Confluencia de Áreas de Conocimiento en un Laboratorio de Sistemas Inteligentes

Klenzi, Raúl; Araya, Jorge; López, Marcelo, Murazzo, María
Instituto de Informática / Departamento Informática / Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan
Domicilio: Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan, CPA: J5402DCS, 0264-260353 0264-4260355
{rauloscarklenzi; jorgemaraya; marcelo.sanjuan.ar; maritemurazzo}@gmail.com

Resumen

Recientemente en el ámbito del Instituto de Informática de la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de San Juan (IdeI-FCEFN-UNSJ) se ha conformado el "Laboratorio de Sistemas Inteligentes para la búsqueda de Conocimiento en Datos Masivos", y está integrado por docentes investigadores pertenecientes al Departamento de Informática de la citada Facultad (DI). Las áreas de conocimiento involucradas en el laboratorio son las derivadas de dos proyectos de investigación CICITCA-UNSJ "Extracción de Conocimiento en Datos Masivos" y "Cloud Computing con herramientas libres para la evaluación de modelos de despliegue híbrido".
Como epílogo de aquellos y prólogo de futuros proyectos en el marco del citado Laboratorio, se presenta el tratamiento de dos casos derivados de aquellas líneas de investigación. Un primer trabajo pretende automatizar el proceso de determinación de código Dewey mediante tareas de minería de texto en tanto un segundo caso propone detectar y modelar un ataque a un servidor de red por denegación de Servicios. En ambos casos las aplicaciones se han resuelto mediante la utilización de módulos y algoritmos específicos pertenecientes al entorno de software libre RapidMiner (RM) 5.3.015.

Palabras clave: Extracción de Conocimiento, TextMining, DNS, software libre

Contexto

En el ámbito del departamento Informática e Instituto de Informática DI-FCEFN-UNSJ, IdeI-FCEFN-UNSJ se han desarrollado en el pasado bienio dos proyectos de áreas complementarias. Por un lado el proyecto CICITCA 21/E951 "Extracción de Conocimiento en Datos Masivos" y por otro el proyecto "Cloud Computing con herramientas libres para la evaluación de modelos de despliegue híbrido" CICITCA 21/E973 y que han sido el corolario de una línea de investigación iniciada en el último lustro. Al momento y en el ámbito del IdeI, por ordenanza 02/2015-CD-FCEFN, se acaba de conformar el laboratorio de Sistemas Inteligentes para la extracción de Conocimiento en Datos Masivos en el que confluyen ambas líneas de investigación y al que conforman docentes e investigadores de aquellos proyectos. Los casos de aplicación que se presentan, hacen las veces del prólogo de este nuevo ámbito de trabajo.
Desde el proyecto "Extracción de Datos Masivos" se han desarrollados aplicaciones, en el área temática del

Descubrimiento de Conocimiento en Bases de Datos (KnowledgeDiscovery Data -KDD-) el cual es un análisis automático exploratorio y modelado de grandes depósitos de datos. KDD es el proceso organizado de determinación de patrones válidos, nuevos, útiles y comprensibles de grandes y complejos conjuntos de datos. Minería de datos (Data Mining -DM-) es el centro del proceso de KDD, e implica la inferencia de algoritmos que exploran los datos, para desarrollar modelos y descubrir patrones previamente desconocidos.Maimon, O., &Rokach, L. (Eds.). (2005).

Estas aplicaciones han permitido concluir diferentes trabajos finales de grado así como avances importantes en otros trabajos de grado y posgrado, los cuales han contemplado las áreas de Data Mining (DM), Text Mining (TM), Web Mining (WM) y Web Analitycs (WA) que medianteherramientas de software libre, en plataformas de hardware multicore, GPU computing, cluster de computadoras y sobre datos del área de la astronomía, redes sociales y relevamientos propios permitió extraer conocimiento en datos masivos.

Así mismo desde el proyecto "Cloud Computing con herramientas libres para la evaluación de modelos de despliegue híbrido" CICITCA 21/E973, entre otras actividades, se destacan la posibilidad constante de medir el flujo de bits que acceden a determinados servidores de nuestra facultad a la vez que se ha logrado poner en funcionamiento un cluster de computadoras, el cual se encuentra en una etapa de evaluación sobre el que se pretende extender las aplicaciones de KDD ya testeadas en plataformas monousuarios.
Desde la medición del flujo de bits que ingresan al servidor se propone una primera instancia de modelación de las tramas que caracterizan a un ataque por denegación de servicios, también llamado ataque DoS (en inglés Denial of Service) o DDoS (Distributed Denial of Service), el cual es un ataque a un sistema de computadoras o red que causa que un servicio o recurso sea inaccesible a los usuarios legítimos. Normalmente provoca la pérdida de la conectividad de la red por el consumo del ancho de banda de la mismao sobrecarga de los recursos computacionales del sistema de la víctima.(Mens, J. P. 2008).

Introducción

Caso de Aplicación 1) Este caso de aplicación es la continuación del trabajo presentado en el WICC 2014 "Minería de Texto en la Determinación Automática de Código Dewey (Una Primer Aproximación)"(Klenzi, R. O., & Araya, J. M.; 2014).
El objetivo de este trabajo consiste en reemplazar la tarea manual de la persona encargada de asignar la correspondiente codificación Dewey (Sistema de Clasificación Decimal creado por Melvil Dewey, un bibliotecario estadounidense, cuyo propósito inicial fue organizar la colección de la biblioteca del Amherst College.) a nuevas publicaciones bibliográfica, mediante un proceso automático basado esencialmente en similitudes sintácticas, derivadas del área de la minería de texto, entre ese nuevo material y material bibliográfico con su Dewey ya asignado.
En aquella presentación y desde una representación vectorial como valija de palabras de las fuentes de texto, se utilizó la métrica de similitud del coseno para contrastar parecidos sintácticos entre títulos e índices temático correspondiente a determinada bibliografía con las áreas

de conocimiento asociadas a los diferentes departamentos de la FCEFN así como con otros títulos e índices bibliográficos ya catalogados según su codificación Dewey y desde cuyo parecido sintáctico permitía proponer, al menos, una fracción del código correspondiente.
Para la implementación se utilizó la herramienta de software libre licencia AGPL RapidMiner 5.3.015 que es un entorno de prueba de algoritmos de aprendizaje de máquina que permite desde el modelo de datos, extraer conocimiento desde los mismos permitiendo la materialización de la totalidad de los pasos que involucran el KDD desde el preprocesamiento de datos hasta las visualización de los modelos (North, M.; 2012).
En esta oportunidad se ha agregado a aquel entorno, la medida de similitud de Okapi, presentando la fórmula directamente a través de una "bolsa de las palabras" pertenecientes a los documentos en lugar de su representación vectorial. Así el documento dj se denota por dj y consulta q se denota por q. anotaciones adicionales son los siguientes.

$$okapi(d_j,q)=\sum_{t_i\in q,d_j}\ln\frac{N-df_i+0.5}{df_i+0.5}\times\frac{(k_1+1)f_{ij}}{k_1(1-b+b\frac{dl_j}{avdl})+f_{ij}}\times\frac{(k_2+1)f_{iq}}{k_2+f_{iq}}$$

Figura 1: Expresión matemáticade la métrica de Okapi.

Dónde:
N: Representa el número de documentos en la colección.
dfi: Representa el número de documentos que contienen al termino i.
k1, b, k2: Son parámetros de ajuste, cuyos valores se ajustan según se obtengan los mejores resultados.
dlj: Es el tamaño del documento en palabras.
Avdl: Representa el tamaño promedio de los documentos en la colección.
fij: Es el peso del término i en el documento j.
fiq: Representa el peso del término ien la consulta q.

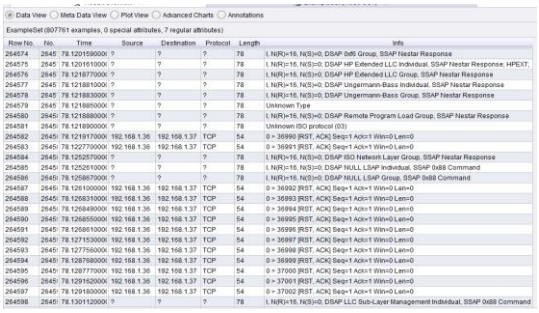
La fórmula de recuperación Okapi dada, se ha demostrado más eficaz que la del coseno para la recuperación de consultas cortas(Liu, B. 2007).

Al momento sólo se proponen codificaciones Dewey para material bibliográfico recientemente editado o adquirido por la biblioteca de la FCEFN afín al área de conocimiento asociada al DI-FCEFN. Por ello una primera parte de la aplicación contenida en el caso 1) y desde la implementación de la métrica de Okapi permite filtrar automáticamente, material bibliográfico ya catalogado, y con mayor similitud sintáctica con los contenidos mínimos de las asignaturas de las carreras pertenecientes al DI-FCEFN. Contra esta base, se compara el material bibliográfico de adquisición reciente rescatándose los cuatros libros de mayor similitud sintáctica y dejando a criterio del catalogador de la biblioteca la decisión final como se presenta en la Figura 2. Allí se observa, a efectos de constatar el funcionamiento del sistema, en la columna DEWEY_EXPERTO los valores asignados por el experto en catalogación de la Biblioteca de la FCEFN así como en la columna DEWEY_PROPUESTOS los valores obtenidos automáticamente. Si bien se observan diferencias, se está en el proceso de mantener nuevas reuniones con el experto a efectos de agregar al simple análisis sintáctico, reglas de análisis propias del catalogador que posibiliten una mejor aproximación a lo expresado por el experto.

Row No.	LIBRO_NUEVO	SIMILITUD	DEWEY_PROPUESTO	DEWEY_EXPERTO
1	COMUNICACION DE DATOS, REDES DE COMPUTADORAS Y SISTEMAS ABIERTOS.	0.1920	004.6	004.6
2	COMUNICACION DE DATOS, REDES DE COMPUTADORAS Y SISTEMAS ABIERTOS.	0.1859	004.67	004.6
3	COMUNICACION DE DATOS, REDES DE COMPUTADORAS Y SISTEMAS ABIERTOS.	0.1731	004.6	004.6
4	COMUNICACION DE DATOS, REDES DE COMPUTADORAS Y SISTEMAS ABIERTOS.	0.1575	004.68	004.6
5	GUIA DE APRENDIZAJE PHP.	0.1895	005.133C	004.62
6	GUIA DE APRENDIZAJE PHP.	0.1865	005.133D	004.62
7	GUIA DE APRENDIZAJE PHP.	0.1417	005.11	004.62
8	GUIA DE APRENDIZAJE PHP.	0.1395	005.73C	004.62
9	INTRODUCCION A LA INVESTIGACION DE OPERACIONES.	0.3659	003	003
10	INTRODUCCION A LA INVESTIGACION DE OPERACIONES.	0.3282	003	003
11	INTRODUCCION A LA INVESTIGACION DE OPERACIONES.	0.3142	003	003
12	INTRODUCCION A LA INVESTIGACION DE OPERACIONES.	0.2243	003	003
13	INVESTIGACION DE OPERACIONES.	0.1497	003	003
14	INVESTIGACION DE OPERACIONES.	0.1473	003	003
15	INVESTIGACION DE OPERACIONES.	0.1464	003	003
16	INVESTIGACION DE OPERACIONES.	0.0916	003	003
17	JAVA 2.	0.4016	005.133D	005.133J
18	JAVA 2.	0.3823	005.133C	005.133J
19	JAVA 2.	0.3562	005.133J	005.133J
20	JAVA 2.	0.3145	005.76	005.133J
21	REPARACION AVANZADA DE PC CON WINDOWS.	0.1338	004.6	004.20288
22	REPARACION AVANZADA DE PC CON WINDOWS.	0.1231	004.22	004.20288
23	REPARACION AVANZADA DE PC CON WINDOWS.	0.1168	004.21	004.20288
24	REPARACION AVANZADA DE PC CON WINDOWS.	0.1102	005.133D	004.20288

Figura 2: Propuestas de Códigos Deweys para nuevo material bibliográfico.

Caso de Aplicación 2) Este segundo caso de aplicación y desde datos representativos de una trama de bits de acceso al servidor se generó el formato csv que se aprecia en la Figura 3.

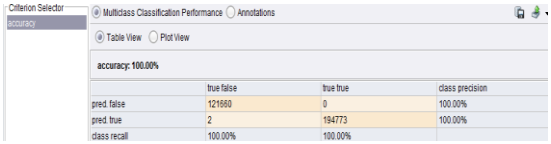


Row No.	No.	Time	Source	Destination	Protocol	Length	Info
264574	2645	78.120159000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP 0x66 Group, SSAP Nestar Response
264575	2645	78.120161000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP HP Extended LLC Individual, SSAP Nestar Response; HPEXT,
264576	2645	78.121877000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP HP Extended LLC Group, SSAP Nestar Response
264577	2645	78.121881000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP Ungermann-Bass Individual, SSAP Nestar Response
264578	2645	78.121883000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP Ungermann-Bass Group, SSAP Nestar Response
264579	2645	78.121885000(	?	?	?	78	Unknown Type
264580	2645	78.121888000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP Remote Program Load Group, SSAP Nestar Response
264581	2645	78.121890000(	?	?	?	78	Unknown ISO protocol (03)
264582	2645	78.121917000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36990 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264583	2645	78.122770000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36991 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264584	2645	78.125257000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP ISO Network Layer Group, SSAP Nestar Response
264585	2645	78.125261000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP NULL LSAP Individual, SSAP 0x88 Command
264586	2645	78.125867000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP NULL LSAP Group, SSAP 0x88 Command
264587	2645	78.126100000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36992 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264588	2645	78.126833000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36993 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264589	2645	78.126849000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36994 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264590	2645	78.126855000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36995 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264591	2645	78.126861000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36996 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264592	2645	78.127153000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36997 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264593	2645	78.127756000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36998 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264594	2645	78.128768000(	192.168.1.36	192.168.1.37	TCP	54	0 > 36999 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264595	2645	78.128777000(	192.168.1.36	192.168.1.37	TCP	54	0 > 37000 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264596	2645	78.129162000(	192.168.1.36	192.168.1.37	TCP	54	0 > 37001 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264597	2645	78.129180000(	192.168.1.36	192.168.1.37	TCP	54	0 > 37002 [RST, ACK] Seq=1 Ack=1 Win=0 Len=0
264598	2645	78.130112000(	?	?	?	78	I, N(R)=16, N(S)=0; DSAP LLC Sub-Layer Management Individual, SSAP 0x88 Command

Figura 3: Trama de datos de entrada

Esta fuente con más de 800.000 registros, sirvió para estudiar y modelar las condiciones que evidencian un ataque por denegación de Servicios. La misma y tras tareas de preprocesamiento donde se filtraron registros con valores faltantes, se transformó el atributo info mediante tareas de filtrado de texto, se estableció una ventana temporal de análisis y desde la aplicación de un clasificador y en RM 5.3.015, se logró modelar lo que en principio permite reconocer un ataque por denegación de servicios.

La aplicación inicialmente emula el análisis de un experto en el análisis de ataques por denegación de servicios y posteriormente modela la salida "label" "flood" bivaluada asignando un 1 al reconocimiento de un ataque y 0 en caso contrario. Sobre esta nueva columna se realiza un clasificador basado en árboles que modela el ataque por denegación de servicios y cuya validación se realiza mediante el módulo específico del entorno de software RM logrando una performance expuesta en la Figura 4.



	true false	true true	class precision
pred. false	121660	0	100.00%
pred. true	2	194773	100.00%
class recall	100.00%	100.00%	

Figura 4: Matriz de confusión que presenta las predicciones de un ataque por denegación de servicios

Líneas de Investigación, Desarrollo e Innovación

El objetivo del laboratorio recientemente creado es profundizar en las tareas de investigación aplicada en lo concerniente a extracción de conocimiento en datos masivos para lo que se deberá propender a la utilización de hardware distribuido o a nivel del cloud computing amalgamando los conocimientos y actividades de cada uno de los grupos de investigación de los que deriva el laboratorio. En particular los datos masivos habrán de ser recabados y provistos por el Grupo de Astronomía Extragaláctica (GAE) de la FCEFN sobre los que habrán de ejecutarse algoritmos de extracción de conocimiento basados en entornos de aprendizaje de máquina que se ejecuten en plataformas paralelo-distribuidas.

Resultados y Objetivos

Los resultados que marcan el prólogo de la formación del laboratorio resultan auspiciosos como lo evidencian los trabajos presentados, y áreas temáticas comunes en que cada uno de los grupos

de investigación original hará de soporte al otro, generando una sinergia favorable.
Según la ordenanza, 02/2015-CD-FCEFN, de creación del laboratorio de sistemas inteligentes para la búsqueda de conocimiento en datos masivos, el mismo tiene como objetivo, extraer conocimiento desde grandes bases de datos mediante la utilización de algoritmos de minería de datos y aprendizaje de máquina soportado por arquitecturas secuenciales y paralelos-distribuidas
Se espera obtener conocimiento, sobre datos del área de la astronomía, que resulten ser fiable y contrastable con aplicaciones estadísticas llevadas adelante por el grupo GAE, pero que desde aplicaciones distribuidas y herramientas de software aprendizaje de máquina y extracción de conocimiento, permitan mejorar la performance de resultados anteriores.

Formación de Recursos Humanos

En el último año desde la actividad de ambos proyectos de investigación se han realizado las siguientes aplicaciones y trabajos de grado y posgrado que han permitido la formación de los siguientes recursos humanos.
En el transcurso del año 2015 desde la defensa de trabajos finales de grado que permitieron coronar las carreras en Licenciaturas en Ciencias de la Computación y Licenciatura en Sistemas de Información, , se defendieron cinco trabajos finales enmarcados en las áreas llevadas adelante por el grupo de investigación en tareas de DM, TM, WA, mediante la aplicación a diferentes tipos de datos, de algoritmos secuenciales y paralelos, mediante la utilización de herramientas de software libre y en hardware multicore y GPU computing.
Al momento y en ambas carreras se están dirigiendo otros seis trabajos finales que seguramente habrán de ser defendidos a la brevedad.
En el ámbito de posgrado se llevan adelante cuatro tesis de maestría, así como tres integrantes del grupo han comenzado a delinear sus líneas de trabajo de Doctorado en Ciencias de la Informáticas.
Por último y como parte de los objetivos planteados en el proyecto 21/E951 y en el ámbito de la FCEFN, se dictó el curso de posgrado “Búsqueda de Conocimiento en Datos” que con una duración de 50hs se dictó en Agosto y Setiembre de 2015, a la vez que se pretende durante el presente ciclo lectivo brindar un curso en el marco de la misma temática pero con profundidad.

Referencias

- Klenzi, R. O., & Araya, J. M. (2014, October). Minería de texto en la determinación automática de código Dewey. XVI Workshop de Investigadores en Ciencias de la Computación.
- Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. SpringerScience& Business Media.
- Maimon, O., &Rokach, L. (Eds.). (2005). Data mining and knowledge discovery handbook (Vol. 2). New York: Springer.
- Mens, J. P. (2008). Alternative DNS Servers: Choice and Deployment, and Optional SQL/LDAP Back-Ends. UIT Cambridge Ltd..
- North, M. (2012). Data mining for the masses.