

**SISTEMA RECOMENDADOR BASADO EN FILTRADO
COLABORATIVO Y MINERIA DE DATOS**

GERMAN PAEZ BELTRAN

**FUNDACION UNIVERSITARIA SAN MARTIN
FACULTAD DE INGENIERIA
PROGRAMA DE INGENIERIA SISTEMAS
BOGOTA, D.C.
2010**

T.i

529

R#2410

**SISTEMA RECOMENDADOR BASADO EN FILTRADO COLABORATIVO Y
MINERÍA DE DATOS**

GERMÁN PÁEZ BELTRÁN

**FUNDACIÓN UNIVERSITARIA SAN MARTÍN
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA SISTEMAS
BOGOTÁ
2010 I**

**SISTEMA RECOMENDADOR BASADO EN FILTRADO COLABORATIVO Y
MINERIA DE DATOS**

**DIANA MILENA GERMAN PAEZ BELTRAN
COD. 042036
germanp28@hotmail.com**

MONOGRAFÍA DE GRADO

**ASESOR TÉCNICO
FABIAN ANDRES GIRALDO GIRALDO
CONSTRUCCION DE SOFTWARE**

**FUNDACIÓN UNIVERSITARIA SAN MARTÍN
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA SISTEMAS
BOGOTÁ
2010 I**

Nota de aceptación

Fabián Andrés Giraldo Giraldo
Asesor

Andrés Solarte
Jurado 1

Christian Benavides
Jurado 2

Bogotá, 5 de Junio de 2010

A mi Familia y Amigos,
Que siempre llenan de felicidad mi entorno
Y me inspiran valor para seguir adelante
En la carrera de la vida sin temor a nuevos retos.

AGRADECIMIENTOS

A Dios, por brindarnos la dicha de la salud y bienestar físico y espiritual

A la señora Bertha Lilia y sus hijos quienes me brindaron un gran apoyo, tanto económico como moral y emocional para lograr sacar adelante mi título profesional

A mis padres y hermanos, como agradecimiento a su esfuerzo, amor y apoyo incondicional, durante la formación tanto personal como profesional.

A mi Asesor que siempre estuvo disponible con sus valiosos consejos para que este trabajo se llevara a cabo.

A mis docentes, por brindarnos su guía y sabiduría en el desarrollo de mi formación profesional

A mis compañeros de universidad que siguieron el paso de graduación conmigo y fueron de gran ayuda en el crecimiento profesional

CONTENIDO

	pág.
SISTEMA RECOMENDADOR BASADO EN FILTRADO COLABORATIVO Y MINERIA DE DATOS	1
SISTEMA RECOMENDADOR BASADO EN FILTRADO COLABORATIVO Y MINERIA DE DATOS	2
AGRADECIMIENTOS	5
CONTENIDO	6
LISTA DE TABLAS	9
LISTA DE FIGURAS	10
RESUMEN	12
INTRODUCCIÓN	13
PROBLEMA	15
2. JUSTIFICACIÓN	17
3. OBJETIVOS	19
3.1 OBJETIVO GENERAL	19
3.2 OBJETIVOS ESPECÍFICOS	19
4. MARCO REFERENCIAL	20
4.1 ANTECEDENTES	20
4.1.1 Algoritmos evolutivos para descubrimiento de reglas de predicción en la mejora de sistemas educativos adaptativos basados en web (Romero, 2005)	20
4.1.2 E-Commerce Recommendation Applications (Schfer, 2005)	22

4.1.3 Hybrid Recommender Systems for Electronic Commerce (Cohen, 2000)	23
4.1.4 Sistema de Recomendación utilizando técnicas de filtrado colaborativo (Nieto, 2006)	25
4.2 MARCO CONCEPTUAL	26
4.2.1 Comercio Electrónico (Brizzio, 1997)	26
4.2.2 E-commerce (Dans, 2008)	26
4.2.3 Oscommerce (Leon, 2000 -2008)	27
4.3 MARCO TEÓRICO	31
4.3.1 Minería de Datos (Vallejos, Romero, 2006)	31
Reglas de secuenciación y Asociación (Escarcega, 2007)	31
4.3.2 Etl (Tallent, 2009)	33
4.3.3 Scriptella (Scriptella, 2008)	33
4.3.4 Weka (Sun,2008)	36
4.3.5 Algoritmos genéticos (Marczyk, 2004)	38
4.5 LIMITACIONES Y ALCANCES	43
5. DISEÑO METODOLÓGICO	45
5.1 Openup (openup, 2007)	45
5.2 DISEÑO	50
5.2.1 Iteración 1. Documentación	50
5.2.2 FASE DE CONSTRUCCION	51

5.2.3 FASE DE TRANSICION	54
6. DESARROLLO	55
7. PRUEBAS Y RESULTADOS	113
8. CONCLUSIONES	114
GLOSARIO	116
BIBLIOGRAFÍA	117

LISTA DE TABLAS

	pág.
Tabla 1. Cruce de padres algoritmos genéticos	39
Tabla 2. Recursos Amazon	41
Tabla 3. Micro-Incremento de la Iteración 1	50
Tabla 4. Iteración 1. Adaptación tienda Oscommerce	51
Tabla 5. Iteración 2. Bajar información Tienda	51
Tabla 6. Iteración 3. Recargar Tienda	52
Tabla 7. Iteración 4. Generar Algoritmo Genético	52
Tabla 8. Iteración 5. Algoritmos de secuenciación	53
Tabla 9. Iteración 6. Algoritmos de asociación	53
Tabla 10. Iteración 7. Generar Recomendación	54
Tabla 11. Fase de Transición	54
Tabla 12. Resultados pruebas algoritmos de recomendación	90
Tabla 13. Listado se secuencias algoritmo GSP	97
Tabla 14. Productos asociación	104

LISTA DE FIGURAS

	pág.
Figura 1. Metodología propuesta para la mejora de cursos web	21
Figura 2. Expresión	24
Figura 3. Modelo entidad relación Oscommerce	29
Figura 4. Diagrama de arquitectura de Scriptella	33
Figura 5. Ilustración copia de datos entre dos bases de datos utilizando Scriptella	34
Figura 6. Script xml	35
Figura 7. Capas OpenUp	46
Figura 8. Una iteración pasa por un ciclo de vida	48
Figura 9. Reducción riesgo (curva roja) y creación de valor (curva verde) durante el ciclo de vida del proyecto	50
Figura 10. Registro de usuario nuevo en la tienda Oscommerce	62
Figura 11. Panel administrativo oscommerce plugin Supertracker	62
Figura 12. Vista productos tienda oscommerce.	65
Figura 13. Consulta tabla oscommerce.ratings	65
Figura 14. Código para lectura de recomendaciones a cada usuario	66
Figura 15. Resultados de recomendaciones de acuerdo a un usuario en particular	67
Figura 16. Modelo entidad relación base de datos de producción	69
Figura 17. Archivo de configuración etl.properties	72
Figura 18. Definición de etiquetas <connection>	72
Figura 19. Bloque de código Etl scriptella	74
Figura 20. sintaxis java para la ejecución de un archivo ETL de Scriptella	75
Figura 21. Diagrama de clases iteración 2	76
Figura 22. Diagrama de paquetes iteración 2	77
Figura 23. Script ETL carga de información	80
Figura 24. Tabla recomendación DB oscommerce	81
Figura 25. Diagrama de estructura de agentes	81
Figura 26. Diagrama de Clases iteración 3 Agentes	83

Figura 27. Formato de vector de puntuaciones para el usuario	85
Figura 28. Modelo de población algoritmo genético	85
Figura 29. Ecuación de distancia entre dos vectores	86
Figura 30. Diagrama de flujo algoritmo genético	87
Figura 31. Método de cruce de cromosomas	88
Figura 32. Cruce de cromosomas método genérico	89
Figura 33. Mutación de los cromosomas	89
Figura 34. Modelo Entidad – Relación Iteración 4	91
Figura 35. Diagrama de clases iteración 4	92
Figura 36. Sentencia sql	94
Figura 37. Formato arff	94
Figura 38. Reconocimiento de variables herramienta WEKA	96
Figura 39. Ejecución reglas de secuenciación WEKA	96
Figura 40. Resultado algoritmo de secuenciación	98
Figura 41. Agrupación de probabilidad de compra de productos	99
Figura 42. Sentencia sql	100
Figura 43. Formato arff	101
Figura 44. Reconocimiento de variables herramienta WEKA	102
Figura 45. Ejecución Apriori Weka	103
Figura 46. Calculo de support algoritmo Apriori	104
Figura 47. Calculo de confianza algoritmo Apriori	105
Figura 48. Resultados algoritmo de asociación	105
Figura 49. Grafica resultados Asociación	106
Figura 50. Diagrama de clases iteración 5	107
Figura 51. Diagrama de paquete iteración 6	108
Figura 52. Presentación final productos recomendados	109
Figura 53. Interface grafica aplicación	110
Figura 54. Paso de mensajes entre agentes	110

RESUMEN

El proyecto de grado "Sistema Recomendador basado en filtrado colaborativo y minería de datos" es una aplicación que tienen por objetivo el descubrimiento, análisis y recomendación de información relevante, tomando en consideración el comportamiento y preferencias de los usuarios de una tienda de comercio electrónico.

La aplicación está basada en la utilización de una herramienta ETL llamada Scriptella con la cual se logra la obtención de información desde la tienda de comercio electrónico, la transformación de datos y el posterior almacenaje en un DataMart, con el objeto de aplicar un algoritmo genético para el proceso de búsqueda de artículos recomendados para cada usuario de la tienda de comercio electrónico, además de la exploración de datos con la utilización de la herramienta WEKA y sus clases de asociación y secuenciación, que permiten el descubrimiento de vínculos entre productos y frecuencias de compra.

INTRODUCCIÓN

El comercio electrónico y los mecanismos de filtrado colaborativo y tratamiento de datos para el aprovechamiento de la información y la obtención de ventajas sobre los competidores, está enmarcado en varios escenarios como lo son la educación, la investigación y el comercio electrónico, que es la razón central del desarrollo de este trabajo.

A lo largo de la historia del comercio electrónico y los sistemas informáticos se ha logrado un avance en la investigación de métodos y estrategias que lo ayuden a comprender el comportamiento del usuario en tiendas de comercio electrónico. Al principio se basaban en simples sistemas estadísticos para mantener la información del flujo de venta de productos y los mejores usuarios. Al pasar el tiempo y con el crecimiento de competencia se hizo necesaria la adopción de nuevos métodos y técnicas para persuadir a los usuarios de tiendas electrónicas para la compra de productos y la extracción de información implícita de gustos e inclinaciones sobre los artículos o productos que se presentan en la tienda de comercio electrónico.

Es por esto que en la lucha de mantenerse a la vanguardia es necesario incorporar mecanismos y herramientas que faciliten a las tiendas de comercio electrónico, a obtener el máximo aprovechamiento de los datos contenidos en la tienda, valiéndose de técnicas avanzadas de optimización y exploración de datos como lo son el filtrado colaborativo basado en algoritmos genéticos y exploración de datos basados en minería de datos. Es de esta oportunidad que surge la necesidad de crear un sistema capaz de explorar y analizar la información de los usuarios para la obtención de nuevos datos relevantes a cada usuario registrado en la tienda de comercio electrónico.

Partiendo de este escenario se creó un sistema multiagente que permita encontrar productos recomendados para cada usuario, valiéndose de técnicas de filtrado colaborativo y minería de datos, utilizando la información dejada por cada usuario de las calificaciones otorgadas a cada producto y la información contenida en el carrito de compra.

El presente trabajo se divide en cuatro etapas que ayudan a completar el objetivo general propuesto, donde la primera etapa comienza con la extracción de información de la tienda de comercio electrónico con la ayuda de una herramienta ETL llamada Scriptella para la obtención de la información, aquí se realizan tratamientos sobre los datos para su posterior almacenaje en una bodega de datos. Partiendo de esta información se inicia la segunda etapa en donde se

procede a la construcción de un algoritmo genético capaz de generar recomendaciones para cada usuario de la tienda, apoyado en la información extraída de las votaciones de productos por cada cliente.

Para la tercera etapa se utiliza la información del carrito de compra de los usuarios con el fin de reglas de asociación de productos y realizar un intercambio de recomendaciones de acuerdo a sus preferencias, esto apoyado en la exploración de información en un DataMart para cada usuario.

En la cuarta etapa se genera un proceso de análisis de información obteniendo información de la bodega de datos, con el fin de extraer las secuencias de compra de productos para poder generar reglas de secuenciación para determinar los productos que un usuario puede llegar a comprar en próximas visitas. Se realiza una modificación sobre la tienda de comercio electrónico con el objetivo de presentar a cada usuario los recomendados, cada vez que este visite un producto en la tienda.

Finalmente se busca, con el desarrollo de este trabajo, que los administradores de tiendas de comercio electrónico se acerquen al máximo al aprovechamiento de la información almacenada, para presentar un beneficio positivo en el movimiento de productos en la tienda, logrando así obtener mayores ganancias sobre los productos vendidos en ella.