Modelo de extracción de información desde recursos web para aplicaciones de la planificación automática

Model information extraction from the web based on design patterns

Carlos Henríquez Miranda[1], Jaime Alberto Guzmán[2]

[1] *Estudiante de Doctorado en Ingeniería de Sistemas. Magíster en Ingeniería de Sistemas –Grupo de Investigación SINTELWEB- Facultad de Minas Universidad Nacional de Colombia*
chenriquezm@unal.edu.co
[2] *Ph.D. Profesor Asociado Escuela de Sistemas Facultad de Minas. Universidad Nacional de Colombia. Sede Medellín.*

Recibido 15/09/12, Aceptado 22/12/2012

RESUMEN

En la actualidad nos encontramos con un alto exceso de información representado en un gran número de documentos electrónicos localizados en distintos lugares y en diferentes formatos. Por ejemplo, la Web, contenedor más grande de conocimiento, ofrece una gran cantidad de información plasmada en diferentes presentaciones como Wikis, blogs, portales, redes sociales entre otras. En el caso de las Wiki se encuentra información proveniente de un grupo de usuarios en diferentes disciplinas, donde se puede extraer conocimiento valioso de sus contenidos. El problema radica en que son representados casi siempre en lenguaje natural, haciendo que la búsqueda y recuperación sea un proceso complejo para los usuarios interesados. Debido a esto, el área de extracción de información se encarga de extraer a partir de recursos, datos útiles dependiendo de una necesidad de información. El enfoque de este trabajo es usar la extracción de información para recuperar automáticamente planes de tareas desde la Web y llevarlos a un proceso de automatización bajo el enfoque de la planificación automática. En este artículo, se presenta el resultado de investigación parcial de un modelo propuesto para la extracción, particularmente se muestran los resultados de las herramientas de extracción y pre procesamiento.

Palabras clave: Extracción información, Planificación Automática, Wrapper, Extracción Web, PDDL

ABSTRACT

Currently we have a high excess of information represented in a large number of digital documents located in different places and in different formats. For example the Web, larger container of knowledge, provides a lot of information embodied in different forms such as Wikis, blogs, websites, social networks and more. In the case of the Wiki is information from a group of users in different disciplines, which can extract valuable knowledge of its contents. The problem is that they are represented mostly in natural language, making search and retrieval be a complex process for users interested. Because of this, extraction area is responsible for extracting information from resources, useful data depending on an information need. The focus of this paper is to use data mining to automatically extract task plans from the Web and take it to an automation process under the automatic planning approach. This article presents the results of partial investigation of a proposed model for the extraction, particularly shows the results of extraction tools and pre processing.

Keywords: Information Extraction, Automatic Planning, Wrapper, Web Extractions, PDDL

1. INTRODUCCIÓN

Hoy día existe un gran exceso de información encontrada en una enorme cantidad de documentos electrónicos sobre diferentes fuentes y formatos, unas veces con conocimiento útil y en otros datos tipo basura. Actualmente existe una gran variedad de herramientas tecnológicas que aprovechan la experiencia de los seres humanos y máquinas para extraer conocimiento a partir de toda esta infoxicación [1]. Particularmente la Web se ha convertido en el mayor repositorio de recursos electrónicos en todas las áreas en diferentes presentaciones, formatos y estructuras. Gran parte de estos recursos se presentan en lenguaje natural en forma de blogs, wikis o redes sociales [2] otras en formas semi y estructuradas. Por esta falta de estructura estándar el proceso de recuperación de información en estos recursos se vuelve una tarea difícil. El área conocida como recuperación de información (IR) aborda este problema encargándose de traer documentos relevantes desde una gran repositorio en respuesta a un criterio definido [3]. Como existe mucha información se han creado buscadores que se encargan de recuperarla por intermedio de consultas clave [4]. Pero no solo es traer documentos relevantes de una consulta especifica, ya que en la Web cualquier individuo con un PC y una conexión puede brindar información valiosa en diferentes áreas como la economía, industria, medicina, robótica, planificación entre otras, sino que deben existir otro tipo de sistemas apoyados en técnicas de Inteligencia artificial (AI) que se encarguen de buscar dentro de los documentos, explorar su contenido y extraer información pertinente de un tema en particular. Estas nuevas herramientas se enmarcan en el área conocida como Extracción de información (IE) que se ocupa de estructurar los contenidos dentro de los textos que son relevantes para el estudio de un dominio particular, llamado dominio de extracción [5]. En otras palabras, el objetivo de un sistema de IE es encontrar y enlazar la información relevante mientras ignora la extraña e irrelevante [6].

Una tarea de IE es definida por su entrada y su extracción objetivo. La entrada puede ser documentos no estructurados como texto libre (por ejemplo, Figura 1) escritas en lenguaje natural o documentos semi-estructurados que se presentan constantemente como tablas o listas detalladas y enumerados, (por ejemplo, Figura 2) y lo que se extrae pueden ser nombres de personas, ciudades, autos, marcas etc., siempre dependiendo del criterio especifico que represente las necesidades de búsqueda.

Figura 1. Información en lenguaje natural.
Figure 1. Natural lenguaje information

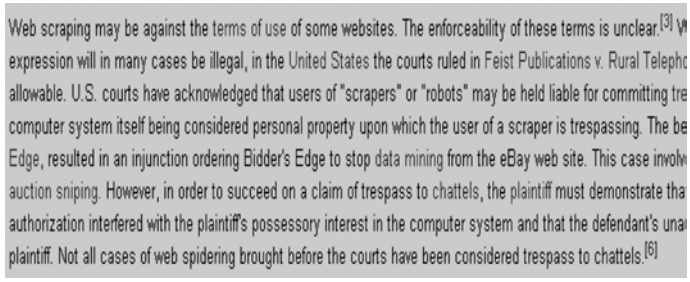
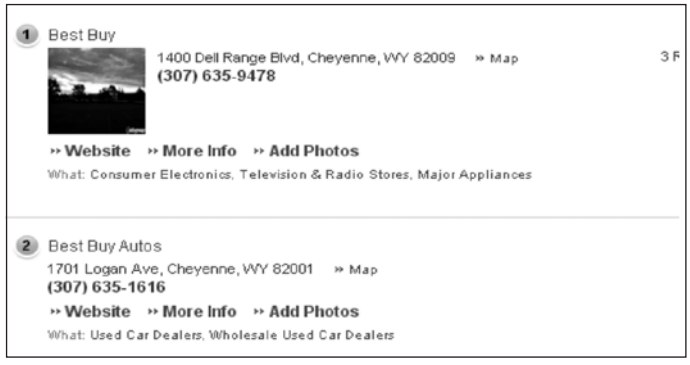

Web scraping may be against the terms of use of some websites. The enforceability of these terms is unclear.[3] W
expression will in many cases be illegal, in the United States the courts ruled in Feist Publications v. Rural Telepho
allowable. U.S. courts have acknowledged that users of "scrapers" or "robots" may be held liable for committing tre
computer system itself being considered personal property upon which the user of a scraper is trespassing. The be
Edge, resulted in an injunction ordering Bidder's Edge to stop data mining from the eBay web site. This case involv
auction sniping. However, in order to succeed on a claim of trespass to chattels, the plaintiff must demonstrate tha
authorization interfered with the plaintiff's possessory interest in the computer system and that the defendant's una
plaintiff. Not all cases of web spidering brought before the courts have been considered trespass to chattels.[6]

Figura 2. Información semiestructurada
Figure 2. Semistructured information



La importancia de la IE radica en encontrar datos útiles desde la Web que han sido colocados por verdaderos expertos en temáticas o áreas complejas, que puedan brindar la solución a muchos problemas existentes hoy día. El área de IE constituye un campo muy amplio ya que presentan una gran utilidad en todas las áreas profesionales donde se maneje cualquier tipo de información.

Entre los trabajos más representativos de la literatura se han abordado diferentes dominios de información, usando múltiples formas para llevar a cabo el proceso de IE. Por ejemplo : en [7] se describe un enfoque para análisis de la comunicación empresarial específicamente extrayendo información de los correos electrónicos, por su parte en [8] se ubican entidades y sus respectivos conceptos a partir de la exploración de la tablas en un documento HTML, en [9] se presenta un paradigma de extracción que facilita el descubrimiento de relaciones extraídas de texto, independiente del dominio y del tamaño del recurso Web, por otro lado en [10] se explota la apariencia visual de la información impulsado por relaciones espaciales que se producen entre los elementos de una página HTML, en [11] donde se construyen herramientas robustas para la extracción de información Web ante cambios en la estructura de la página basadas en un modelo de costo mínimo, finalmente en [12] se propone un enfoque de descubrimiento de patrones para la rápida generación de extractores de

información. En este estudio, particular atención tiene [13] y [14] donde a partir de un sitio Web se obtienen un conjunto de acciones referentes a planes y se representan en un lenguaje declarativo conocido como PDDL [15], para que sirvan de solución en diferentes problemas referentes con la planificación automática (AP). La AP busca que un ordenador, agente, robot realice tareas de forma automática sin intervención humana. Una de las motivaciones para la planificación automática es el diseño de herramientas de procesamiento de información que dan acceso a la eficiente planificación de recursos [16]. El trabajo de [13] extrae desde una página de WikiHow información relevante previo estudio de la ubicación de la entidad a extraer.

En este artículo se parte de [13] y [14] para proponer un modelo inicial para la extracción de información de planes de tareas desde la Web a partir de nuevas herramientas y técnicas computacionales tratando de mejorar la precisión de la extracción. La motivación de este trabajo consiste en que a partir de un documento plasmado por un conjunto de expertos planificadores en lenguaje natural se puedan extraer planes de acciones que puedan resolver problemas enfocados a la AP tales como: planes de turismo, logística de transporte, gestión de ascensores, vigilancia entre otras. Luego representarlos en un lenguaje como PDDL para llevar a cabo la tarea automáticamente.

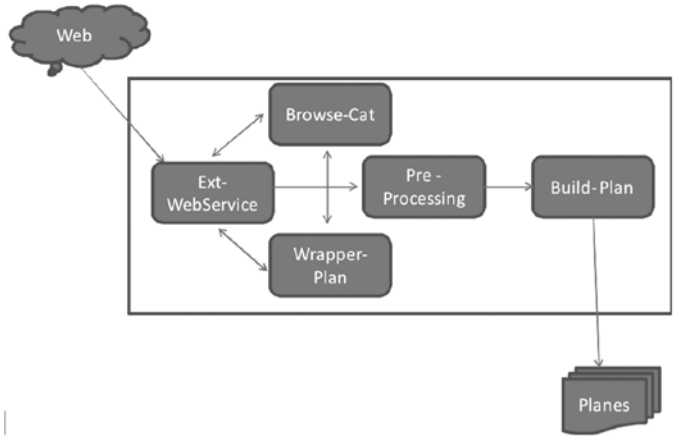
Las herramientas de extracción mostradas aquí se basan en wrapper o envoltorios que se encargan de obtener la traza de datos útiles a partir de un documento. En nuestro caso especial la exploración se hace en un documento Web en donde el wrapper explora toda su estructura HTML en busca de información relevante dependiendo de patrones iniciales [17]. Con los wrapper se combinan servicios Web basados en Web RestFul que permiten tener diseños más simples, bajo consumo de recursos, uri por recurso y generalmente por ser servicios fáciles de construir y adoptar [18] [19] [20].

Adicionalmente se explorarán otras herramientas para la eliminación de ruido y clasificación del sentido de las entidades encontradas. El resto del artículo se organiza de la siguiente manera. En la sección numero 2 se propone el modelo de extracción y sus componentes. En la sección 3 se muestran los resultados y la discusión. Finalmente las conclusiones se abordan en la sección 4.

2. MODELO DE EXTRACCIÓN

El modelo inicial es presentado en la Figura 3 y tiene como objetivo que a partir de un sitio Web (por ejemplo WikiHow) que tenga asociado contenido de planes de tareas extraerlos para posteriormente permitir su automatización. Está dividido en cuatro componentes: El primero (Ext-Web Servicie) que recibe una URL (dirección donde está la pagina) y junto con otros componente entrega una trama de entidades al componente Pre-Processing. El segundo (Browse-Cat) toma una página (formato HTML) que contiene una categoría de la WikiHow y obtiene todos los planes de esa categoría. El tercero (Wrapper - Plan) recibe una página y retorna la totalidad de pasos o acciones para ejecutar ese plan. El cuarto toma la lista de pasos y pasa por una etapa de pre-procesamiento eliminado las palabras que hacen ruido y clasificando las mismas semánticamente. El último (Build-plan) toma el plan refinado en formato textual para convertirlo a un plan en formato PDDL (aun en construcción).

Figura 3. Modelo Extracción Planes
Figure 3. Plan model's extraction



2.1 Ext –webservicie

Este componente es la interface del modelo que ofrece dos servicios Web RestFul. El primero recibe una URL que contiene la dirección del recurso que representa lo que se va a extraer y devuelve dependiendo de la petición una lista de acciones de un plan o una lista de planes asociados a una categoría. Para hacer esto se apoya en los dos procesos subsecuente Wrapper-plan y Browse-Cat. El proceso puede hacerse tomando una categoría y procesando todos los planes de esa categoría o simplemente la extracción de los pasos de un plan.

2.2 Browse-cat

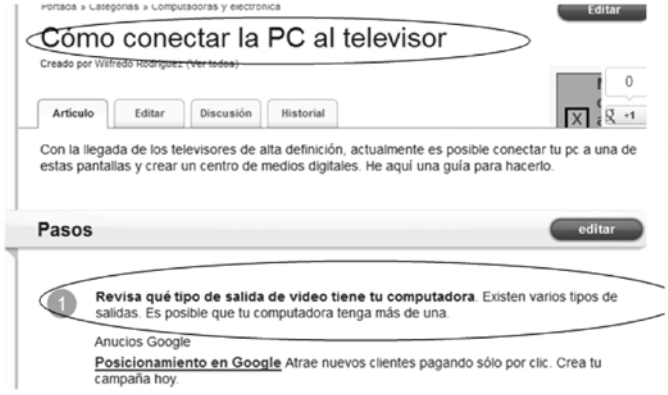
Este componente recibe un recurso que representa la página HTML que contiene una categoría de planes (ver Figura 4). A partir de la página que es ingresada se obtiene un conjunto de URL que representan las páginas o recursos donde se encuentran los planes de tareas.

Figura 4. Categoría de WikiHow
Figure 4. WikiHow's category

2.3 Wrapper –plan

Este componente recibe la página a procesar de su predecesor y obtiene a partir de ella el plan y un conjunto de acciones. Para este componente se analizó previamente la estructura de la página y se hizo una aplicación (wrapper) que explora los tags(etiqueta de HTML). Este componente recibe tanto las páginas en la versión en español[1], y en Ingles[2]. En la figura 5 se muestra un recurso de Wikihow (como conectar la PC al televisor) en donde se obtendrá el plan y el conjunto de pasos. En la figura 6 podemos observar el resultado de la ejecución del componente cuando se le pasa la url: *"http://es.wikihow.com/conectar-la-PC-al-televisor"*.

Figura 5. Plan en WikiHow
Figure 5. WikiHow's Plan



[1] http://es.wikihow.com/Portada
[2] http://www.wikihow.com/Main-Page

Figura 6. Resultado componente Wrapper Plan
Figure 6. Wrapper component Outcome Plan

```
Plan es  Cómo conectar la PC al televisor
1 Revisa qué tipo de salida de video tiene tu computadora
2 Revisa qué tipo de entrada tiene tu televisor
3 Consigue el cable correcto para llevar a cabo la conexión
4 Consigue un cable de audio si es necesario
5 Apaga la computadora
6 Desconecta el monitor
7 Coloca tu computadora cerca del televisor
```

2.4 Pre-processing

Este componente toma toda la traza enviada por el Ext-Web Servicie después del proceso que se hace en Wrapper Plan y realiza las actividades de clasificación y limpieza. Lo primero es tomar cada entidad que hace parte de un paso del plan y clasificarla en alguna de las siguientes categorías: nombres, sustantivos, adjetivos y adverbios. Estas categorías se toman de WordNet base de datos léxica en inglés que agrupa entidades en conjuntos de sinónimos cognitivos (synsets), cada uno expresando un concepto distinto [21]. Esta clasificación se realiza usando la herramienta *Stanford POS Tagger* creada por investigadores del grupo de procesamiento del lenguaje natural en la Universidad de Stanford [22]. Luego de esta etapa se limpian los datos borrando algunas palabras sin significado como artículos, pronombres, preposiciones, etc. Estas palabras en el idioma ingles se conocen como stopWords [23].

2.5 Build-plan

Luego del pre- procesamiento se toma el listado restante, la traza completa y se construye los planes en el lenguaje PDDL. Este componente está en construcción.

3. RESULTADOS

La implementación del modelo se hizo en *Java 7.0* con uso de librerías especiales *HttpClient 3.1* usando *Netbeans IDE 7.11* bajo la arquitectura Rest usando Web services y como servidor *GlassFish 3.1.2* . Para las categorías de las entidades se uso *WordNet 2.1* para Windows y la interfaz para java *JWI* versión 2.2.3 (*Java Wordnet Interface*). Para la clasificación de entidades se uso la herramienta *Stanford POS Tagger.*

Para el experimento se tomo el sitio Web WikiHow y se pasaron al modelo las direcciones (URL) de diferentes categorías (5) que organizan un conjunto de planes. En la Figura 7 se muestra el resultado específico de la categoría: *Pasatiempos*, el cual arrojó una totalidad de 201 planes de tareas en los cuales se extrajeron una totalidad de 1803 pasos.

Figura 7. Partes de ejecución de extracción categorías
Figure 7. Performance parts's extraction categories.

```
Categoría planes: http://es.wikihow.com/Categor%C3%ADa:Pasatiempos 201 Recuperados

Total pasos 1803

http://es.wikihow.com/hacer-rosas-de-seda%2C-sat%C3%ADn-o-list%C3%B3n

http://es.wikihow.com/hacer-un-palo-de-lluvia

http://es.wikihow.com/Fazer-medalh%C3%B5es-%C3%A9tnicos

http://es.wikihow.com/Fazer-uma-l%C3%A2mpada-com-a-estopa

http://es.wikihow.com/Ganar-el-modo-clasico-en-super-smash-bros-brawl

http://es.wikihow.com/Ser-Las-Tortugas-Ninja-en-Club-Penguin
```

Por cada categoría pasada se tomaron las URL de los planes, se exploraron y extrajeron un grupo de pasos. En la Figura 8 se muestra el recurso Web "*How to Prepare Your Boat for Transport*" al cual se aplicó el proceso de extracción.

Figura 8. Página WikiHow
Figure 8. WikiHow's page

En cada recurso para realizar el proceso de extracción de información asociada a lo que se requiere, se identificaron patrones en donde se localizaban el plan y sus pasos. Se revisaron muchas páginas similares y se obtuvo el patrón para la extracción. Por ejemplo para sacar el titulo, en la pagina se encuentra que este aparece rodeado de los tags *<TITLE></TITLE>*. Para los pasos se emplea un procedimiento similar. En la figura 9 se muestra el resultado del proceso. Note que solo se extrae lo útil e importante según el fin específico.

Figura 9. Partes ejecución de extracción plan
Figure 9. Parts's Plan execution extraction

```
Plan es  How to Prepare Your Boat for Transport: 20 steps
1 Take the time to plan for your boat move
2 Choose a Professional Boat Transport Company experienced in moving y
3 Communicate with your Boat Transportation Company
4 Measuring Your Boat for Transport
5 Understand how payment is to be made for both the boat transport ser
6 Remove and properly store the following items
7 Items to be removed inside your boat
8 Items to be removed outside your boat
9 Check for Zebra Mussels
10 Decide whether to shrink-wrap your boat or not
```

Luego se tomo la traza de cada plan dividida en el conjunto de pasos y se entro a la etapa de pre-procesamiento. Cada paso consta de una serie de palabras que llamamos entidades. Cada entidad se envía a clasificar según la categoría planteada (VERB, NOUN, ADVERB, ADJECTIVE). Para nuestros propósitos lo más importante en este espacio es encontrar cual entidad es el verbo representativo de la acción de cada paso para futuros propósitos de creación del plan en PDDL (por ejemplo ver Figura 10).

Figura 10. Partes ejecución clasificación entidades
Figure 10. Parts execution of entities classification

```
 Entidades clasificadas por paso
Loading default properties from tagger models/wsj-0-18-left3words.tagger
Reading POS tagger model from models/wsj-0-18-left3words.tagger ... done [1,1 sec].
 Paso 1
 Take VERB  the N/A  time NOUN  to N/A  plan NOUN  for N/A  your N/A  boat NOUN  move NOUN
 Paso 2
 Choose VERB  a N/A  Professional NOUN  Boat NOUN  Transport NOUN  Company NOUN  experienced ADJECTIVE
 Paso 3
 Communicate VERB  with N/A  your N/A  Boat NOUN  Transportation NOUN  Company NOUN
 Paso 4
 Measuring VERB  Your N/A  Boat NOUN  for N/A  Transport NOUN
 Paso 5
 Understand NOUN  how ADVERB  payment NOUN  is VERB  to N/A  be VERB  made VERB  for N/A  both N/A  th
```

Como se nota en la figura 10, cada entidad es clasificada según nuestros propósitos. Las que no tienen ningún significado para nuestro estudio aparecen como N/A. Después del proceso de clasificación encontramos que existen varios verbos identificado en algunos pasos, lo que acarrea problemas para la futuro representación en PDDL. Así que en el siguiente paso se tiene en cuenta este inconveniente y además se eliminan las palabras que causan ruido (stopWords) quedando una traza de entidades más limpias. (Figura 11).

Figura 11. Partes ejecución del borrado de stopwords
Figure 11. Parts execution to stopwords's delete

```
Entidades borradas                63
Total entidades procesadas        203
Despues de Eliminacion de Entidades  31.034481%
 Paso 1
 Take VERB  time NOUN  plan NOUN  boat NOUN  move NOUN
 Paso 2
 Choose VERB  Professional NOUN  Boat NOUN  Transport NOUN  Company NOUN  experienced ADJECTIVE
 Paso 3
 Communicate VERB  Boat NOUN  Transportation NOUN  Company NOUN
 Paso 4
 Measuring VERB  Boat NOUN  Transport NOUN
 Paso 5
```

Para la prueba de las herramientas creadas se examinaron un conjunto de categorías de la WikiHow y e l resultado se ve en la tabla 1.

Tabla 1. Resultados de extracción por categoría
Table 1. Result extraction category

Categoría	Planes Hallados	Pasos por plan
http://es.wikihow.com/Categor%C3%ADa:Pasatiempos	201	1803
http://es.wikihow.com/Categor%C3%ADa:Vida-familiar	201	1681
http://es.wikihow.com/Categor%C3%ADa:En-el-trabajo	54	465
http://es.wikihow.com/Categor%C3%ADa:Arte-y-entretenimiento	201	1622
http://es.wikihow.com/Categor%C3%ADa:Viajes	93	892

Adicionalmente se tomaron varios recursos de planes al azar, cuyo resultado se muestra en la tabla 2.

Tabla 2. Resultados de eliminación StopWords
Table 2. Results stopwords's delete

Plan a explorar	Entidades Procesadas	StopWords	% StopWords eliminados
How to Make the Most of Your Money 19 Tips - wikiHow	356	131	36.79
How to Prepare Your Boat for Transport	203	63	31.03
How to Make Caribbean Carob Cake 18 steps - wikiHow	174	72	41.37
How to Add Medicinal Plants to Your Garden 12 steps - wikiHow	113	36	31.85
How to Make Shot Glass Candles with Step-by-Step Pictures	104	36	34.61
How to Clean a House 13 steps (with pictures) - wikiHow	58	15	25.86

A partir de los resultados obtenidos podemos decir que para todas las categorías analizadas en el experimento arrojaron el conjunto de planes y acciones esperadas según el recurso brindado. El máximo de planes hallados fue 201 que es el máximo número de planes por categoría y los números de pasos más altos encontrados fueron de 1803. Para todos los planes hubo una extracción de la lista de pasos sin ningún inconveniente tanto en pruebas de páginas en español y en Ingles. Al hacer un análisis ad-hoc de forma manual de algunos planes de salida, se muestra que el sistema funcionó muy bien en la extracción del conjunto de planes por categoría y el conjunto de pasos por cada plan.

Después de la extracción se nota que las herramientas de clasificación y limpieza funcionan adecuadamente y que el % de entidades sin significado en los planes es alta (mayor que 26%). Cabe anotar que este es un trabajo inicial antes de entrar a los siguientes procesos que son el análisis semántico más detallado del sentido de las entidades y su respectiva desambiguación (Build-Plan).

También cabe anotar que se tuvo ciertos problemas cuando la categoría poseía más de 201 tipos de planes porque no se logró explorar el nuevo recurso Web que se recargaba con técnicas de desarrollo Web (Ajax). Sin embargo consideramos que por ser el primer prototipo se logro nuestro objetivo con una alta eficiencia logrando extraer la información requerida. De todas formas las tecnologías que hacen parte de las herramientas usadas están sujetas a mejoras en el futuro. En cuanto a los wrapper implementados cumplieron sus objetivos pero no están preparados para posibles cambios en la estructura HTML. Para eso se buscará el uso de técnicas inductivas (wrapper induction).

4. CONCLUSIONES

- En este artículo se han mostrado los resultados iniciales de las herramientas de extracción construidas que responden a un modelo propuesto de extracción de información. Estas pretenden unirse al grupo de herramientas que permitirán la construcción de forma automática de planes a partir de la existencia del conocimiento existente en la Web.

- En los resultados obtenidos se muestra que la extracción de información es un área que puede trabajarse para la solución de muchos problemas en diferentes dominios de información como medicina, comercio, industria, transporte, academia, publicidad entre muchas otras.

- Estas herramientas son un prototipo inicial que busca mejorar en el caso de los wrapper haciendo uso

de técnicas inductivas que permitan la extracción de información independiente del formato del recurso. Adicionalmente se busca más adelante realizar la clasificación de los planes usando nuevas tecnologías que aporta la Web semántica como las ontologías.

REFERENCIAS

[1] Cornella, A. (2011). *Infoxificacion* [Internet]. Disponible desde: <http://www.infonomia.com/articulo/ideas/7150> [Acceso 1 Septiembre de 2012].

[2] Pérez, L. (2011). *Redes Sociales, Blogs y Wikis: Tendencias y realidades* [Internet]. Disponible desde: <http://www.slideshare.net/gentedeinternet/blogs-redes-sociales-y-wikis> [Acceso 05 de Mayo de 2012].

[3] Martínez, F., Recuperación de información: Modelos, sistemas y evaluación, EL KIOSKO JMC, Murcia España, 2004.

[4] Olivera, M. D., Métodos y técnicas para la indización y la recuperación de los recursos de la World Wide Web, Boletín de la Asociación Andaluza de Bibliotecarios, Año nº 14, Nº 57, 11-22, 1999.

[5] Téllez, A. Extracción de Información con Algoritmos de Clasificación. Tesis de Maestria, Instituto Nacional de Astrofísica, Óptica y Electrónica, 2005.

[6] Cowie, J. Information Extraction. Magazine Communications of the ACM, volumen 39 Numero 1, 80 – 91, 1996.

[7] Laclav´ık, M., Dlugolinsky´, S. and Seleng, M. . Email analysis and information extraction. Computing and Informatics, volumen 30, numero. 1, 57-87, 2011.

[8] Dalvi, B., Cohen, W. W., and Callan, J. WebSets: Extracting Sets of Entities from the Web Using. WSDM '12 Proceedings of the fifth ACM international conference on Web search and data mining , 243-252, 2012.

[9] Banko, M., Cafarella, M., Soderland, S., Broadhead, M., and Etzioni, O. Open Information Extraction from the Web. Magazine Communications of the ACM, volumen 51 numero 12, 68-74, 2008.

[10] Penna, G., Magazzeni, D., and Orefice, S. Visualextraction of information from webpages. Journal of Visual Languages & Computing, volumen 21, numero 1, 23–32, 2010.

[11] Liu, D., Wang, X., Li, L., and Yan, Z.. Robust Web Extraction Based on Minimum Cost Script Edit Model. Procedia Engineering , volumen 29, 1119–1125, 2012.

[12] Chang, C.-H., Hsu, C.-N., & Lui, S.-C. Automatic information extraction from semi-structured Web pages by pattern discovery. Decision Support Systems. Volumen 35 numero 1, 129 – 147, 2003.

[13] Addis, A., Armano, G., & Borrajo, D. (2009). Recovering Plans from the Web.

[14] Addis, A.M; Borrajo, D. (2011). From Unstructured Web Knowledge to Plan Descriptions. En A. Soro, Information retrieval and mining in distributed enviroments. Volumen 324, 41-59, 2011.

[15] Kvarnström, J. (2012). *TDDD48 Automated planning (6 ECTS*) [Internet]. Disponible desde <:http://www.ida.liu.se/~TDDD48/labs/2012/pddl.en.shtml> [Acceso 05 de Mayo 2012].

[16] Ghallab, M., Nau, D., and Traverso, P. Automated Planning: Theory and Practice. Morgan Kaufmann Publishers, Usa, 2004.

[17] Rodríguez, M. (2004). *Curso: Arquitecturas de bases de datos en la distribución UPM [Internet]*. Disponible desde< http://sinbad.dit.upm.es/docencia/doctorado/curso0304/Wrappers.pdf> [Acceso 6 de Junio de 2012].

[18] Tyagi, S. (2006). *De RESTful Web Services* [Internet]. Disponible desde <http://www.oracle.com/technetwork/articles/javase/index-137171.html> [Acceso octubre de 2012].

[19] Navarro, R. (2007). *Rest vs Web Service* [Internet]. Disponible desde <http://users.dsic.upv.es/~rnavarro/NewWeb/docs/RestVsWebServices.pdf> [Acceso Agosto de 2012].

[20] Rodríguez, A. (2008). *RESTful Web services: The basics* [Internet]. Disponible desde <http://www.ibm.com/developerworks/webservices/library/ws-restful/> [Acceso Octubre de 2012].

[21] Princeton University. (2012). *WordNet A Lexical database for english* [Internet]. Disponible desde <http://wordnet.princeton.edu/> [Acceso Noviembre 2012].

[22] Toutanova, K., and Manning., C. D, Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Volumen 13, 63-70, 2000.

[23] RANKS.NL, *English Stopwords* [Internet].Disponible desde <http://www.ranks.nl/resources/stopwords.html> [Acceso Octubre 2012].