Selecting Information Sources for Collaborative Filtering and Content-Based Filtering

Silvana Aciar[1], Josefina López Herrera[1] and Josep Lluis de la Rosa[1]
[1] Departament d' Electronica, Informatica i Automatica
Univertitat de Girona
17071 Campus Montilivi
Girona, Spain
{saciar, peplluis} @eia.udg.es, lopez.herrera@udg.es

Abstract. This paper addresses the problem of identifying and selecting relevant sources of information in order to enhance the accuracy of recommender systems. Recommender systems suggest to the users the items they will probably like. The large amount of information available nowadays on Internet makes the process of detecting user's preferences and selecting recommended products more and more difficult. In this paper we present a methodology to identify and select sources holding relevant information for recommender systems. This methodology is applied using two recommender methods: Content-Based Filtering (CBF) and Collaborative Filtering (CF) and showed in a real case-study, how the accuracy of the recommendations made with these methods and the selected sources increase.

1 Introduction

Information overload is one of the most important problems met by the Internet's users nowadays. The great amount of old and new information to analyze, contradictions in the available information generate noise and make difficult the identification process of relevant information. The information overload phenomena is determined by the lack of methods to compare and process the available information Recommender Systems address this problem filtering the most relevant information for the user's purpose. These systems receive as input the preferences of the users, analyze them and deliver recommendations.

Recommender systems are used in a network overloaded of information. In such a case, the search of specific information for recommender systems is a difficult task. The literature in the field of recommender systems is focused towards the methods that are used to filter the information to make the recommendations. Methods such as Content-Based Filtering (CBF) [4] [5] and Collaborative Filtering (CF) [8][9] are significant examples in this field. This paper presents a methodology for the identification of information sources, comparing the sources and to selecting most relevant information to make recommendations. This methodology allows the optimization of the search of the information obtained only from the relevant sources for the recommendations. A Multi-Agent System (MAS) called ISIIRES (Identifying, Selecting, and Integrating Information for Recommender Systems) has been designed for this

© A. Gelbukh, S. Torres, I. López (Eds.)
Advances in Artificial Intelligence
Research in Computing Science 20, 2006, pp. 119-130

purpose [2]. ISIIRES identifies information sources based on a set of intrinsic characteristics and selects the most relevant to be recommended. In this paper we present the result obtained applied the methodology using two recommender methods: CBF and CF and showed in a real case-study, how the accuracy of the recommendations increases.

The paper is organized as follows. In Section 2 are described the ISIIRES methodology and the Multi-Agent System to implement the methodology. In Section 3, the application of the recommender methods is showed. In Section 4 a Case Study with some result is described and finally, conclusions are drawn in Section 5.

2 Identifying, Selecting, Integrating Information for Recommender Systems (ISIIRES)

A methodology for the identification of information sources, comparing the sources and to selecting most relevant information to make recommendations has been proposed and has been described by Aciar et.al. in [2]. Four blocks compose the methodology which is shown in figure 1.

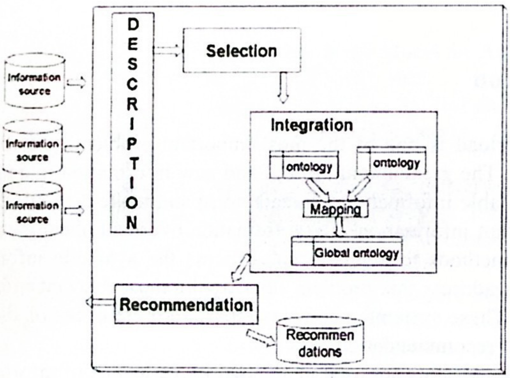

Fig. 1. Functional blocks of the methodology

2.1 Description

A set of intrinsic characteristics of the sources has been defined for the identification of relevant information to make recommendations [3]. These characteristics are:

- an abstract representation of the information contained in the sources
- criteria to compare and to select the sources.

2.2 Selection

The selection of the information sources is made based on the intrinsic characteristics and a value of trust from last recommendations. An initial value of trust = 1 is assigned to the sources that have not been used in last recommendations [1].

2.3 Integration

Mapping between the ontologies is made to integrate the information from the selected sources which are distributed and heterogeneous. The mapping has been made defining a global ontology comparing the ontologies of each source looking for similar concepts among them [2].

2.4 Recommendation

The methodology is focussed on the selection of the sources. When the relevant information sources have been selected it is possible to apply any of recommender methods. In this paper the methodology is applied using two recommender methods: CBF and CF. The Multi-Agent System (MAS) shown in figure 2 has been designed to implement the methodology [3].

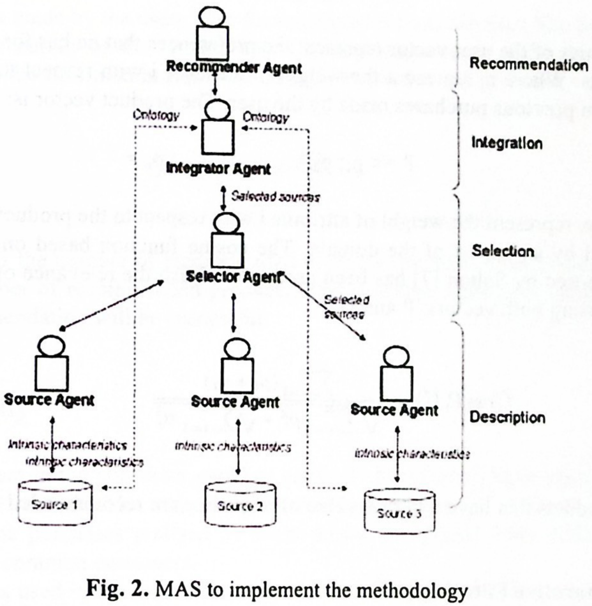

Fig. 2. MAS to implement the methodology

Four types of agents compose the MAS: Source Agent (SA), Selector Agent (SEA), Integrator Agent (IA) and Recommender Agent (RA).

- The SA managers the sources obtaining intrinsic characteristics.
- The SEA selects the sources that provide the most relevant information to make the recommendation based on the intrinsic characteristics and a trust value.
- The IA integrates the information from selected sources establishing a mapping between the ontologies of the selected sources.
- The RA is an interface agent that interacts with the users. Once the sources have been selected, it makes the recommendations and evaluates the results to keep them for future selections.

3 Applying Different Methods to Recommend

Two main methods have been used to compute recommendations: Content-Based Filtering and Collaborative Filtering, these methods are implemented in this paper with information from the selected sources.

3.1 Content-Based Filtering (CBF)

In this method the attributes of products are extracted and they are compared with a user profile (preferences and tastes). Vectors are used to represent the user profile and products in this work. The user vector is:

$$U =< u_1; u_2; \ldots\ldots\ldots\ldots; u_n >$$

The values of the user vector represent the preferences that he has for the attributes of products. Where u_i represent the weight of attribute i with respect to the user obtained from previous purchases made by the user. The product vector is:

$$P =< p_1; p_2; \ldots\ldots\ldots\ldots\ldots; p_n >$$

Where p_i represent the weight of attribute i whit respect to the product. This weight is assigned by an expert of the domain. The cosine function based on the vectorial space proposed by Salton [7] has been used to establish the relevance of products for the users using both vectors: P and U

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

The products that have a higher value of relevance are recommended the users.

3.2 Collaborative Filtering (CF)

The information provided by users with similar interests or necessities is used to determine the relevance that the products have for the user. Similarity between users is calculated for this purpose and the recommendations are made based only in this similarity, the bought products are not analyzed as it is made in the FBC. In this paper the cosine vector similarity [7] is used to compute the distance between the representation of the present user and the other users. All users are represented by vectors.

$$U =< u_1; u_2; \ldots\ldots\ldots\ldots; u_n >$$

Where u_i are the preferences of the user which is represented by the weight assigned by him to any attribute of the product, such as colour, type of product, etc. The similarity measurement is calculated by:

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

Where U and V are the user vectors.

3.3 Evaluating Recommendations

The purchases made by the users after the recommendations are used like feedback to evaluate the accuracy of the system. The accuracy is evaluated using the precision equation from the Information Retrieval field [6] adapted to our problem.

$$Precision = \frac{Pr}{R} \quad (3)$$

Where Pr is the number of recommended products that have been bought and R is the total number of recommended products. The precision represents the probability that a recommendation will be successful.

4 Case-Study

Eight data bases in the consumer package goods domain (retail) have been used. The data bases are related tables containing information of the retail products, 1200 customers and the purchases realized by them during the period 2001-2002. All data bases contain common customers.

The sources used in the experiments are the sources selected in the previous phase of the methodology, see Aciar et. al [1] for more detail about the selection of the sources.

4.1 Content-Based Filtering (CBF))

An expert of the supermarket has defined the relevant attributes of the product used in the CBF method. Based in these attributes shown in figure 3 has been established the user preferences represented by a vector:

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

Fig. 3. Relevant attributes in the consumer package goods domain (retail) defined by an expert.

$$U =< u_1; u_2; \ldots\ldots\ldots\ldots; u_n >$$

The weight ui has been obtained from last purchases of the user using the tf-idf method (Term Frequency Times Inver Document Frequency) [6].

$$u_i = t_i * log_2(\frac{N}{n_i}) \quad (4)$$

Where ti is the frequency of attribute i in the purchases, ni is the number of users who have been bought a product with attribute i and N is the total number of users. The weight pi of the product vector has been assigned by the expert in the supermarket.

$$P =< p_1; p_2; \ldots\ldots\ldots\ldots; p_n >$$

The weights ui and pi of each vector are shown in figure 4. 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 5).

The products with a value of relevance > 6 have been recommended the users. Theses recommendation are shown in figure 6.

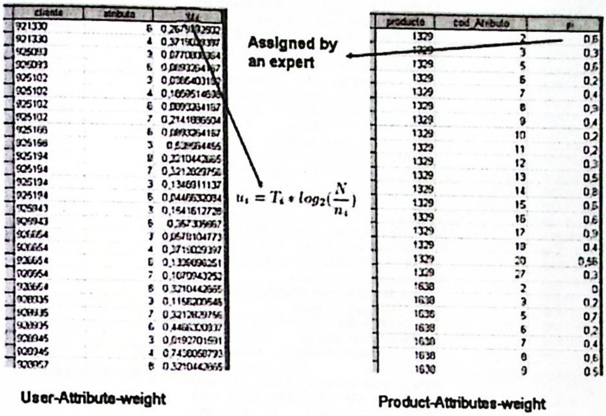

Fig. 4. Weights to obtain the user vector and the product vector

cliente	producto	Relevance
91850	32318	0,3633179926
91850	34971	0,3633179926
91850	34864	0,3166948446
91850	34781	0,6468153614
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
921330	59930	0,8585791209

Fig. 5. Relevance of the products for the users

cliente	producto_recom	Descripcion
561339	92750	SERVILL BLANCA CAPRABO
561339	166960	CREMA CACAO 1 C 475 G NOCILLA
561339	23998	COCTEL DE FRUTAS 1 L GRANINI
561339	61256	MAGDALENA VALENC 350 G CAPRABO
561339	113700	BIO C/F BOSQUE X4 500G DANONE
561339	121764	ACEITE OLIVA 1 1 L CAPRABO
561339	115678	BEBIDA SOJA CALCIO 1L DIET-RADIS
561339	127622	SOLUBLE NAT. EXT 100 G CAPRABO
561339	143112	LATA SIN CAFEINA 33CL COCA COLA
561339	936000	AJOS SECOS 1/4 I
561339	63473	CERVEZA LATA 33CL CAPRABO
561339	542310	PATATAS TRADICION 170G EAGLE SNAC
561339	479628	MOZZARELLA RALL. 200 G M.D
561339	274372	VINO TINTO BRIK 1 L DON SIMON
561339	19122	CAVA BRUT NATURE 75CL C CARALT
561339	544330	PAN FRESCO FAMILI 600 SILUETA
561339	133361	O MOZZARELLA LONCHAS 2 CAPRABO
561339	23170	12 ROLLO HIGIENI CAPRABO
561339	124971	ACEITE VIRGEN EXTRA 1 YBARRA
561339	51664	MAQ.DESECH.LADY EXTRA-II
561339	121192	LATA CERVEZA 33CL RGEMESSTER
561339	63269	PAN DE LECHE 400 G BELLA EASO
561339	440362	ZUMO MELOCOTON×3 600ML KASFRUIT
561339	9000003	Frutas y Verduras
561339	61256	MAGDALENA VALENC 350 G CAPRABO
561339	9000003	Frutas y Verduras
561339	143112	LATA SIN CAFEINA 33CL COCA COLA
561339	510004	LAUREL BOLSA 10 G DANI
561339	24677	CHAMPU NORMAL 400ML WELLA BALS
561339	143111	LATA 33CL COCA COLA
561339	483052	PARMESAN RALLADO 40 G KRAFT
561339	374306	TOMATE TRITURADO 410 G APIS
561339	121690	ACEITE OLIVA 1 L LA GLORIA

Fig. 6. Recommendations made using CBF

4.2 Collaborative Filtering (CF)

The attributes shown in figure 3 defined by the expert of the supermarket have been used to obtain the user vectors.

$$U = < u_1; u_2; \ldots\ldots\ldots\ldots; u_n >$$

The weight ui has been obtained from last purchases of the user using the tf-idf method (Term Frequency Times Inver Document Frequency) [6] like in the CBF

$$u_i = t_i * log_2(\frac{N}{n_i}) \tag{5}$$

Where ti is the frequency of attribute i in the purchases, ni is the number of users who have been bought a product with attribute i and N is the total number of users. The weights ui obtained for each user are shown in figure 7.

The similarity between users has been established with equation 2 using the vectors representing the users (See figure 8)

U1_cliente	U1_codi_Atri	u_i
1000285	2	0,33268222156325
1000285	3	0,01974198931049
1000285	4	0,18281134822244
1000285	5	0
1000285	6	0,17524844956754
1000285	7	0,09670115880460B
1000285	7	0,19340231760921B
1000285	8	0,31207210034415A
1000285	9	0,45265841221335B
1000285	11	0
1000285	13	0
1000285	14	0,58611256312311
1000285	15	0,14875623213221
1000285	16	0
1000285	17	0
1000285	18	0,236589655555
1000285	19	0,2155145112452
1000285	20	0

cliente	Cod_Atributos	u_i
2998776	3	0,0394839706
2998339	6	0,0438121112
2998366	6	0,0438121112
299807	6	0,0438121112
2998515	6	0,0438121112
299840	3	0,0692259679
2998766	3	0,0789679572
2998022	3	0,0789679572
2998045	3	0,0789679572
299807	3	0,0789679572
2998339	3	0,0789679572
2997940	3	0,0789679572
2998045	6	0,0876242225
2998743	6	0,0876242225
2998045	7	0,0967011500
2998515	7	0,0967011500
299819	7	0,0967011500
299840	7	0,0967011500

$$u_i = T_i * log_2(\frac{N}{n_i})$$

User-Attribute-weight **User-Attribute-weight**

Fig. 7. Weights for the user vectors – CF

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

Fig. 8 Similarity between users - CF

The products bought by other users with a value of similarity > 6 have been recommended to the user. Theses recommendation are shown in figure 9.

Cliente	Productos_recomendados	Descripcion
1000285	931830	PLATANO CANARIO EXTRA BANDEJA
1000285	39667	PAGES TALLAT 430GUN ROCAS 76
1000285	1818	ESPINACA CORTADA 400 G FINDUS
1000285	68155	MAYONESA FRASCO 450ML CAPRABO
1000285	131086	TOMATE MONTSERRAT BJA
1000285	36072	CAMINO MEGATRUCK C/PAL GOZAN 138
1000285	455393	MERMEL NARANJA 350 G HELIOS
1000285	124928	ENSALADILLA 750 G FINDUS
1000285	16638	TORTILLA PATATA/CEBOLL G CAMPI#A
1000285	68386	FRESH TABLETS BREF WC
1000285	130690	PEPINILLOS 345 G HELIOS
1000285	412400	PASAS CALIFORNIA 200 G CAPRABO
1000285	59906	SALSA LIGERA 225ML YBARRA
1000285	13470	ESTRO+ESP.NO RAY SCOTCH BRI
1000285	22862	MACARRONES HUEVO 250 G EL PAVO
1000285	110646	PATATAS CHURRERIA 170 CAPRABO
1000285	8493	TOMATE FRITO BRK 400 G CAPRABO
1000285	127515	PIMIENTOS ROJOS 185 G CAPRABO
1000285	121665	ACEITE OLIVA 0,4 1 L BORGES
1000285	22858	SPAGHETTI HUEVO 250 G EL PAVO
1000285	15426	GUAN.SATINADO P SCOTCH BRI
1000285	113727	Q.DESN.NATURALX2150 G DANONE
1000285	4022	DET.LIQUIDO MAO 1000ML FINOSEDIL
1000285	179100	HARINA 1000 G CAPRABO
1000285	456908	CONFT.MANZANA 345 G HERO
1000285	340785	MEJILLON ESCABX3 240 G RIANXEIRA
1000285	46127	BIOF.DES.CO/MUX4 500 G DANONE
1000285	281557	VINO BLANCO PESC 75CL PERELADA
1000285	122939	PLATANOS FLOW-PACK
1000285	13285	ESTROP.VERDE GTE SCOTCH BRI
1000285	13966	FILETE TARRO 100 G AZKUE
1000285	466344	CALDO PESCADO08P 84 G KNORR

Fig. 9. Recommendations made using CF

4.3 Evaluating Recommendations

The experiments have been made implementing both methods: CBF and CF with the information of all the sources (8 data bases) without the methodology. The precision of recommendations has been evaluated using equation 3 obtaining the results shown in figure 10 and figure 11.

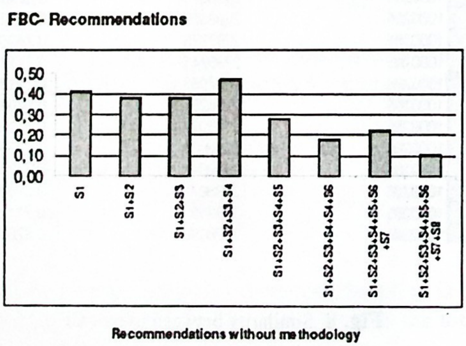

Fig. 10. Accuracy of the recommendations using CBF with all sources

In figures 12 and 13 can be observed the accuracy of the recommendations made using the CBF and the CF with information of the selected sources in the methodology. In the graphs are showed, how the accuracy of the recommendations made with these methods and the selected sources increase. The selection of the sources is established based on intrinsic characteristics of each source.

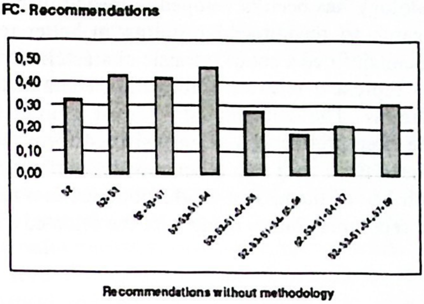

Fig. 11. Accuracy of the recommendations using CF with all sources

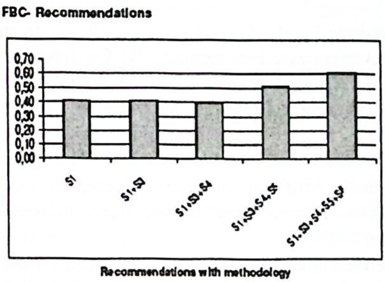

Fig. 12. Accuracy of the recommendations using CBF only with the selected sources in the methodology

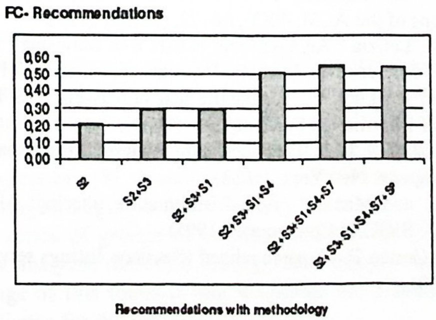

Fig. 13. Accuracy of the recommendations using CF only with the selected sources in the methodology

5 Conclusions

The large amount of information available nowadays on Internet makes the process of detecting user's preferences and selecting recommended products more and more difficult. A methodology has been developed to make this task easier and to optimize the information search to recommend resulting in better recommendations. In the methodology has been defined a set of intrinsic characteristics of information sources. The characteristics represent relevant information contained in the sources used to make recommendations. The selection of relevant sources is made based in these characteristics. The user preferences are established from the selected sources. The methodology has been used with two recommender methods: CBF and FC obtaining good results in each one of them. The results obtained in a real case-study show how the accuracy of the recommendation made with the selected sources increase.

Referentes

1. Aciar, S, Lpez Herrera, J and . de la Rosa, J.Ll '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, (Noviembre 2005).
2. Aciar, S, Lpez Herrera, J and . de la Rosa, J.Ll 'Integrating information sources for recommender systems', Artificial Intelligence Research and Development - IOS Press. VIII Catalan Association for Artificial Intelligence Congress (CCIA 2005),Alghero, Sardinia, Italy, (October 2005).
3. Aciar, S, Lpez Herrera, J and . de la Rosa, J.Ll 'Sma para la búsqueda inteligente de información para recomendar', I Congreso Español de Informática,(Cedi 2005), Simposio de Inteligencia Computacional, SICO2005 (IEEE Computational Intelligence Society, SC)Granada, Espana, (Septiembre 2005).
4. Balabanovic,M and Shoham, Y. 'Fab: Content-based, collaborative recommendation', Communications of the ACM, 40(3), 66–72, (March 1997).
5. Lieberman, H. 'Letizia : An agent that assists web browsing', Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, (1995).
6. Salton, G. and Buckley, C. 'Automatic text processing', he Transformation, Analysis and Retrieval of Information by Computer, (1989).
7. Salton, G. and McGill, M.J. 'Introduction to modern information retrieval', McGraw-Hill Publishing Company, NewYork, (1983).
8. Shardanand, U. and Maes, P. 'Social information filtering: Algorithms for automating word of mouth', SIGCHI Conference, (1995).
9. Smyth, B. and Cotter, P. 'A personalized television listings service', Communications of the ACM, (2000).