd
8. zoom >= 1 − − > gd
9. picture >= 1 − − > gd
10. shots >= 1 − − > gd
11. [TRUE] −− > gd

Table 4.8: Rule set to classify the sentences of the Kodak EeasyShare V610 camera in the Good category

Selected rule set
1. problem >= 1 − − > bd
2. doesn't −− > bd
3. haven't >= 1 − − > bd
4. noise −− > bd
5. no −− > bd
6. focusing −− > bd
7. battery −− > bd
8. [TRUE] −− > bd

Table 4.9: Rule set to classify the sentences of the Kodak EeasyShare V610 camera in the Bad category

Selected rule set
1. first >= 1 − − > ql
2. week −− > ql
3. best >= 1 − − > ql
4. [TRUE] −− > ql

Table 4.10: Rule set to classify the sentences of the Kodak EeasyShare V610 camera in the Quality category

```
<?xml version="1.0" ?>

<Camera>
      <userOpinion>
            <nameUser> albanna </nameUser>
            <dateOpinion> 29 Sep 06 </dateOpinion>
            <nameCamera> Kodak V610 </nameCamera>
      </userOpinion>
      <Opinions>
            <OpinionQuality>
                  <UserSkill> </UserSkill>
                  <UserExperience>
                        <TimeToUseThisCamera> </TimeToUseThisCamera>
                        <TimeToUseDigitalCamera> </TimeToUseDigitalCamera>
                        <NumberOfDifferentCamera> </NumberOfDifferentCamera>
                  <UserExperience>
            </OpinionQuality>
            <Features>
                  <Quality>
                  <Quality_Situation> bad </Quality_Situatio>
                  <RedEye> </RedEye>
                  <Quality_Image> good </Quality_Image>
                  <Purple_Fringing> </Purple_Fringing>
                  <Quality_Video> </Quality_Video>
                  </Quality>
                  <Phisical>
                  <Price> good </Price>
                  <Size> good </Size>
                  <Zoom> good </Zomm>
                  <LCD> good </LCD>
                  <Lens> </Lens>
                  <Battery> bad </Battery>
```

Figure 4.10: Ontology representing the review of the user albanna about the camera Kodak V610

case study in section 4.5.2. In this way, an XML ontology was obtained for each user-reviewer representing his opinion and experience for a particular digital camera. Figure 4.5.2 shows the ontology obtained when applying text mining to the opinion of the user named "alabanna" with regard to the Kodak V610 digital camera. In this ontology, you can see the ratings given for the camera features such as, for example, that the image quality is good, but not the battery. It can also be seen, that there is no information on this person's skills or experience with the camera.

In this way, ontologies were obtained representing opinions on a digital camera from each user-reviewer. Using the data contained in the ontologies, Opinion Quality ratings are calculated. In the specific case of the user "alabanna", where there is no data on his/her ability or experience with the digital camera, the Opinion Quality rating (OQ) will be computed as the average of the lowest ratings granted to the variables used for its calculation (Table 4.5) which is $OQ = (0.5 + 0.3 + 0.3 + 0.3)/4 = 0.35$. The Feature Quality rating (FQ) is also calculated for each camera feature and for each opinion by the "user-reviewers". So far we have - for each one of the cameras - OQ and FQ ratings from each one of the "user-reviewers".

The ratings for Overall Feature Quality (OFQ) and Overall Assessment (OA) are calculated at the time requests are received from the "user-system" so that recommendations can be made to them. In the next section, this process is explained in detail.

4.6.2 Obtaining the recommendations

The user-systems were asked which digital camera they wanted and were able to add information about what they considered were most important features of the camera or ask other users for more information about these features. To stimulate participation and to improve the accuracy and reliability of the information supplied to the system, there was a prize draw of two digital cameras worth a maximum of USD 520 $. This maximum becomes a restriction in our system which is apparent in the following situation: The user-systems made requests for cameras that cost less than 520 $ and the system should only recommend those that cost less than 520 $. In any case, 65 % of the cameras come into this category.

The user-systems made requests in various ways, some introducing the brand name of the camera they wanted, for example, the user named U1 wrote: "Nikon, 2.5" screen", zoom lens", user U3 asked for a "Kodak Easy Share V610", and user U24 wrote, "I want a small camera (that fits into my pocket), is lightweight, has at least 8 mp and optical zoom lens of 3x" while user U26 simply wrote "an automatic camera". All the 33 requests received are listed in Table 4.11:

In response to these requests the recommendations were calculated. For each one of the requested cameras and depending on the characteristics included in the requests, OFQ and OA ratings were obtained. The resulting recommendations and the OA ratings of the cameras requested are listed in Table 4.12:

As you can see, for each one of the users, the system found the highest-rated camera, i.e., the one with the highest OA (Overall Assessment) rating. Out of all the cameras, the three with the highest OA ratings, according to all features, are: Kodak P712 (OA=0.256), Panasonic Lumix DMC-FZ7 (OA=0.220) and Canon PowerShot SD800 IS (OA=0.206). An OA with a positive rating means that most of the user-reviewers for that camera were happy with many of the camera's features. In contrast, most of the user-reviewers thought the features of the camera Sony Cyber-shot DSC-T50, were not good, leaving it with a negative OA (OA=-0.285). Hence, when user U4 requested this camera, he was recommended the most highly rated Sony camera with features similar to those he requested.

"User-System"	***request***
U1:	*camera Nikon, screen 2.5, zoom*
U2:	*camera NIKON Coolpix S10*
U3:	*camera Kodak EASYSHARE V610*
U4:	*camera Sony Cyber-shot DSC-T50*
U5:	*camera NIKON Coolpix S10, battery and screen*
U6:	*camera Kodak EasyShare P712*
U7:	*camera Sony Cyber-shot DSC-h5*
U8:	*camera Canon IXUS 900 TI*
U9:	*camera Canon PowerShot S3 IS*
U10:	*digital Camera: Sensor: CMOS, 8,0 Megapixel Size: 14.8 x 2 2.2mm SLR screen: 1.8 " TFT Color Shutter speed: 1/4000-30 sec Measures: 12.7 x 6.35 x 9.39 cm. Connections: USB y Video Bateria:Ion-Litio*
U11:	*camera Olympus SP-500 UZ*
U12:	*camera with good optic, speed and screen of 3" or higher*
U13:	*automatic camera that make all*
U14:	*camera Canon Digital Ixus 800 IS*
U15:	*camera Sony Cyber-shot DSC-h5*
U16:	*camera SONY Cyber-shot DSC-H2*
U17:	*camera digital*
U18:	*camera Sony Cyber-shot DSC-h5*
U19:	*camera Canon Digital Ixus 850 IS*
U20:	*camera Sony DSC-W100*
U21:	*camera digital*
U22:	*camera Sony Cyber-shot DSC-h5*
U23:	*camera Panasonic Lumix DMC-FZ7*
U24:	*small camera,that it be light, 8 mp or higher and optical zoom de 3x or higher*
U25:	*camera Sony DSC-T10*
U26:	*camera digital*
U27:	*camera Canon IXUS 55*
U28:	*camera CANON EOS 400D*
U29:	*camera Canon Digital Ixus 850 IS*
U30:	*camera Canon PowerShot SD600*
U31:	*camera Kodak EasyShare Z710*
U32:	*camera Canon PowerShot A640*
U33:	*digital camera ultra compact best rated*

Table 4.11: "User-System" request

User: **U1** your request was: **camera Nikon, screen 2.5, zoom**
The Nikon camera most highly rated by the users is ***Nikon Coolpix P4*** *We therefore recommend that you buy this camera.*
Overall Assessment of Nikon Coolpix P4 = **0.173**

User: **U2** your request was: **camera NIKON Coolpix S10**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of NIKON Coolpix S10 = **0.011**

User: **U3** your request was: **camera Kodak EASYSHARE V610**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of EASYSHARE V610 = **0.156**

User: **U4** your request was: **camera Sony Cyber-shot DSC-T50**
Most of the users are not happy with the features of this camera. We therefore recommend the Sony camera best rated and with the feature most similar that you have requested, this camera is: ***Sony Cyber-shot DSC-T10***
Overall Assessment of Sony Cyber-shot DSC-T50 = **-0.285**
Overall Assessment of Sony Cyber-shot DSC-T10 = **0.149**

User: **U5** your request was: **camera NIKON Coolpix S10, battery and screen**
Most of the users are satisfied with the screen of this camera. No opinions have been given on the battery. In general, most users are happy with many of the features of this camera. We therefore recommend that you buy it.
Overall Assessment of NIKON Coolpix S10 = **0.011**

User: **U6** your request was: **camera Kodak EasyShare P712**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it. It is one of the three most highly rated cameras.
Overall Assessment of Kodak EasyShare P712 = **0.256**

User: **U7** your request was: **camera Sony Cyber-shot DSC-h5**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Sony Cyber-shot DSC-h5 = **0.112**

User: **U8** your request was: **camera Canon IXUS 900 TI**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Canon IXUS 900 TI = **0.111**

User: **U9** your request was: **Canon PowerShot S3 IS**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Canon PowerShot S3 IS = **0.206**

User: **U10** your request was: **Digital Camera: Sensor: CMOS, 8.0 Megapixel; Size: 14.8x 22.2mm SLR Screen: 1.8 Inches TFT Color Shutter speed: 1/4000 - 30 sec Measurements:12.7 x 6.35 x 9.39 cm. Connections: USB i Video Battery: Ion-Litio**
No camera priced less than 400 Euro has been found with these features. We recommend the camera most highly rated by the users and which most closely meets your requirements. This camera is: ***Sony DSC-W100***
Overall Assessment of Sony DSC-W100 = **0.128**

User: **U11** your request was: **Olympus SP-500 UZ**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Olympus SP-500 UZ = **0.134**

User: **U12** your request was: **camera with good optical, faster and screen of 3" or more**
The camera best rated for users according these features is ***Sony DSC-H5****. We therefore recommend that you buy it.*
Overall Assessment of Sony DSC-H5 = **0.112**

User: **U13** your request was: **Automatic camera that make all**
We recommend to you three of the best rated cameras: ***Kodak P712, Panasonic Lumix DMC-FZ7 and Canon PowerShot SD800 IS***
Overall Assessment of Kodak P712 = **0.256**
Overall Assessment of Panasonic Lumix DMC-FZ7 = **0.220**
Overall Assessment of Canon PowerShot SD800 IS = **0.206**

User: **U14** your request was: **Canon Digital Ixus 800 IS**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Canon Digital Ixus 800 IS = **0.206**

User: **U15** your request was: **camera Sony Cyber-shot DSC-h5**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Sony Cyber-shot DSC-h5 = **0.112**

User: **U16** your request was: **camera SONY Cyber-shot DSC-H2**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of SONY Cyber-shot DSC-H2 = **0.113**

User: **U17** your request was: **Camera digital**
We recommend to you three of the best rated cameras: ***Kodak P712, Panasonic Lumix DMC-FZ7 and Canon PowerShot SD800 IS***
Overall Assessment of Kodak P712 = **0.256**
Overall Assessment of Panasonic Lumix DMC-FZ7 = **0.220**
Overall Assessment of Canon PowerShot SD800 IS = **0.206**

User: **U18** your request was: **camera Sony Cyber-shot DSC-h5**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Sony Cyber-shot DSC-h5 = **0.112**

User: **U19** your request was: **camera Canon Digital Ixus 850 IS**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it. It is one of the three most highly rated cameras.
Overall Assessment of Canon Digital Ixus 850 IS = **0.206**

User: **U20** your request was: **camera Sony DSC-W100**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Sony DSC-W100 = **0.128**

User: **U21** your request was: **Camera digital**
We recommend to you three of the best rated cameras: ***Kodak P712, Panasonic Lumix DMC-FZ7 and Canon PowerShot SD800 IS***
Overall Assessment of Kodak P712 = **0.256**
Overall Assessment of Panasonic Lumix DMC-FZ7 = **0.220**
Overall Assessment of Canon PowerShot SD800 IS = **0.206**

User: **U22** your request was: **camera Sony Cyber-shot DSC-h5**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Sony Cyber-shot DSC-h5 = **0.112**

User: **U23** your request was: **camera Panasonic Lumix DMC-FZ7**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it. It is one of the three most highly rated cameras.
Overall Assessment of Panasonic Lumix DMC-FZ7 = **0.220**

User: **U24** your request was: **small camera, that it be light, 8 mp or higher and optical zoom de 3x or higher**
The camera best rated for users according these features is ***Sony DSC-N2*** *We therefore recommend that you buy it.*
Overall Assessment of Sony DSC-N2 = **0.152**

User: **U25** your request was: **camera Sony DSC-T10**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it.
Overall Assessment of Sony DSC-T10 = **0.149**

User: **U26** your request was: **Camera digital**
We recommend to you three of the best rated cameras: ***Kodak P712, Panasonic Lumix DMC-FZ7 and Canon PowerShot SD800 IS***
Overall Assessment of Kodak P712 = **0.256**
Overall Assessment of Panasonic Lumix DMC-FZ7 = **0.220**
Overall Assessment of Canon PowerShot SD800 IS = **0.206**

User: **U27** your request was: **camera Canon IXUS 55**
Most of the users are happy with the features of this camera. We therefore recommend that you buy it. Overall Assessment of Canon IXUS = **0.053**

User: **U28** your request was: **camera CANON EOS 400D** *No camera priced less than 400 Euro has been found with these features. We recommend the camera most highly rated by the users and which most closely meets your requirements. This camera is:* ***Canon PowerShot A640*** Overall Assessment of Canon PowerShot A640 = **A640**
User: **U29** your request was: **camera Canon Digital Ixus 850 IS** *Most of the users are happy with the features of this camera. We therefore recommend that you buy it. It is one of the three most highly rated cameras.* Overall Assessment of Canon Digital Ixus 850 IS = **0.206**
User: **U30** your request was: **camera Canon PowerShot SD600** *Most of the users are happy with the features of this camera. We therefore recommend that you buy it.* Overall Assessment of Canon PowerShot SD600 = **0.122**
User: **U31** your request was: **camera Kodak EasyShare Z710** *No opinion about this camera has been found. We recommend the camera Kodak most highly rated by the users and which most closely meets your requirements. This camera is:****Kodak EasyShare P712*** Overall Assessment of Kodak EasyShare P712 = **0.256**
User: **U32** your request was: **camera Canon PowerShot A640** *Most of the users are happy with the features of this camera. We therefore recommend that you buy it.* Overall Assessment of Canon PowerShot A640 = **0.157**
User: **U33** your request was: **Digital camera ultra compact best rated** *We recommend to you three of the best rated cameras:* ***Kodak PowerShot SD800 IS, Kodak V610 and Sony DSC-N2****.* Overall Assessment of Canon PowerShot SD800 IS = **0.206** Overall Assessment of Kodak V610 = **0.156** Overall Assessment of Sony DSC-N2 = **0.152**

Table 4.12: Recommendation results

The users who simply requested a digital camera, without specifying features or brand names, were recommended the aforementioned three most highly-rated cameras, all features considered. Another case worth commenting on is that of user U10 who requested a digital camera with very specific features such as "Sensor: CMOS, 8.0 Megapixel; Size: 14.8x 22.2mm; SLR Screen: 1.8 Inches; TFT Colour Shutter speed: 1/4000 -30 sec; Measurements 12.7 x 6.35 x 9.39 cm; Connections: USB and Video Battery: Ion-Lithium". No camera with these features is available at a price of less than $ 520 (price is one of the restrictions of the system), so the camera recommended to him was that which had most features in common with his/her request and had the highest OA rating.

4.6.3 Evaluating the set of recommendations

In order to evaluate the recommendations, the system uses precision and recall measures that are used to evaluate the performance of a recommender system [Montaner et al., 2002]. The precision represents the probability that a recommendation is successful, whereas the recall represents the probability that an item has of being recommended. To get the data to calculate these values, the user-system, having received their recommendation, are then asked to fill in a questionnaire with the following questions:

1. *Is this camera that you wanted?*
2. *Is the camera that you requested?*
3. *Would you buy this camera based on the opinions of other users?*
4. *Would you like to have a system that searches for other web users' opinions and shows you them like this?*

Answers to the first two questions were used to calculate the precision and recall of the system, whereas the last two were designed to evaluate whether the system itself was worth having.

This questionnaire was apply to 33 user-system and the following answers were obtained (see Table 4.13):

With this information, the precision was calculated using the formula defined by Salton and Buckley, [Salton and Buckley, 1988] and which was described in Chapter 3:

$$Precision = \frac{Pr}{P} \tag{4.7}$$

User	Question 1	Question 2	Question 3	Question 4
U1	no	yes	no	no
U2	no	yes	no	no
U3	yes	yes	yes	yes
U4	no	no	yes	yes
U5	yes	yes	yes	yes
U6	yes	yes	yes	yes
U7	yes	yes	no	yes
U8	yes	yes	yes	yes
U9	yes	yes	yes	yes
U10	yes	no	yes	yes
U11	yes	yes	yes	yes
U12	no	yes	yes	yes
U13	yes	yes	no	yes
U14	yes	yes	yes	yes
U15	no	yes	yes	yes
U16	yes	yes	yes	yes
U17	no	no	yes	yes
U18	yes	yes	yes	yes
U19	yes	yes	yes	yes
U20	yes	yes	yes	yes
U21	yes	yes	yes	yes
U22	yes	yes	yes	yes
U23	yes	yes	yes	yes
U24	yes	yes	yes	yes
U25	no	yes	no	yes
U27	yes	yes	yes	yes
U28	no	no	yes	yes
U29	yes	yes	yes	yes
U30	yes	yes	yes	yes
U31	yes	no	no	yes
U32	yes	yes	yes	yes
U33	yes	yes	yes	yes

Table 4.13: Responses given by the users to the four questions

Where Pr is the number of successful recommendations and P is the number of recommendations carried out. In this experiment, a successful recommendation is one for which the system users have confirmed that the recommended camera was the camera they had requested. Bearing this in mind, Pr is calculated as the number of users who said yes to the second question of the questionnaire (column 3 of Table 4.13). P is the number of recommendations that were made; in this experiment there were 33 recommendations. Using this data, the precision of the recommender system is:

$$Precision = \frac{25}{33} = 0.76 \tag{4.8}$$

Recall is computed follows:

$$Recall = \frac{Pr}{PD} \tag{4.9}$$

Where Pr is the number of successful recommendations and PD is the number of possible recommendations. In this experiment, a possible recommendation is one which the users confirmed that the recommended camera was the camera they wanted. Hence, PD is calculated as the number of users who said yes to the first question of the questionnaire (column 2 of Table 4.13). Thus, the recall of the recommender system is:

$$Recall = \frac{25}{28} = 0.89 \tag{4.10}$$

Since the ratings for precision and recall are high, the recommendations were successful. For those recommendations that were not successful, we tried to identify the reasons for the failure. Using the information given in answer to the questions, we have identified two reasons:

1. The system recommended a camera that they had requested, but which they did not actually want. These are the users who said no to the first question and yes to the second. There were 5 such cases and in each case, the system recommended the camera that they had requested or that matched the features specified in requests. Two of the reasons why this camera was not the one they actually wanted were a) because the user had failed to specify some feature he/she had wanted or b) the camera requested cost more than $ 520.

2. The system recommended a camera that the user neither wanted nor had requested. These were the users who responded no to both the first and second question. There were 3 such cases. Analysis showed that the system recommended the camera that

was most similar to the one the user requested - rather than the one actually requested - because the OA rating of the requested camera was negative or it cost more than $ 520. Another case involved the user not specifying clearly what he/she wanted and since there was no data about what he/she wanted, the system recommended the three highest scoring cameras which were rejected. This was a one-off case, since most of the times that a user expressed no preference and was therefore recommended the three top cameras, the users were happy with the recommendation. Nevertheless, these are aspects to consider in future work on improving the performance of this recommender system.

Another aspect evaluated was the importance of a system such as this that retrieves internet-based user preference data and presents such data in a structured way which is easily accessible, automatic, fast and simple. To do so, we used the answers given by the system users to the third and fourth questions of the questionnaire. 82% of the users said that they would base their purchase of a product (in this case, a digital camera) on the opinions of other users. 94% of the users said they would like to have a system that searched for opinions about products on the Internet. In this sense, the system would save time searching for opinions that will help the user to decide whether or not to buy a product.

4.7 Conclusions

This chapter proposed a novel approach to structure online information and create recommendations in recommender systems, which utilizes online consumer review comments. The research work using reviews in a recommender system is still in its infancy. To the best of our knowledge, this is the first attempt to build a recommender system based on review comments in free form text. In [Ricci and Wietsma, 2006][Wietsma and Ricci, 2005] the authors use the reviews to give some explanation about the recommendation of a product. We have proposed a potentially and novel approach for the retrieval of review's information.

This work makes 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.

In the case study, an ontology has been defined for the domain of digital camera reviews and has been used for demonstration of the work with some examples. The example shows that the information extracted from unstructured data contained in the reviews can be mapped into predefined ontology using text mining techniques based on decision rules. Experiments have shown that good results are obtained when classifying simple and curt sentences from the review, long and complicated sentences can not been classify into one category to them map it into the ontology, so the further work have to address this problem. However a real case study using review about many digital cameras have been implement. The recommendation have been evaluated resulting in good result also. The implementation of this method allows a recommender system to use valuable textual information for recommendation.

In conclusion, in this chapter, we have provided a new application of reviews to recommender systems. The experimental study shows that it is a promising approach. These reviews allow us obtain ratings of products that have not been rated by a sufficient number of consumers unstructured sources to overcome the cold start problem. Once the information have been structured is apply the methodology ACQUAINT to obtain its relevance and trust measure to select the most relevant and trustworthy. Due to time constraint this step will be in further work.

Chapter 5

Conclusions and Future Work

In this chapter, we present the conclusions of our research. First, we discuss the scientific contributions made in the course of this doctoral thesis. Then, we will comment the future studies to be made and the related lines of research for future application. Finally, we present a list of related publications as well as the scientific collaborations involved in the preparation of the thesis.

5.1 Summary

This thesis 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 recommendations more effective and applicable to real life [Adomavicius, 2005]. One of the problems that recommender systems try to solve, whatever the method used to make recommendations, is evaluating products that have never been bought by a user. This evaluation is made on the basis of his/her preferences. In the case of CBF, preferences of a user are obtained from evaluations that he/she gave for other products. In CF, his/her preferences are taken as being similar to those of similar users. User preference data can be implicitly stored in different sources of information. In this work, a methodology has been created to select sources with the most suitable data for retrieving user preferences so that more effective recommendations can be made. We have also created a method for retrieving and structuring user preferences available on Internet.

The success of a recommendation method depends on the amount of data available to make the evaluations. The lack of information originate the so-called "cold start"

problems when there is no user information to make the first recommendation and the problem of "Sparsity" when there is insufficient information about user preference to make recommendations [Stuart et al., 2002].

The search for and selection of **relevant and trustworthy** sources that allow us to get more user information is one of the subjects analyzed in this thesis. The suitability of the information sources used in recommender systems has not been measured until now. In Chapter 3, we presented the ACQUAINT methodology which uses two criteria to select the most suitable sources: the relevance (R) and trust(T) of the sources. The relevance is obtained base on characteristics of the sources. These characteristics must be representative of the information required to know whether or not the sources contain user information. The trust of a source is a value that represents the degree of success of the recommendations made using this source in the past. Applying this methodology in two case studies showed that the set of characteristics defined is representative of the information contained in sources. The results obtained have shown that, recommendations made with information from several sources - selected based on these two criteria (R and T) - are more precise and close to the optimal result. The methodology is general and easy to apply.

In addition to this methodology, we have also proposed a mechanism that allows us to retrieve and structure the user information from web pages. Once this information is structured, the ACQUAINT methodology is applied. In Chapter 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 the user's preference. The information was structured using ontologies.

5.2 Contributions

There are two main contributions that stem from this work. The first is a method to identify and select relevant and trustworthy sources that can be used in a recommender system, thus providing a possible solution to the problem of sparsity. The second is a method that retrieves and structures user's preferences.

We can subdivide these 2 main contributions into more specific contributions:

1. A set of intrinsic characteristics has been created which indicate whether or not a source contains information that the recommenders need [Aciar et al., 2005c] [Aciar et al., 2007a].

2. A measure of the relevance of sources based on characteristics defined in point 1, has been defined [Aciar et al., 2005b] [Aciar et al., 2005d].

3. A trust measure that allow us to know how trustworthy a source may be has been applied . This measure is calculated according to the success of previous recommendations using a source [Aciar et al., 2005a].

4. An algorithm that the system uses to select from a set of candidate sources the most relevant and trustworthy base on measures defined in previous points, has been created [Aciar et al., 2006a].

5. A ontology to structured the user's information from web page has been defined [Aciar et al., 2006c] [Aciar et al., 2007b].

6. A process to map the user information into the ontology using text mining techniques has been created [Aciar et al., 2006b] [Aciar et al., 2007b].

5.3 Future work

There are two main directions for future work in this research. The first would seek to improve or emphasize the functionalities of the proposed methodology. This includes implementing this methodology in a multiagent system utilising advantages of agent paradigm in order to make the search for data sources fully automatic. The second, which is a little more specific, would deal with implementing the method used to retrieved unstructured user information. In this category, we would include the redefinition of the ontology to retrieve information from different domains. In addition, there are aspects of this work that could be applied to various lines of research currently worked within research group where this thesis was developed. These possible connections are mentioned in this section.

We will now describe the work that should be done to improve the contributions made in this thesis:

5.3.1 Applying the methodology using other information sources

Suppose we had a scenario with many sources available - structured and unstructured. Unstructured sources are becoming more and more available now that the Web is evolving into the so-called Web 2.0, due to recent innovations including blogs, wikis and social

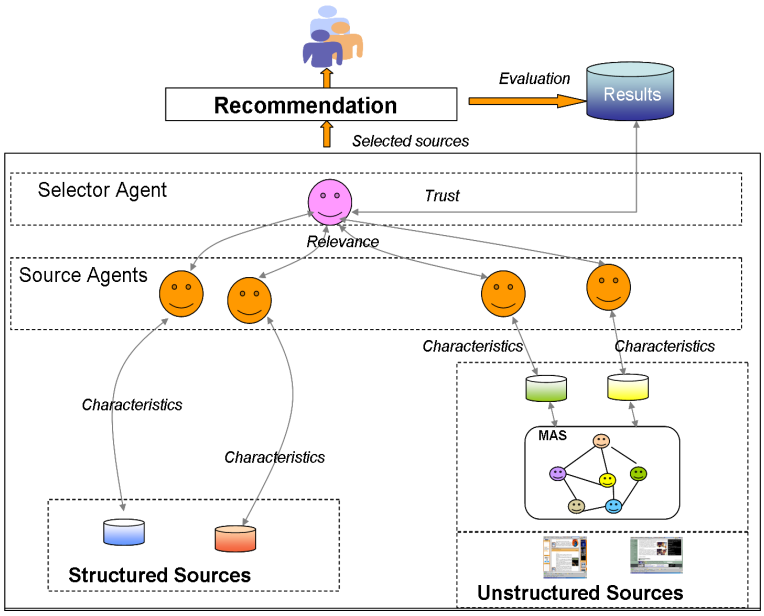


Figure 5.1: Multi agent system to implement the methodology proposed in this PhD thesis

networks. These recent developments make the Web more participatory, with users contributing more information and able to upload yet more personal information, such as their experiences, hobbies, videos, etc. All this information could be useful for recommender systems. In order for our methodology to be applied automatically, we would have to implement a multiagent system.

An intelligent agent can be defined as " A computer system, situated in some environment, that is able to do autonomous actions in order to achieve objectives." [Jennings and Wooldridge, 1998]. An agent is reactive and proactive responding to the changes produced in its environment and is collaborative due its capacity to interact with other agents. These properties of intelligent agents make them ideal for implementing the methodology.

Figure 5.1 shows the multiagent system that would be required to automate the methodology and which could be implemented in a real application for a recommender system in the Web. A first definition of this system can be found in [Aciar et al., 2005d] [Aciar et al., 2007a].

- **Source Agents**: each information source is manipulated by an agent, which is the one in charge of obtaining the characteristics of the source, calculating them using the equations defined in Chapter 3. Source agent knows the information contained

in its source. This agent works automatically, without human intervention.

- **Unstructured Source Agents**: since there may be unstructured sources, there needs to be a special type of agent for these sources. These agents will be responsible for obtaining and structuring the information from them. In order to structured the information have to be applied the method proposed in Chapter 4. Due to the complexity of the method, the source agent should be a multiagent system, composed by various agents to address the complexity in the implementation of the method. A more detailed analysis will be required when this part of the methodology will be implemented. Once the unstructured sources have been structured, it will be manipulated by a source agent as described in the previous point.

- **Selector Agent**: responsible for selecting the most suitable source for a recommender system. This agent employs the selection algorithm defined in Chapter 3. In order to carry out this task, it will have to calculate the relevance and trust for each one of the sources. In order to obtain the relevance, it requests information about the characteristics from source agent. In order to calculate the trust, it needs to read the result of previous recommendations using the data sources.

- **Recommender System**: In the Figure 5.1, the recommender system is modelled as a black box. The system input consists of the data sources provided by the selector agent and the output, of interest to us, is the result of the recommendations made with these sources, this information will be used for our methodology to calculate the trust measure. This black box could be a recommender system or a recommender agent, or in fact, any other method of recommendation could be used. As shown in Chapter 3, the methodology has given good results using two of the most popular methods in the recommender systems: Content Based Filtering (CBF) and Collaborative Filtering (CF).

These are the basic aspects that need to be taken into consideration when implementing the methodology in a real multi-agent system, although at the moment of implementing and due to the complexity of the application, it may be necessary to create other agents.

5.3.2 Web usage miming

The methodology could be applied using information about user access. The sources used in this proposal are data bases containing information on users' purchases and the web pages where they write text, but not with data obtained regarding access or navigation of web sites by the users.

5.3.3 Improving the URR method

According to the method used to obtain user's information from web page, some improvements need to be made to resolve certain issues that have arisen. For example, during the real case experiments, it was observed that quite often in the process of text mining, there are phrases that are complicated and cannot be classified in a particular category. These phrases should be broken down into shorter phrases so that a good classification can be made. One important aspect to consider is the definition of the ontology. In the particular case of digital cameras used in these experiments, the ontology needs to be more detailed. For example, we have only one concept to identify image quality, but this quality depends on several other concepts such as the camera's stabilizing system, the flash used, etc.

5.3.4 Applications in future lines of research

In addition to the work that needs to be done to resolve some of the problems encountered while implementing our proposals, the contributions of the present thesis are related to other lines of investigation:

- In relation to the research work whose objective is to confer an economic value to knowledge that is exchanged by agents in a multiagent system [Carrillo et al., 2006], the relevance of the information contained in the source can be considered as the value of the knowledge that the source agent contains. This relevance value, R, could serve as the exchange currency between the agents. In this way, the measurement of relevance could be used as another way to appraise and to exchange knowledge possessed by agents, in addition to those defined by Carillo et al, [Carrillo et al., 2006].

- S. Delfin et al, [S. Delfin et al., 2006] have studied issues to do with privacy and reputation in recommender systems. Generally, each recommender agent is associated to a user and this agent possesses information on the tastes and preferences of that user. To deal with privacy problems, Delfin defines the concept of "dissociation" in which a recommender agent can be associated to several users. Since each agent can constitute a source of user information, so we can apply our methodology to select agents with relevant user's information.

- And a last further line is the need to set up stable knowledge flow path among the selected agents by R if this set agents are continuously used. This knowledge flow path is called "Life Pipe ". Mechanisms to guarantee some properties such as quality, quantity and stability of the flow have to been used. Swarm Intelligence (SI)

could be analysed to model the interaction between these agents and to guarantee the Life Pipe proprieties.

5.4 Related publications

The work developed for this thesis has led to two publications in journals and several contributions to international conferences and congresses. These are listed below.

The two main contributions of this thesis have been published in the following international scientific journals:

- S. Aciar, D. Zhang, S. Simoff and J. Debenham, "Informed Recommender: A Recommender System That Bases Recommendations on Consumer Product Reviews". IEEE Intelligent Systems. Special Issue on Recommender Systems - May/June 2007. Vol. 22, No. 3, pp. 39-47
- S. Aciar, J. L. de la Rosa, M. Royo-Vela and C. Serarols-Tarrés . "Increasing effectiveness in e-commerce: recommendations applying intelligent agents". International Journal of Business and Systems Research. February 2007.

There follows a list of the contributions to international conferences and congresses as a consequence of the work carried out during this thesis:

- S. Aciar, J. L. de la Rosa and J. López Herrera. "Information Sources Selection Methodology for Recommender Systems Based on Intrinsic Characteristics and Trust Measure". AAAI'07 Workshop on Recommender Systems in e-Commerce. Vancouver, Canada. July 2007.
- S. Aciar, J. López Herrera and J. L. de la Rosa. "Recommendations Using Information from Selected Sources with the ISIRES Methodology". Frontiers in Artificial Intelligence and Applications. AI Research and Development vol. 146. pp. 258 -265. Ed. IOS Press. Amsterdam, The Netherlands, 2006.
- S. Aciar, D. Zhang, S. Simoff and J. Debenham, "Recommender System Based on Consumer Product Reviews," IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06), pp. 719-723, 2006.
- S. Aciar, D. Zhang, S. Simoff and J. Debenham, "Informed Recommender Agent: Utilizing Consumer Product Reviews through Text Mining," IEEE/WIC/ACM In-

ternational Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops)(WI-IATW'06) pp. 37-40, 2006.

- S. Aciar, J. López Herrera and J. L. de la Rosa."Integrating Information Sources for Recommender Systems". Frontiers in Artificial Intelligence and Applications AI Research and Development vol. 131; pp. 405-412. Ed. IOS Press .Amsterdam, The Netherlands, 2005.

- S. Aciar, J. López Herrera and J. L. de la Rosa. "SMA para la búsqueda inteligente de información para recomendar", I Congreso Español de Informática, (Cedi 2005), Simposio de Inteligencia Computacional, SICO'2005 (IEEE Computational Intelligence Society, SC) Granada, España, Septiembre, 2005.

- S. Aciar, J. López Herrera and J. L. de la Rosa. "Identifying Information from heterogeneous and distributed information sources for recommender systems". Fourth Mexican International Conference on Artificial Intelligence (MICAI 2005). Mexico. November, 2005.

- S. Aciar, J. López Herrera and J. L. de la Rosa. "FC en un SMA para seleccionar fuentes de información relevantes para recomendar". XI Conferencia de la Asociación Española para la Inteligencia Artificial (caepia'05). Workshop de Inteligencia Computacional: Aplicaciones en Marketing y Finanzas. Santiago de Compostela. Noviembre, 2005.

Bibliography

[Abbattista et al., 2002] Abbattista, F., Degemmis, M., Fanizzi, N., Licchelli, O., Lopes, P., Semeraro, G., and Zambetta, F. (2002). Learning customer profiles for content-based filtering in e-commerce. *Italian Artificial Intelligence Conference.*

[Aciar et al., 2007a] Aciar, S., de la Rosa, J. L., Royo-Vela, M., and Serarols-Tarrés, C. (2007a). Increasing effectiveness in e-commerce: recommendations applying intelligent agents. *International Journal of Business and Systems Research, to apperar in February 2007*, pages 308–315.

[Aciar et al., 2005a] Aciar, S., Herrera, J. L., and de la Rosa, J. L. (2005a). Fc en un sma para seleccionar fuentes de información relevantes para recomendar. *XI Conferencia de la Asociación Española para la Inteligencia Artificial (caepia'05). Workshop de Inteligencia Computacional: Aplicaciones en Marketing y Finanzas. Santiago de Compostela, España.*

[Aciar et al., 2005b] Aciar, S., Herrera, J. L., and de la Rosa, J. L. (2005b). Identifying information from heterogeneous and distributed information sources for recommender systems. *Fourth Mexican International Conference on Artificial Intelligence (MICAI 2005). Mexico.*

[Aciar et al., 2005c] Aciar, S., Herrera, J. L., and de la Rosa, J. L. (2005c). Integrating information sources for recommender systems. *Frontiers in Artificial Intelligence and Applications AI Research and Development*, 131:405–412.

[Aciar et al., 2005d] Aciar, S., Herrera, J. L., and de la Rosa, J. L. (2005d). Sma para la búsqueda inteligente de información para recomendar. *I Congreso Español de Informática, (Cedi 2005), Simposio de Inteligencia Computacional, SICO'2005 (IEEE Computational Intelligence Society, SC) Granada, España.*

[Aciar et al., 2006a] Aciar, S., Herrera, J. L., and de la Rosa, J. L. (2006a). Recommendations using information from selected sources with the isires methodology. *Frontiers in*

Artificial Intelligence and Applications AI Research and Development ISSN 0922-6389, 146:258–265.

[Aciar et al., 2006b] Aciar, S., Zhang, D., Simoff, S., and Debenham, J. (2006b). Informed recommender agent: Utilizing consumer product reviews through text mining. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops)(WI-IATW'06)*, pages 37–40.

[Aciar et al., 2006c] Aciar, S., Zhang, D., Simoff, S., and Debenham, J. (2006c). Recommender system based on consumer product reviews. *IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, pages 719–723.

[Aciar et al., 2007b] Aciar, S., Zhang, D., Simoff, S., and Debenham, J. (2007b). Informed 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. (JRC= 2.5).*

[Adomavicius, 2005] 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*, 17(6):734–749.

[Amazon, 2006] Amazon (2006). Amazon.com. available at [www.amazon.com] accessed on december 15, 2006.

[Ansari et al., 2000] Ansari, A., Essegaier, S., and Kohli, R. (2000). Internet recommendations systems. *Journal of Marketing Research*, 37(3):363–375.

[Babaguchi et al., 2003] Babaguchi, N., Ohara, K., and Ogura, T. (2003). Effect of personalization on retrieval and summarization of sports video. *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing*, 2:940–944.

[Balabanovic and Shoham, 1997] Balabanovic, M. and Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Comm. ACM*, 40(3):66–72.

[Billsus et al., 2002] Billsus, D., Brunk, C., Evans, C., Gladish, B., and Pazzani, M. (2002). Adaptive interfaces for ubiquitous web access. *Communications of the ACM*, 45(5):34–38.

[Billsus and Pazzani, 1998] Billsus, D. and Pazzani, M. (1998). Learning collaborative information filters. *In Proc. of the 15-th International Conference on Machine Learning, Wisconsin*, pages 46–54.

[Billsus and Pazzani, 2000] Billsus, D. and Pazzani, M. (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, 10(2-3):147–180.

[Boklaschuk and Caisse, 2001] Boklaschuk, K. and Caisse, K. (2001). Evaluation of educational web sites. resources. available www: [http://www.usask.ca/education/coursework/802papers/bokcaisse/bokcaisse.htm] accessed on december 11, 2006.

[Boll, 2002] Boll, S. (2002). Modular content personalization service architecture for e-commerce applications. *Advanced Issues of E-Commerce and Web-Based Information Systems, 2002. (WECWIS 2002). Proceedings. Fourth IEEE International Workshop*, pages 213–220.

[Burke, 2000] Burke, R. (2000). *Knowledge-Based Recommender Systems.* Number 32. Marcel Dekker, a. kent edition.

[Burke, 2002] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

[Calvanese et al., 1998] Calvanese, D., Giacomo, G., and Lenzerini, M. (1998). Information integration: Conceptual modeling and reasoning support. *In: CoopIS'98.*

[Calypool et al., 2001] Calypool, M., Brown, D., Le, P., and Waseda, M. (2001). Inferring user interest. *IEEE Internet Comput*, 5:32–39.

[Carenini et al., 2003] Carenini, G., Smith, J., and Poole, D. (2003). Towards more conversational and collaborative recommender systems. *In Proceedings of the 7th International Conference on Intelligent User Interfaces. IUI.*

[Carrillo et al., 2006] Carrillo, C., de la Rosa, J. L., Moreno, A., Muntaner, E., Delfin, S., and Canals, A. (2006). Social currencies and knowledge currencies. *Artificial Intelligence Research and Development*, 146:266–274.

[Castillo et al., 2003] Castillo, J., Silvescu, A., D.Caragea, J.Pathak, and Honavar, V. (2003). Information extraction and integration from heterogeneous, distributed, autonomous information sources - a federated ontology-driven query-centric approach. *IEEE International Conference on Information Reuse and Integration (IRI 2003)*, pages 183–191.

[Castillo, 2002] Castillo, J. R. (2002). Ontology-driven information extraction and integration from autonomous, heterogeneous, distributed data sources - a federated query-centric approach. *Masters Thesis.*

[Chevalier and Mayzlin, 2003] Chevalier, J. and Mayzlin, D. (2003). The effect of word of mouth on sales: Online book reviews. *NBER Working Paper Series, National Bureau of Economic Research, USA.*

[Cho et al., 2002] Cho, Y., Kim, J., and Kim, S. (2002). A personalized recommender system based on web usage mining and decision tree. *Expert Systems with Applications*, 23:329–342.

[Ciocoiu et al., 2001] Ciocoiu, M., Nau, D. S., and Gruninger, M. (2001). Ontologies for integrating engineering applications. *Journal of Computing and Information Science in Engineering*, 1(1):12–22.

[Clerkin et al., 2002] Clerkin, P., Hayes, C., and Cunningham, P. (2002). Automated case generation for recommender systems using knowledge discovery techniques.

[Cosley et al., 2003] Cosley, D., Lam, S., Albert, I., Konstan, J., and Riedl, J. (2003). Is seeing believing? how recommender interfaces affect users opinions. *CHI Lett*, 5.

[Curien et al., 2006] Curien, N., Fauchart, E., Laffond, G., and Moreau, F. (2006). *Online consumer communities: escaping the tragedy of the digital commons.* Cambridge University Press.

[Cykana et al., 1996] Cykana, P., A., and Stern, M. (1996). Dod guidelines on data quality management. In Wang, R. Y., editor, *Conference on Information Quality (IQ)*, pages 154–171. MIT.

[Dellarocas, 2003] Dellarocas, C. (2003). The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10):1407–1424.

[Deshpande and Karypis, 2004] Deshpande, M. and Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177.

[Edwards, 1998] Edwards, J. (1998). The good, the bad and the useless: Evaluating internet resources. *Available WWW: [http://www.ariadne.ac.uk/issue16/digital] Accessed on December 11, 2006.*

[Ehrig and Sure, 2004] Ehrig, M. and Sure, Y. (2004). Ontology mapping - an integrated approach. *Proceedings of the First European Semantic Web Symposium, LNCS*, 3052:76–91.

[Endo and Noto, 2003] Endo, H. and Noto, M. (2003). A word-of-mouth information recommender system considering information reliability and user preferences. *IEEE International Conference on Systems, Man and Cybernetics*, 3:2990– 2995.

[Española, 2006] Española, R. A. (2006). Amazon.com. available at [http://buscon.rae.es] accessed on december 2, 2006.

[Evans et al., 2006] Evans, A., Fernandez, M., Vallet, D., and Castells, P. (2006). Adaptive multimedia access: from user needs to semantic personalisation. *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS 2006)*, page 4.

[Farzan and Brusilovsky, 2006] Farzan, R. and Brusilovsky, P. (2006). Social navigation support in a course recommendation system. In Wade, V., Ashman, H., and Smyth, B., editors, *Adaptive Hypermedia and Adaptive Web-Based Systems: 4th International Conference, AH 2006*, volume 4018 of *Lecture Notes in Computer Science*, pages 91–100, Berlin. Springer Verlag.

[Felfernig et al., 2007] Felfernig, A., Friedrich, G., Jannach, D., and Zanker, M. (2006-2007). An integrated environment for the development of knowledge-based recommender applications. *International Journal of Electronic Commerce*, 11(2):11.

[Goh, 1997] Goh, C. H. (1997). Representing and reasoning about semantic conflicts in heterogeneous information sources. *Phd thesis MIT*.

[Group, 2006] Group, I. S. (2006). Ibm websphere information integration. *white paper from www-306.ibm.com/software/data/integration.*

[Gruber, 1993] Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

[Guarino et al., 1999] Guarino, N., Masolo, C., and Vetere, G. (1999). Ontoseek: content-based access to the web. *IEEE Intelligent Systems*, 14:70–80.

[Haübl and Trifts, 2000] Haübl, G. and Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, 19(1):4–21.

[Herlocker et al., 2000] Herlocker, J., Konstan, J., and Riedl, J. (2000). Explaining collaborative filtering recommendations. *ACM Conf. Computer Supported Cooperative Work*, pages 241–250.

[Herlocker et al., 2004] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Information Systems*, 22(1):5–53.

[Hilderman and Hamilton, 2000] Hilderman, R. and Hamilton, H. (2000). Principles for mining summaries using objective measures of interestingness. *In Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00). Vancouver, BC.*, pages 72–81.

[Hughes, 2000] Hughes, A. (2000). Strategic database marketing : The master plan for starting and managing a profitable. *Customer-Based Marketing Program.*

[Ilebrekke, 2002] Ilebrekke, X. (2002). A comparative study of ontology languages and tools. *Proceedings of the 14th International Conference on Advanced Information Systems Engineering*, pages 761–765.

[Jarrar, 2005] Jarrar, M. (2005). Towards methodological principles for ontology engineering. *PhD Thesis. Vrije Universiteit Brussel., Brussels.*

[Jennings and Wooldridge, 1998] Jennings, N. and Wooldridge, M. (1998). *Agent Technology: Foundations, Applications and Markets.* Springer Verlag.

[Kalfoglou and Schorlemmer, 2005] Kalfoglou, Y. and Schorlemmer, M. (2005). Ontology mapping: The state of the art. In Kalfoglou, Y., Schorlemmer, M., Sheth, A., Staab, S., and Uschold, M., editors, *Semantic Interoperability and Integration*, number 04391 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

[Kibum et al., 2002] Kibum, K., Carroll, J., and Rosson, M. (2002). An empirical study of web personalization assistants supporting end-users in web information systems. *Human Centric Computing Languages and Environments, 2002. Proceedings. IEEE 2002 Symposia*, pages 60–62.

[Kirk et al., 1995] Kirk, T., Levy, A., Sagiv, Y., and Srivastava, D. (1995). The information manifold. *In Proceedings of the AAAI Spring Symposium: Information Gathering from Heterogeneous Distributed Environments*, pages 85–91.

[Knoblock et al., 2001] Knoblock, C., Minton, S., Ambite, J., Ashish, N., Muslea, I., Philpot, A., and Tejada, S. (2001). The ariadne approach to web-based information integration. *International Journal on Intelligent Cooperative Information Systems (IJCIS)*, 10(1):145–169.

[Koeller, 2001] Koeller, A. (2001). Integration of heterogeneous databases: Discovery of meta-information and maintenance of schema-restructuring views. *Dissertation Degree of Doctor of Philosophy in Computer Science.*

[Levy, 2000] Levy, A. (2000). Logic-based techniques in data integration. *Logic Based Artificial Intelligence, Edited by Jack Minker Kluwer.*

[Linden et al., 2003] Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing, 2003.*

[Marlin, 2003] Marlin, B. (2003). Modeling user rating profiles for collaborative filtering. *Proc. 17th Ann. Conf. Neural Information Processing Systems (NIPS 03).*

[Massa and Avesani, 2004] Massa, P. and Avesani, P. (2004). Trust-aware collaborative filtering for recommender systems. *International Conference on Cooperative Information Systems (CoopIS).*

[Melville et al., 2002] Melville, P., Mooney, R., and Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. *Proc.18th Nat'l Conf. Artificial Intelligence.*

[Michalowski et al., 2004] Michalowski, M., Ambite, J., Thakkar, S., Tuchinda, R., Knoblock, C., and Minton, S. (2004). Retrieving and semantically integrating heterogeneous data from the web. *IEEE Intelligent Systems*, 19(3):72–79.

[Michelson and Knoblock, 2006] Michelson, M. and Knoblock, C. A. (2006). Phoebus: A system for extracting and integrating data from unstructured and ungrammatical sources. *In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), Intelligent Systems Demonstrations, Boston, MA.*

[Middleton et al., 2002] Middleton, S., Alani, H., Shadbolt, N., and Roure, D. (2002). Exploiting synergy between ontologies and recommender systems. *Proceedings of Semantic Web Workshop 2002 At the Eleventh International World Wide Web Conference Hawaii.*

[Middleton et al., 2004] Middleton, S., Shadbolt, N., and de Roure, D. (2004). Ontological user profiling in recommender systems. *ACM Trans. Information Systems*, 22(1):54–88.

[Montaner et al., 2002] Montaner, M., López, B., and de la Rosa, J. (2002). Developing trust in recommender agents. *In Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02). Cristiano Castelfranchi and W. Lewis Johnson (Eds). Bologna (Italy).*, ACM Press. 1:304–305.

[Mooney and Roy, 2000] Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. *Conference on Digital Libraries archive Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204.

[Motro and Rakov, 1998] Motro, A. and Rakov, I. (1998). Estimating the quality of databases. *Flexible Query Answering Systems*, page 298.

[MovieLens, 2006] MovieLens (2006). Movielens available at [http://movielens.umn.edu]. accessed on december 15, 2006.

[Naumann, 2004] Naumann, F. (2004). Information quality criteria. In *Quality-Driven Query Answering for Integrated Information Systems*, volume 146, pages 29–30. Springer Berlin / Heidelberg.

[Naumann et al., 2004] Naumann, F., Freytag, J., and Leser, U. (2004). Completeness of integrated information sources. *Information Systems*, 29:583–615.

[Nguyen and Ricci, 2004] Nguyen, Q. and Ricci, F. (2004). User preferences initialization and integration in critique based mobile recommender systems. *Proceedings of Workshop on Artificial Intelligence in Mobile Systems 2004. In conjunction with UbiComp 2004, Nottingham, UK.*

[Noriega et al., 1998] 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).*

[Noy and M.A.Musen, 2002] Noy, N. and M.A.Musen (2002). Evaluating ontology-mapping tools: Requirements and experience. *In Workshop on Evaluation of Ontology Tools at EKAW'02 (EON2002).*

[Noy, 2004] Noy, N. F. (2004). Semantic integration: a survey of ontology-based approaches. *ACM SIGMOD Record*, 33(4):65–70.

[Oard and Kim, 1998] Oard, D. and Kim, J. (1998). Implicit feedback for recommender systems. *Proceedings of the AAAI Workshop on Recommender Systems.*

[O'Donovan and Smyth, 2005] O'Donovan, J. and Smyth, B. (2005). Trust in recommender systems. *Proceedings of the 10th international conference on Intelligent user interfaces San Diego, California*, pages 167–174.

[Palmer, 2002] Palmer, J. (2002). Designing for web site usability. *IEEE Computer*, 35(7):102–103.

[Patel et al., 1998] Patel, J., Teacy, W. T. L., Jennings, N. R., and Luck, M. (1998). A probabilistic trust model for handling inaccurate reputation sources. *In Proceedings of Third International Conference on Trust Management , Rocquencourt, France.*

[Pazzani, 1999] Pazzani, M. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, pages 393–408.

[Rashid et al., 2002] Rashid, M., amd D. Cosley, I. A., Lam, S. K., McNee, S. M., Konstan, J. A., and Riedl, J. (2002). Getting to know you: learning new user preferences in recommender systems. *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134.

[Ricci et al., 2002] Ricci, F., Arslan, B., Mirzadeh, N., and Venturini, A. (2002). Itr: a case-based travel advisory system. *Proceedings of the 6th European Conference on Case Based Reasoning [ECCBR 2002], Aberdeen, Scotland.*

[Ricci and Wietsma, 2006] Ricci, F. and Wietsma, R. T. A. (2006). Product reviews in travel decision making. *Information and Communication Technologies in Tourism Proceedings of the International Conference in Lausanne, Switzerland, Springer Verlag*, pages 296–307.

[Rieh, 2002] Rieh, S. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161.

[S. Delfin et al., 2006] S. Delfin, C. C., Muntaner, E., Moreno, A., Ibarra, S., and de la Rosa, J. L. (2006). Improving privacy of recommender agents by means of full dissociation. *Artificial Intelligence Research and Development*, pages 308–315.

[Salton and Buckley, 1988] Salton, G. and Buckley, V. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

[Schafer et al., 1999] Schafer, J., Konstan, J., and Riedl, J. (1999). Recommender systems in ecommerce. *Proceedings of the 1st ACM conference on Electronic commerce Denver, Colorado, United States.*, pages 158 –166.

[Schein et al., 2002] 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*, pages 253–260.

[Schrah et al., 2006] Schrah, G. E., Dalal, R., and Sniezek, J. A. (2006). No decision-maker is an island: integrating expert advice with information acquisition. *Journal of Behavioral Decision Making*, 19:43–60.

[Semeraro et al., 2005] Semeraro, G., Lops, P., and Degemmis, M. (2005). Wordnet-based user profiles for neighborhood formation in hybrid recommender systems. pages 291–296.

[Senecal and Nantel, 2004] Senecal, S. and Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80, Elsevier*, 80:159–169.

[Shahabi and Chen, 2003] Shahabi, C. and Chen, Y. S. (2003). An adaptive recommendation system without explicit acquisition of user relevance feedback. *Distributed and Parallel Databases*, 14(2):173–192.

[Smyth and Cotter, 2000] Smyth, B. and Cotter, P. (2000). A personalized television listings service. *Communications of the ACM.*

[Stabb et al., 2002] Stabb, S., Werther, H., Ricci, F., Zipf, A., Gretzel, U., Fesenmaier, D., Paris, C., and Knoblock, C. (2002). Intelligent systems for tourism. *IEEE Intelligent Systems*, 17(6):53–66.

[Stuart et al., 2002] Stuart, M., Harith, A., Nigel, S., and Dave, D. (2002). Exploiting synergy between ontologies and recommender systems. *In Proceedings Semantic Web Workshop, Hawaii, USA.*

[Tam and Ho, 2003] Tam, K. Y. and Ho, S. Y. (2003). Web personalization: is it effective? *IT Professional*, 5(5):53–57.

[Towle and Quinn, 2000] Towle, B. and Quinn, C. (2000). Knowledge based recommender systems using explicit user models. *In Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, AAAI Technical Report WS-00-04*, pages 74–77.

[Vézina and Militaru, 2003] Vézina, R. and Militaru, D. (2003). Collaborative filtering: Theoretical positions and research agenda in marketing. *Work paper 2003-012, Faculté des sciences de l'administration, Université Laval, Quebec, Canada.*

[Weiss et al., 2004] Weiss, M. S., Indurkhya, N., Zhang, T., and Damerau, F. (2004). Text mining: Predictive methods for analyzing unstructured information. *Springer-Verlag, New York.*

[Wietsma and Ricci, 2005] Wietsma, R. and Ricci, F. (2005). Product reviews in mobile decision aid systems. *Workshop on Pervasive Mobile Interaction Devices, in conjunction with Pervasive 2005, Munich, Germany.*

[wikipedia.org, 2007] wikipedia.org (2007). Available at [http://en.wikipedia.org] accessed on january 5, 2007.

[Wu et al., 2003] Wu, D., Im, I., Tremaine, M., Instone, K., and Turoff, M. (2003). A framework for classifying personalization scheme used on e-commerce websites. *Proceedings of the 36th Annual Hawaii International Conference*, page 12.

[Xu et al., 2005] Xu, B., Zhang, M., Pan, Z., Yang, H., Gervasi, O., Gavrilova, M., Kumar, V., Lagana, A., Lee, H., Mun, Y., Taniar, D., and Tan, C. (2005). Content-based recommendation in e-commerce. *Lecture notes in computer science (Lect. notes comput. sci.) ISSN 0302-9743.*

[Yang et al., 2004] Yang, W., Wang, Z., and You, M. (2004). An improved collaborative filtering method for recommendations' generation. *in Proc. of IEEE International Conference on Systems, Man and Cybernetics.*

[Yu and Singh, 2002] Yu, B. and Singh, M. P. (2002). Towards a probabilistic model of distributed reputation management. *4th Workshop on Deception, Fraud and Trust In Agent Societies. Montreal.*

[Yu and Singh, 2003] 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.

[Zaýane et al., 2003] Zaýane, O., Srivastava, J., M.Spiliopoulou, and Masand, B. M. (2003). Mining web data for discovering usage patterns and profiles. *Proc. WEBKDD 2002. O.R. Zaýane, J. Srivastava, M.Spiliopoulou, B. M. Masand, eds.*

[Zhang and Simoff, 2006] Zhang, D. and Simoff, S. (2006). Informing the curious negotiator: Automatic news extraction from the internet. *Data Mining: Theory, Methodology, Techniques, and Applications. Williams, G. and Simoff, S. (eds)*, pages 176–191.