

Integrando Información de Fuentes Relevantes para un Sistema Recomendador

Silvana Aciar , Josefina López Herrera and Javier Guzmán Obando

Agents Research Laboratory

University of Girona

{saciar, jguzmano}@eia.udg.es, josefina.lopez@udg.es

Abstract. Este artículo presenta un marco de trabajo para la integración inteligente de un número de diferentes fuentes de información entradas para el proceso de recomendación basado en cualquier método de recomendación como lo es el Filtrado Colaborativo (FC) o el Filtrado Basado en Contenidos (FBC). El artículo presenta algunos resultados mostrando que la integración de fuentes genera resultados superiores integrando solo las fuentes seleccionadas basadas en sus características intrínsecas. Resultados experimentales en el dominio de ventas de bienes de gran consumo (Supermercados) fueron obtenidos.

Keywords. Sistemas Recomendadores, Integración de la Información

1. Introducción

La sobrecarga de información es uno de los problemas más importantes que encuentran hoy en día los usuarios de Internet. La gran cantidad de vieja y nueva información para analizar, contradicciones en la información disponible genera ruido y hace difícil el proceso de identificación de información relevante.

Este fenómeno es determinado por la falta de métodos para comparar y procesar la información disponible. Los sistemas recomendadores hacen frente a este problema filtrando la información más relevante para las necesidades de los usuarios. Estos sistemas reciben como entradas las preferencias de los usuarios, las analizan y realizan recomendaciones. En cualquier caso, la búsqueda de información específica para estos sistemas es una tarea difícil. Los trabajos realizados en el campo de los sistemas recomendadores están enfocados hacia los métodos que se usan para filtrar la información para hacer las recomendaciones. Métodos como Filtrado Basado en Contenidos (FBC)[4] [5] y Filtrado Colaborativo (FC) [8][9] son ejemplos significantes en este campo.

Este artículo presenta los resultados de aplicar ambos métodos el Filtrado Colaborativo (CF)y el Filtrado Basado en Contenidos (FBC) con información de las fuentes seleccionadas basdas en sus características intrínsecas. Nosotros usamos las fuentes seleccionadas por la metodología ISIIRES (Identifying, Selecting, and Integrating Information for Recommender Systems) [2]. ISIIRES identifica fuentes de información basadas

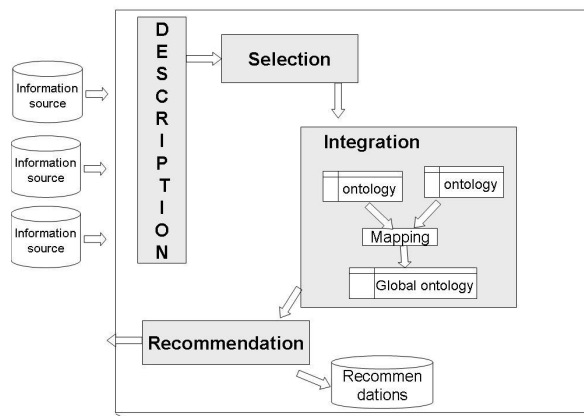


Figure 1. Functional Blocks of the Methodology ISIRES

en un conjunto de características intrínsecas y selecciona las más relevantes para utilizarlas en un sistema recomendador.

Este artículo está organizado como sigue: En la sección 2 es descrito ISIRES como punto de partida para implementar los métodos de recomendación FC y FBC. En la sección 3, es detallada la aplicación de ambos métodos. En la sección 4 es descrito un caso de estudio y los resultados obtenidos y finalmente, las conclusiones son presentadas en la sección 5.

2. Metodología: ISIRES

Una metodología para la identificación de fuentes de información, la comparación y la selección de las más relevantes para realizar las recomendaciones ha sido propuesta por Aciar et.al. en [2]. La metodología está compuesta por cuatro bloques los cuales son mostrados en la figura 1.

1. Descripción:

Un conjunto de características intrínsecas de las fuentes ha sido definido para la identificación de fuentes de información relevantes para realizar las recomendaciones [3]. Las características que describen la información contenida en las fuentes para usar en un sistema recomendador son: Completitud de la información contenida en la fuente, Diversidad de grupos de usuarios, Actualización de la información, Frecuencia de las transacciones realizadas por los usuarios y la cantidad de información de los usuarios disponible en la fuente. Estas características representan:

- Una representación abstracta de la información contenida en las fuentes.
- Criterios para comparar y seleccionar las fuentes.

2. Selección:

La selección de las fuentes se realiza en base a las características intrínsecas y un valor de trust de las recomendaciones pasadas que se realizaron [1].

. Integración:

Un mapping entre ontologías es realizado para la integración de la información de las fuentes seleccionadas las cuales son distribuidas y heterogéneas [2].

4. Recomendación:

La metodología está enfocada en la selección de las fuentes. Cuando las fuentes relevantes han sido seleccionadas es posible aplicar cualquier método de recomendación. En este artículo la metodología se aplica utilizando dos métodos de recomendación: FC y FBC.

Las siguientes secciones muestran los resultados obtenidos en la aplicación de ambos métodos con información de las fuentes seleccionadas usando la metodología ISIRES versus la aplicación de los métodos usando la información de todas las fuentes.

3. Aplicando diferentes métodos para recomendar

Dos de los principales métodos para recomendar son utilizados para computar las recomendaciones: El FC y el FBC, estos métodos son implementados en este artículo con la información de las fuentes seleccionadas usando la metodología ISIRES [2].

3.1. Filtrado Basado en Contenido (FBC) :

En este método los atributos de los productos son extraídos y ellos son comparados con el perfil del usuario (preferencias y gustos). Vectores son usados para representar los perfiles de los usuarios y productos en este trabajo.

El vector usuario es:

$$U = \langle u_1, u_2, \dots, u_n \rangle$$

Los valores del vector usuario representan las preferencias que el tiene por los atributos de los productos. Donde u_i representa el peso del atributo i con respecto a los usuarios. Este peso es obtenido de las compras previas que el usuario ha hecho.

El vector producto es:

$$P = \langle p_1, p_2, \dots, p_n \rangle$$

Donde p_i representa los pesos de los atributos i con respecto al producto. Este peso es asignado por un experto del dominio. La función del coseno que se basa en el espacio vectorial propuesta por Salton [7] usando los dos vectores P y U , es utilizada para establecer la relevancia que tienen los productos para los usuarios.

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

Los productos que tienen más alto valor de relevancia son recomendados a los usuarios.

3.2. Filtrado Colaborativo (FC):

La información provista por los usuarios con intereses similares es usada para determinar la relevancia que los productos tienen para los usuarios. La similitud entre los usuarios es calculada para este propósito y las recomendaciones son realizadas solamente en base a esta similitud. Para computar la distancia entre los usuarios es también usado el vector coseno [7]. Los usuarios son representados por vectores.

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

Donde U y V son vectores representando a los usuarios.

3.3. Evaluando las Recomendaciones

Las compras realizadas por los usuarios después de las recomendaciones son usadas para evaluar la precisión del recomendador. La precisión es evaluada usando la ecuación de precisión utilizada en el área de recuperación de información [6]. La precisión representa la probabilidad de que una recomendación sea exitosa.

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

Donde RP es el número de productos recomendados que han sido comprados y R es el número total de productos recomendados.

4. Caso de Estudio

Para realizar el caso de estudio fueron utilizadas ocho tablas relacionadas de una base de datos en el dominio de venta de bienes de gran consumo (supermercado). La base de datos contiene información de los productos, 1200 clientes y las compras realizadas en el periodo 2001-2002. Estas tablas son utilizadas como fuentes de información. Las fuentes utilizadas en los experimentos son las fuentes utilizadas en [1].

4.1. Filtrado Basado en Contenido (FBC)

La figura 2 muestra los atributos relevantes de los productos los cuales fueron definidos por un experto del supermercado. Las preferencias de los usuarios han sido establecidas en base a aquellos atributos representados por un vector:

$$U = \langle u_1, u_2, \dots, u_i \rangle$$

El peso u_i ha sido obtenido usando el método tf-idf (Term Frequency Times Inverse Document Frequency) [6], cuya calculación ha sido realizada a partir de las compras previas de los usuarios.

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

	codi	nóm
+	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, conservaciór
+	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 2. Atributos relevantes en el dominio de ventas de gran consumo (supermercado) definido por un experto en el dominio

cliente	producto	Relevance
91850	32318	0,3633179926
91850	34971	0,3633179926
91850	34864	0,3166948446
91850	34781	0,6488153614
91850	34599	0,3989170281
91850	34533	0,3166948446
91850	32998	0,3166948446
91850	32993	0,3166948446
91850	32865	0,3989170281
91850	32492	0,3633179926
91850	32491	0,3633179926
91850	32490	0,3633179926
91850	32378	0,3633179926
91850	33535	0,3989170281
921330	44683	0,4541988345
921330	48587	0,8585791209
921330	48497	0,5599386555
921330	48472	0,8585791209
921330	48194	0,4541988345
921330	47748	0,4541988345
921330	47343	0,5599386555
921330	46127	0,5599386555
921330	46087	0,5599386555
921330	45976	0,8585791209
921330	43890	0,8585791209
921330	45479	0,4541988345
921330	44452	0,4541988345
921330	49326	0,5599386555
921330	45606	0,5599386555
921330	55178	0,8585791209
921330	65745	0,4541988345
▶ 921330	63473	0,4541988345
921330	60339	0,4541988345
921330	59930	0,8585791209

Figure 3. Productos relevantes para los usuarios

t_i es la frecuencia del atributo i en las compras, n_i es el número de usuarios quienes han comprado un producto con el atributo i y N es el número total de usuarios.

El vector producto está compuesto por el peso p_i el cual ha sido asignado por el experto en el supermercado.

$$P = \langle p_1, p_2, \dots, p_i \rangle$$

La relevancia de cada producto para los usuarios ha sido establecida usando la ecuación 1 usando los vectores representando a los usuarios y los productos (vea figura 3).

VectorClienteU1_cliente	VectorClienteU2_cliente	SIM
1000285	299840	0,7426971197
1000285	2997940	0,7426971197
1000285	2998776	0,7426971197
1000285	2998773	0,7426971197
1000285	2998786	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 4. Similitud entre los usuarios - FC

A los usuarios se les recomiendan los productos que tienen un valor de relevancia > 6 .

4.2. Filtrado Colaborativo (FC)

Los mismos atributos que se muestran en la figura 2 fueron utilizados en el FC para obtener una representación de las preferencias de los usuarios. Los vectores para cada usuario son:

$$U = \langle u_1, u_2, \dots, u_n \rangle$$

El peso u_i ha sido obtenido de las compras previas de los usuarios usando el tf-idf (Term Frequency Times Inverse Document Frequency) [6] como lo fue utilizado en el FBC.

$$u_i = t_i * \log_2\left(\frac{N}{n_i}\right) \quad (5)$$

Donde t_i es la frecuencia del atributo i en las compras, n_i es el número de usuarios quienes han comprado un producto con el atributo i y N es el número total de usuarios.

La similitud entre los usuarios ha sido establecida aplicando la ecuación 2 a los vectores representando a los usuarios (vea figura 4)

Son recomendados Los productos comprados por otros usuarios con un valor de similitud > 6 .

4.3. Evaluando las Recomendaciones

Los experimentos han realizados implementando ambos métodos: FBC y el FC con la información de todas las fuentes con la metodología. La precisión del recomendador ha sido evaluada utilizando la ecuación 3 obteniendo los resultados mostrados en la figura 5 y la figura 6.

En las figuras 7 y 8 puede observarse la precisión del recomendador aplicando el FBC y el FC solamente a las fuentes de información seleccionadas en la metodología.

En los gráficos son mostrados como incrementa la precisión de las recomendaciones realizadas con estos métodos y las fuentes relevantes. En la figura 5 y figura 6 es mostrada

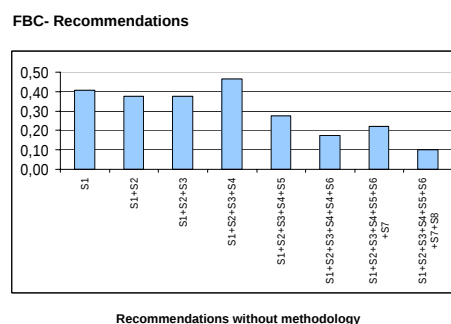


Figure 5. Precisión del recomendador utilizando FBC con todas las fuentes de información

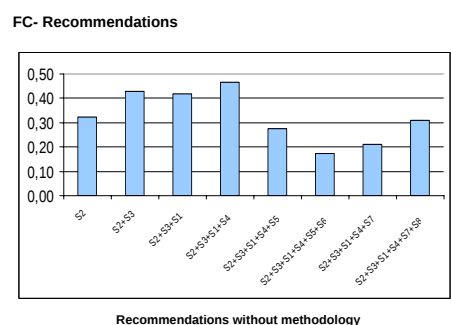


Figure 6. Precisión del recomendador utilizando el FC con todas las fuentes de información

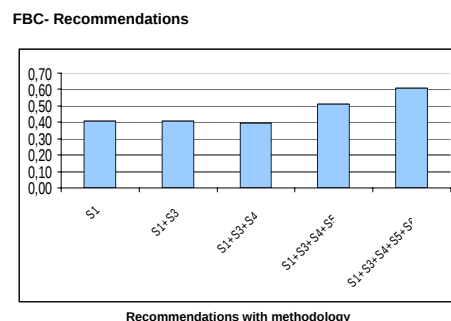


Figure 7. Precisión del recomendador aplicando el FBC solamente a las fuentes de información relevantes

la precisión de las recomendaciones usando todas las fuentes y puede observarse que la agregación de la información en las recomendaciones causa que la precisión sea más baja.

En la figura 7 y 8 es mostrada la precisión de las recomendaciones utilizando solo la información de las fuentes de información seleccionadas por ISIRES. Los resultados muestran que la agregación de la información de estas fuentes incrementa la precisión. La selección de las fuentes se basa en las características intrínsecas de cada una de ellas.

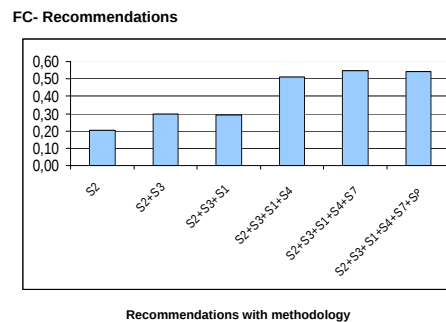


Figure 8. Precisión del recomendador aplicando el FC solamente a las fuentes de información relevantes

5. Conclusiones

La gran cantidad de información disponible en Internet hace que el proceso para inferir las preferencias de los usuarios y la selección de productos para ser recomendados sea más difícil. En este artículo ha sido utilizada una metodología que hace esta tarea más fácil y optimiza la búsqueda de información de usuarios resultando en mejores recomendaciones. En esta metodología ha sido definido un conjunto de característica intrínsecas de las fuentes. Las características representan información relevante contenida en las fuentes usadas para realizar las recomendaciones. La selección de las fuentes relevantes es realizada en base a ellas y las preferencias de los usuarios son establecidas usando la información solo de las fuentes seleccionadas. La metodología ha sido utilizada aplicando dos métodos de recomendación: el Filtrado Colaborativo (FC) y el Filtrado Basado en Contenido (FBC) obteniendo buenos resultados en cada uno de ellos. Los resultados obtenidos en un caso de estudio real muestran cómo la precisión de las recomendaciones realizadas con la integración de la información de las fuentes genera resultados superiores integrando solo la información de las fuentes seleccionadas en base a sus características intrínsecas.

References

- [1] S. Aciar, J López Herrera, and J. Ll. 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, Espana*, (Noviembre 2005).
- [2] S. Aciar, J López Herrera, and J. Ll. de la Rosa, 'Integrating information sources for recommender systems', *Artificial Intelligence Research and Development - IOS Press. VIII Catalan Association for Artificial Intelligence Congress (CCIA2005),Alghero, Sardinia, Italy*, (October 2005).
- [3] S. Aciar, J López Herrera, and J. Ll. de la Rosa, 'Sma para la búsqueda inteligente de información para recomendar', *I Congreso Español de Informática,(Cedi 2005), Simposio de Inteligencia Computacional, SICOŠ2005 (IEEE Computational Intelligence Society, SC)Granada, Espana*, (Septiembre 2005).
- [4] M. Balabanovic and Y. Shoham, 'Fab: Content-based, collaborative recommendation', *Communications of the ACM*, **40**(3), 66–72, (March 1997).

- [5] H. Lieberman, 'Letizia : An agent that assists web browsing', *Proceedings of the International Joint Conference on Artificial Intelligence, Montreal*, (1995).
- [6] G. Salton and C. Buckley, 'Automatic text processing', *he Transformation, Analysis and Retrieval of Information by Computer*, (1989).
- [7] G. Salton and M.J. McGill, 'Introduction to modern information retrieval', *McGraw-Hill Publishing Company, NewYork*, (1983).
- [8] U. Shardanand and P. Maes, 'Social information filtering: Algorithms for automating word of mouth', *SIGCHI Conference*, (1995).
- [9] B. Smyth and P. Cotter, 'A personalized television listings service', *Communications of the ACM*, (2000).