RELACIONANDO COMENTARIOS TEXTUALES Y VALORES NUMÉRICOS EN ENCUESTA DE SATISFACCIÓN DE USUARIOS

Mag. Raúl Klenzi, Lic. L. Gutierrez, Alum. Tamara Pinto
Instituto de Informática (IdeI) / Departamento Informática (DI) / Facultad de Ciencias Exactas Físicas y Naturales (FCEFN) / Universidad Nacional de San Juan (UNSJ)
Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", San Juan
rauloscarklenzi, gutierrez.laura, tamara932 @gmail.com

Resumen

En este trabajo se procesa información extraída de encuesta de satisfacción de usuarios alumnos de bibliotecas universitarias. En este contexto la encuesta consiste en valorar 17 atributos según una escala Likert de 5 estados asociados a conceptos que van desde insatisfecho (1) a totalmente satisfecho (5) respectivamente. Además, y como parte de la encuesta, cada encuestado puede expresar en formato de texto comentarios adicionales. El objetivo central de este trabajo consiste en relacionar los atributos inductores que describen la encuesta desde las respuestas numéricas, con aquellos que se obtienen desde un análisis de Text Mining (TM) aplicado a los comentarios. La encuesta procesada, como caso de estudio, se realizó a alumnos de la biblioteca de la Facultad de Ingeniería de la Universidad Nacional de San Juan (FI-UNSJ)

Palabras clave: Data Mining, Text Mining, Satisfacción de Usuarios.

Contexto

En el marco de proyectos anteriores y del actual “Minería de datos (DM) en la Determinación de Patrones de Uso y Perfiles de usuarios” 21/E889 se realizaron encuestas a los usuarios de Bibliotecas tratando de encontrar su grado de satisfacción respecto de los servicios y funcionamiento de las mismas.

El constructo que permitió realizar la encuesta, se conforma de 17 atributos, asociados a diferentes servicios que ofrece la biblioteca, que deben ser respondidos, con valoraciones numéricas entre 1 y 5 según sus percepciones. A la vez se permite, a los usuarios, escribir todo aquello que considere válido a analizar por parte de los encuestadores en formato de texto libre a modo de comentario.

Como tareas inherentes al proyecto se han procesado las encuestas encontrando los inductores (atributos más relevantes) que conforman la imagen que los usuarios tienen de su biblioteca. La instancia a considerar en la presente propuesta, es procesar los comentarios que se redactaron por parte de los usuarios mediante técnicas de TM.

El propósito de este procesamiento será contrastar, para aquellas encuestas que contienen comentarios, la coincidencia o complementación del conocimiento extraído de la valoración numérica realizada por los usuarios a los diferentes atributos.

Introducción

Muchas encuestas en las que los atributos por los que se consulta deben ser respondidos en forma discreta y numérica, permiten en un apartado de la misma, que el encuestado brinde una opinión en formato texto que amplíe o aclare lo expresado en sus respuestas numéricas.

Este es el caso del constructo validado en **[1]**, el cual se relevó a usuarios alumnos de la biblioteca de la FI-UNSJ. Estos atributos cubren entre otros, aspectos edilicios, de personal, de material bibliográfico y tecnológico.

La hipótesis, que mueve la presente propuesta, es que todo comentario expresado en formato de texto en principio se considera una "queja" o "reclamo" sobre alguna condición de la biblioteca que, a criterio del encuestado, puede mejorarse y que por lo tanto deberá estar relacionada con aquellos atributos menos valorados numéricamente por el mismo.

Mediante tareas de segmentación aplicadas a las respuestas numéricas de las encuestas se dividen las respuestas en dos grupos. Por un lado el grupo de los encuestados satisfechos con las prestaciones de la biblioteca y por otro lado el grupo de aquellos usuarios que consideran que la biblioteca puede mejorar sus servicios. Tras esta agrupación y desde una tarea de clasificación se obtienen los atributos inductores que describen la encuesta detectando además, cuáles de éstos definen la pertenencia a uno u otro segmento.

Por otro lado, mediante tareas de TM y métricas de similitud sintáctica entre el nombre de los atributos y los comentarios expresados en cada encuesta, se encontrará información complementaria sobre los atributos evaluados.

Con el objetivo de verificar la hipótesis manifestada anteriormente, es de esperar que aquellos atributos inductores obtenidos desde el procesamiento numérico y que definen al segmento de "posibles mejoras" coincidan con los de mayor similitud sintáctica, encontrados desde el procesamiento de los comentarios.

Líneas de investigación y desarrollo

El proyecto en el que está enmarcado el presente trabajo, ha llevado adelante durante el último año, diferentes líneas de investigación con el afán de que la investigación aplicada ayude a la toma de decisiones de la autoridad competente.

Así, desde la aplicación de diferentes técnicas de DM y TM se están procesando títulos bibliográficos de la biblioteca midiendo la similitud sintáctica de los mismos con los contenidos de las diferentes carreras que se dictan en la FCEFN-UNSJ tratando de proponer una primera aproximación a la determinación de código Dewey para bibliografía que no lo posee o es edición, en español, de la propia universidad.

El uso de métricas de TM es también un camino seguido en la presente propuesta de trabajo y es un eslabón que pretende ir cerrando la cadena de aplicaciones que, en el tema de satisfacción de usuarios de bibliotecas universitarias y desde el procesamiento de la encuesta a usuarios alumnos y docentes de la biblioteca, encaró el grupo de investigadores pertenecientes al proyecto.

En la mayoría de los casos los trabajos se han llevado adelante mediante la utilización de herramientas de software libre del área de DM. Los resultados se han alcanzado mediante el uso de algoritmos de DM que posee la herramienta RapidMiner cuya última versión es la 5.3.005 **[5]**.

Desarrollo

La aplicación consta de tres pasos:

1) El procesamiento de datos correspondientes a la fracción numérica de la encuesta, consiste en la aplicación sucesiva del algoritmo de segmentación W-SimpleKmeans a un grupo de 46

encuestas de alumnos, de las 150 relevadas, que poseen comentarios y que permite segmentar y etiquetar en dos grupos, las opiniones de los encuestados.
Seguidamente el algoritmo clasificador WJ48 permite describir la encuesta con sus etiquetas asociadas, con la menor cantidad de atributos, denominados inductores reconociendo además, mediante aplicaciones sucesivas del operador Single Rule Induction, cuáles de ellos definen cada segmento.

2) Esta parte del desarrollo consiste en encontrar una medida de similitud sintáctica entre el comentario de una encuesta denominada consulta o **req**uest y cada uno de los nombres asociados a los atributos de la encuesta denominado **ref**erencia, realizando además, un mínimo análisis de sinonimias.

3) El tercer y último paso consiste en analizar comparativamente los resultados obtenidos en l) y 2) tratando de verificar la hipótesis.

La elección de los algoritmos se fundamenta en la necesidad de procesar registros incompletos permitido por W-SimpleKmeans y W-J48 como así también la rápida interpretación de resultados que posee una estrategia de árboles basada en Ganancia de Información Relativa **[4]**.

La tarea de segmentación llevada adelante en 1) asignó 28 registros al cluster0 y 18 al cluster1. A los efectos de asignar un significado a cada cluster, la Tabla 1 presenta el valor de sus respectivos centroides. Allí se aprecia, según el valor de los centroides, que los usuarios tienen una buena percepción de los servicios ofrecidos por la biblioteca, dado que ambos valores están por encima de la media. De todas maneras se puede considerar al segmento con menor valor de centroide, como aquel en el que los encuestados consideran que la biblioteca es factible de mejorar en sus prestaciones.

Row No.	SEGMENTO	CLUSTER	average(CENTROIDE)
1	POSIBLES MEJORAS	cluster1	3.0893518518518515
2	SATISFACTORIO	cluster0	3.8625550220088027

Tabla 1

Tras la tarea de segmentación y desde la aplicación del algoritmo clasificador se obtuvo el siguiente árbol

W-J48

J48 unpruned tree

CANTIDADPERSONAL <= 2: POSIBLES MEJORAS (8.0)
CANTIDADPERSONAL > 2
| CANTIDADMATERIAL <= 2: POSIBLES MEJORAS (6.16/0.16)
| CANTIDADMATERIAL > 2
| | IDONPERSONAL <= 3
| | | SERVPRESTAMODOMICILIO <=3:SATISFACTORIO (2.0)
| | |SERVPRESTAMODOMICILIO >3:POSIBLESMEJORAS (4.0)
| | IDONPERSONAL > 3: SATISFACTORIO (25.84)

Number of Leaves : 5
Size of the tree : 9

Árbol de clasificación descriptor de la encuesta

En el clasificador anterior se observa que solamente 4 atributos, de los 17 iniciales, denominados inductores, describen la totalidad de la encuesta. Con estos cuatro atributos inductores como universo, se aplican sucesivamente algoritmos del tipo Single Rule Induction (Single Attribute), cuyo objetivo es describir la encuesta, con el menor error posible, mediante la consulta de un único atributo y paso seguido se elimina el atributo encontrado.
De los atributos inductores el que mejor describe la encuesta es Idoneidad de personal (**IDONPERSONAL**) que asigna correctamente 39 de los 46 registros. Eliminado **IDOPERSONAL**, de los restantes, es Cantidad de material (**CANTIDADMATERIAL**) el que describe mejor la encuesta con 37 registros asignados correctamente de los 46. La secuencia de pasos antes descriptos y hasta agotar los atributos inductores permite obtener la Tabla 2.

ATRIBUTO	SEGMENTO	
	SATISFAC-TORIO	POSIBLES MEJORAS
IDONPERSONAL 39/46 (84,78%)	26/31 (83,87%)	13/15 (86,66%)
CANTIDADMATERIAL 37/46 (80,43%)	28/37 (75,67%)	9/9 (100%)
CANTIDADPERSONAL 36/46 (78,26%)	28/38 (73,68%)	8/8 (100%)
SERVPRESTAMO-DOMICILIO 28/46 (60,86%)	28/46 (60,86%)	0

Tabla 2

En la Tabla 2 se observa que los atributos **CANTIDADMATERIAL** y **CANTIDADPERSONAL** describen con mayor exactitud el segmento "POSIBLES MEJORAS"

Para la realización de 2), en trabajos previos presentados en WICC 2012 **[2] [3]**, se constató que entre las diferentes métricas de similitud utilizables y que posee la herramienta de software disponible, la del coseno es la que mejores resultados brinda. En este caso los valores de similitud (distance) varían entre 0 (**req** y **ref** sin similitud) y 1 en que ambos son iguales. Así mismo y dado que el comentario escrito por el encuestado está en formato libre, es necesario generar una base de datos de sinónimos que permita buscar aproximaciones entre el **req**uest y la **ref**erencia. Si bien la herramienta RM posee un módulo de wordnet que facilita las tareas de procesamiento de sinónimos y significado de palabras, el mismo por el momento funciona correctamente para el idioma inglés. Por ello la tarea de procesamiento de sinónimos se embebió en el módulo (Process document from file Replace Token) de la propia herramienta como se muestra en la Figura 1.

En la Figura 1 se aprecia que palabras como *avisador* o *anunciador*, que pudieran aparecer en los comentarios, automática-mente se reemplazan, en tareas de preprocesamiento, por *cartelera* que es una palabra contenida en algún nombre de atributo de la encuesta.

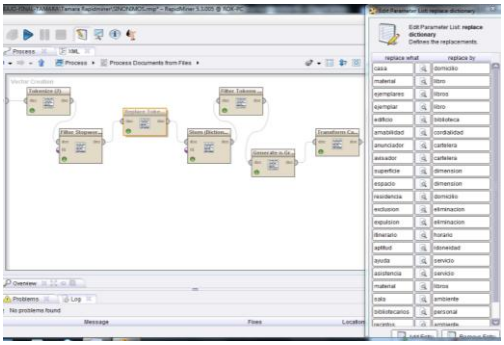

Figura 1

La aplicación de la secuencia de módulos que se presenta en Figura 2 permitió obtener las métricas, entre nombres de atributos y comentarios, como se observa en la Tabla 3.

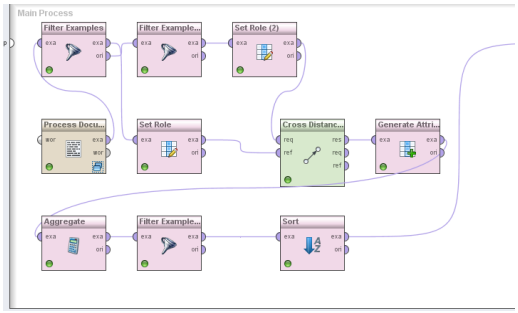

Figura 2

En la Tabla 3 se aprecia que para el comentario "Con respecto a los días sábado…" la mayor similitud (distance) se da con el atributo Horario de la Biblioteca y solamente 4 de los 17 atributos presentan un valor de similitud (distance) mayor que cero.

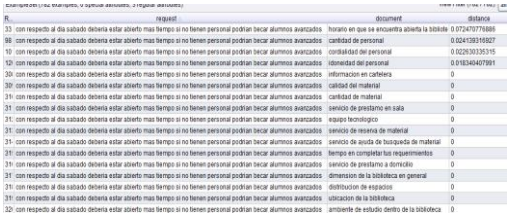

Tabla 3

Finalmente la aplicación desarrollada en RM presenta los resultados, que a modo de resumen figuran en la Tabla 4

document	ATRIBUTOQUEJA	count(CUENTA)
cantidad de material	SI	23
cantidad de personal	SI	23
servicio de ayuda de busqueda de material	SI	23
calidad del material	SI	22
horario en que se encuentra abierta la biblioteca	SI	22
servicio de reserva de material	SI	22
ambiente de estudio dentro de la biblioteca	SI	18
cordialidad del personal	SI	16
idoneidad del personal	SI	16
dimension de la biblioteca en general	SI	14
ubicacion de la biblioteca	SI	14
servicio de prestamo en sala	SI	6
servicio de prestamo a domicilio	SI	5
distribucion de espacios	SI	1
informacion en cartelera	SI	1

Tabla 4

La Tabla 4 muestra que la mitad de los comentarios (23) tienen una medida de similitud distinta de cero con atributos referidos a: cantidad de material, cantidad de personal y servicio de ayuda de búsqueda de material, respectivamente. Esto verifica la hipótesis que justamente los atributos inductores del segmento "posibles mejoras" coinciden con los encontrados desde las métricas de similitud entre nombre de atributo y comentario.

Resultados y Objetivos

La aplicación permitió comprobar la hipótesis de que los comentarios, mayormente, expresan disconformidades de los usuarios para con un servicio bajo evaluación.

En el marco de trabajos futuros y en el contexto de la aplicación, se pretende extender el análisis a encuestas relevadas en otras bibliotecas de la UNSJ y ampliarla utilizando el módulo de RM de wordnet adaptado a lenguaje español que acrecentaría la posibilidad de mejorar métricas y eliminaría el trabajo cuasi manual de reemplazos implementado, comenzando a transitar el camino desde el análisis sintáctico al análisis semántico.

Formación de Recursos Humanos

La formación de recursos humanos es un tema de vital importancia. En este marco, se están dirigiendo un conjunto de trabajos finales de grado de becas de finalización de carrera de la Agencia Nacional de Promoción Científica y Tecnológica.

En particular, se está trabajando con datos de títulos bibliográficos pertenecientes a la biblioteca de la FCEFN-UNSJ a los que en forma automática se intenta asignar la codificación Dewey correspondiente. Así mismo y desde una tesis de posgrado, se intenta encontrar automáticamente también, y mediante similitud sintáctica qué contenidos mínimos establecidos en la resolución 786/09 son contenidos brindados por los planes de estudios de las carreras de informática de la FCEFN-UNSJ

Al momento, integrantes del proyecto dirigen 7 trabajos de grado y 2 de posgrado en temáticas afines a las tratadas en el proyecto.

Referencias

[1] Beguerí, G. "Logística como garantía de satisfacción del usuario". Tesis de maestría- Universidad Nacional de Cuyo. (2007)

[2] Klenzi, R., Gutierrez, L., Villafañe V. "Técnicas de Recuperación de Información en la Determinación de Pertinencias Bibliográficas". WICC-2012.

[3] Liu B., "Web DataMining. Exploring Hyperlinks, Contents, and Usage Data" Springer-Verlag Berlin Heidelberg 2007

[4] North M. "Data Mining for the Masses" ISBN: 0615684378. A Global Text Project Book. Ed 2012.

[5] https://rapid-i.com/content/