Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 5, Mes Mayo, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 14-24
http://publicaciones.uci.cu

Tipo de artículo: Artículo original
Temática: soluciones informáticas
Recibido: 25/03/17 | Aceptado: 02/05/17 | Publicado: dd/mm/aa

Herramienta para recopilar la información en las noticias publicadas en los sitios web de internet

Tool to collect the information in the news published on internet web sites

Ing. Ariannis Vargas Pérez [1*]

[1] Centro de Tecnologías de Gestión de Datos, Facultad de Ciencias y Tecnologías Computacionales, Universidad de las Ciencias Informáticas. ariannis@uci.cu

[*] Autor para correspondencia: ariannis@uci.cu

Resumen

Con la alta disponibilidad de datos producidos a diario en el mundo de la informática, las técnicas convencionales para su obtención no suelen aprovechar al máximo la información de valor en los mismos. Se hace necesario para ello, implementar nuevas técnicas y herramientas que sirvan de ayuda para solucionar esta problemática. El objetivo de este trabajo, enmarcado en la Universidad de las Ciencias Informáticas, es el desarrollo de una herramienta informática que permita llevar a cabo el proceso de monitoreo y seguimiento de la información contenida en las publicaciones web y sus comentarios utilizando el método de Web Scraping o Raspado Web, sirviendo de apoyo empresas o entidades en la toma de decisiones respecto a la información recopilada. En el documento se abordan, aspectos esenciales como son: características, ventajas, arquitectura e importancia de las herramientas de recopilación de información web y aporta también, el análisis y diseño necesario para lograr el desarrollo de la solución. La herramienta desarrollada permite a los usuarios: la obtención de los datos que conforman tanto las publicaciones web, como los datos en cada uno de sus comentarios; la actualización de las publicaciones existentes en la base de datos del sistema, así como el envío de reportes por correo electrónico con los resultados de cada operación. La investigación describe los beneficios alcanzados al utilizar la herramienta, en cuanto a la reducción de tiempo y esfuerzo para ejecutar el proceso de monitoreo y seguimiento desde la Web.

Palabras clave: Herramientas, Información, Recopilación, Web Scraping.

Abstract

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 5, Mes Mayo, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 14-24
http://publicaciones.uci.cu

With the great availability of data daily produced in the computer world, conventional techniques for obtaining them not usually make a good use of the most valuable information in them. For that reason it's necessary to implement new techniques and tools that help to solve this problem. The aim of this document, developed at the University of Informatics Sciences, is the development of an informatics tool that allows carrying out the process of monitoring and tracking information contained in the web publications and its comments using the method of Web Scraping, as support for companies or entities in making decisions about the information collected. This document also deals with key issues such as: features, benefits, architecture and the importance of the web information collection tools, and provides the necessary analysis and design to achieve the development of the tools. The developed tool allows users to: obtaining data that web publications, including data from each of your comments, updating existing publications in the database system as well as sending email reports with the of each operation. The research describes the benefits achieved by to use the tool in the reduction of time and effort to execute the process of monitoring and tracking from the Web.

Keywords: *Collection, Information, Tools, Web Scraping.*

Introducción

Los avances científicos-técnicos han constituido un papel fundamental en la evolución de diversos sectores de la sociedad, con el objetivo de dar respuesta a problemas existentes dentro de la misma. Este hecho trae consigo el surgimiento de nuevos conceptos relacionados con la informática y las tecnologías de la información. En los últimos años, el uso de las llamadas Tecnologías de Información y Comunicación (TIC) (MARIA *et al.* 2012), integración de la computación, las telecomunicaciones y las técnicas para el procesamiento de datos; se ha incrementado. Los procesos empresariales que antes se realizaban manualmente hoy son optimizados por dichas tecnologías.

Internet en los últimos años se ha convertido cada vez más en una extensa fuente de conocimientos y datos utilizables por entidades o empresas para su beneficio (SANTOS 2013). Mediante su estudio se pueden obtener aspectos significativos para la empresa o entidad como(ZALDÍVAR 2014):

Sistematicidad con que fluye la información.

- Detectar los temas y tendencias que identifiquen a la comunidad universitaria, al país y el resto del mundo.
- Efectuar análisis cualitativo-cuantitativo sobre un tema específico.

Estos aspectos pueden ser analizados mediante la extracción, almacenamiento y monitoreo de la información contenida en las publicaciones de sitios, comunidades y portales Web de la universidad, país y el resto del mundo.

Dicho proceso en la actualidad presenta la dificultad de ser realizado de forma manual por el personal destinado al mismo, extrayendo la información desde el HTML de las páginas accedidas campo por campo de la información deseada (título, fecha, autor, contenido, imágenes, comentarios, etc.). Esto se debe principalmente a en ocasiones no poseer acceso a las diferentes fuentes de la información de los diversos sitios en la Web, entiéndase como fuente de información como la base de datos del sitio o los diversos mecanismos para compartir dicha información (GARZA-RÍOS *et al.* 2012).

El centro Ideo-Informática (CIDI), perteneciente a la Universidad de las Ciencias Informáticas (UCI), presenta como misión proveer soluciones, productos y servicios relacionados con las tecnologías de Internet, en función de la defensa de la ideología socialista en la red de redes. Dentro de las tareas del centro se encuentra la extracción, almacenamiento y monitoreo de la información contenida en las publicaciones de sitios, comunidades y portales Web de la universidad, país y el resto del mundo. Este centro aprovecha cada una de los indicadores que brinda el estudio de la información recopilada para el estudio interno y la toma de decisiones sobre temas que influyan en la comunidad universitaria, del país y el resto del mundo.

El proceso de monitoreo y seguimiento efectuado actualmente por el centro CIDI, es dividido en dos fases:

- Recopilación: Proceso de identificación de las nuevas publicaciones en los sitios web y la obtención de los datos, haciendo referencia a: título, dirección, autor, cantidad de visitas, fecha, fuente de origen, cuerpo de la publicación y sus comentarios; de los cuales se obtienen: el autor, la fecha de realización y el texto del mismo.
- Actualización: Corresponde a la actualización de los datos recopilados anteriormente de las publicaciones. Se actualiza la cantidad de visitas, la cantidad de comentarios, se agregan los comentarios nuevos y algún cambio en el cuerpo de la publicación.

Debido al crecimiento sostenido de la información publicada en la Web, hace ineficiente dicho proceso de recopilación; convirtiéndose en una tarea agotadora, monótona y sobre todo lenta ya que se necesitan varias horas para su recopilación solamente . Esta situación afecta directamente el consumo de tiempo y personal destacando la necesidad de buscar nuevos métodos para la obtención de dicha información.

A partir de la situación presentada se adopta como objetivo: Desarrollar una herramienta que permita mejorar el proceso de recopilación de la información en las publicaciones de sitios web de Internet, que contribuya a obtener mejores soluciones de soporte a la toma de decisiones para las empresas y organismos cubanos. Con este sistema se pretende que en lo adelante las soluciones desarrolladas en la universidad y/o centros de la misma puedan

concentrarse en diseñar mecanismos de análisis de la información ya recopilada. De esta forma garantizar una mejor experiencia a los especialistas y ejecutivos de las empresas y organismos cubanos en el análisis de la información, la planificación y la toma de decisiones.

Materiales y métodos

Las tecnologías de extracción de datos e información de la Web han supuesto una revolución en el campo de la interacción usuario-ordenador. Facilitan el acceso a una cantidad impresionante de información que puede ser transformada, convirtiéndose en beneficios para los intereses del usuario final. Con el trascurso del tiempo se han desarrollado un conjunto de métodos y técnicas para la recopilación de información en la Web y su posterior uso.

A continuación, se abordan sobre algunos métodos y técnicas utilizados para la recopilación de información en Internet:

Fig.1 Métodos y técnicas de recopilación de información en la Web

- Manual: Método más primitivo y básico de recopilar los datos de las páginas web, mediante el copiar y pegar (Ctrl+C y Ctrl+V) de los datos seleccionados. Este proceso es lento y por lo general, solo se realiza una vez.

- Búsquedas en Internet: Método mediante el cual buscadores web, como Google.com, Yahoo.es y otros, obtienen la información asociada a un parámetro de búsqueda para su posterior recopilación.

- Programación HTTP: Permite la creación de formas para facilitar el acceso y obtención de la información contenida en la Web. Entre sus principales usos en sitios web encontramos los RSS un formato XML para sindicar o compartir contenido en la Web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito. Este método a pesar de ser bastante utilizado, depende totalmente de la información que el sitio desee brindar o compartir y si el sitio presenta o no el uso de RSS, por lo que esta opción no cumple lo requerido para la solución.

Los métodos anteriormente descritos son factibles para la obtención de información como un conjunto o un todo, pero si se desea obtener algunos datos específicos de dicha información, se debería recurrir al método manual para su realización. Para resolver las dificultades anteriormente mencionadas surgen los métodos de Web Scraping, en español también conocidos como métodos de Raspado Web.

Métodos de web Scraping

Los métodos de Web Scraping permiten realizar la búsqueda, descarga y procesamiento de información de manera programática y automática. Existe una amplia gama de empresas que utilizan programas de Web Scraping, en español también se conoce como raspado web, para el proceso de obtención de la información contenida en el HTML de una página web. También se usan para realizar diferentes actividades como la investigación en línea, seguimiento de los cambios de datos del sitio web, la extracción de datos desde diferentes sitios web, entre otros.

De manera general, en este proceso intervienen los conceptos de crawler, scraper y spider. Aunque no existe una definición precisa para estos términos, hay algunas diferencias en su funcionamiento y los casos en que son usados (RIQUELME *et al*. 2006). A continuación, se abordan con más detalles los términos mencionados:

Crawler: Recorren los enlaces en la Web usando un sitio de partida y permite crear copias del contenido de los sitios visitados, de manera similar a un motor de búsqueda.

Spider: Conocidos en español como arañas web, permiten iterar a través de los enlaces en las páginas web hasta el nivel de profundidad indicado. Los enlaces son identificados mediante sus etiquetas <a> por lo que es requerido un análisis sintáctico del HTML.

Entre las tareas más comunes de los Spider o Arañas Web se encuentran:

- Crear el índice de una máquina de búsqueda.
- Analizar los enlaces de un sitio para buscar links rotos.
- Recolectar información de un cierto tipo, como precios de productos para recopilar un catálogo o el texto de publicaciones o noticias en la Web para su posterior análisis.
- Indexar páginas web, en tareas de mantenimiento comprobando enlaces o validando el código HTML.
- Reunir tipos específicos de información procedente de páginas web, como es el caso de cosechar direcciones de correo electrónico para enviar correos basura.

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 5, Mes Mayo, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 14-24
http://publicaciones.uci.cu

Scraper: Realizan la extracción de información de sitios específicos, buscando expresiones regulares, palabras clave, elementos, atributos, entre otros.

Para llevar a cabo la automatización de la búsqueda y procesamiento de información en la Web, los 3 procesos deben trabajar en conjunto (GARZÓN 2010), dando origen a las herramientas o aplicaciones de Web Scraping o raspado web. Las aplicaciones de un programa de raspado web son prácticamente ilimitadas, especialmente en la era moderna de la tecnología de la información, donde el Internet es la principal fuente de información (GONZÁLEZ 2013). Su funcionamiento, de forma general, consiste en obtener los datos desde las páginas web, insertarlos en la base de datos (BD) de la entidad u organización y monitorizar toda la operación realizada.

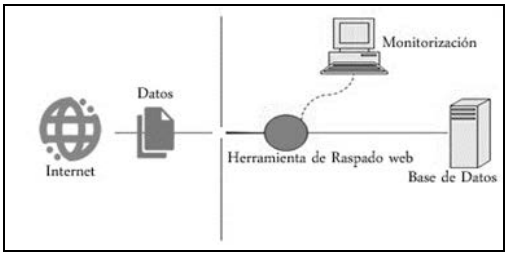

Fig.2 Funcionamiento general de un programa de Web Scraping

Existen varios tipos, dependiendo de su esquema de operación, entre los más básicos se encuentran:

- Uno a muchos: La estructura de uno a muchos es común de un scraper que busca por ejemplo información de precios de boletos de avión a través de diferentes sitios.

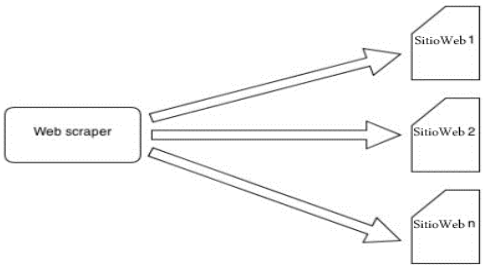

Fig.3 Esquema de web scraper de uno a mucho

- Uno a uno: Estructura más típica, se accede a la información de un sitio web específico.

Fig.4 Esquema de web scraper de uno a uno

Herramientas y tecnologías

En el desarrollo de la herramienta se emplearon las siguientes tecnologías:

- Python: Lenguaje de scripting orientado a objetos e independiente de la plataforma en que se ejecute. La utilización de este lenguaje de programación varía desde aplicaciones de escritorio a servidores de red o incluso páginas web. Es un lenguaje interpretado, que no necesita compilar el código fuente para poder ejecutarlo, ofreciendo como ventaja rapidez de desarrollo y mayor velocidad de ejecución (ANGEL 2008).

- Webscraping: Biblioteca para el lenguaje Python. Esta facilita la obtención de los datos deseados de las páginas web seleccionadas de una manera fácil y rápida. Entre sus principales funcionalidades se encuentran mecanismos para la utilización de expresiones regulares, obtención de elementos mediante su dirección xpath, tratamiento de texto como eliminación de etiquetas HTML, entre otras. La característica que destaca su uso es que posee mecanismos para la creación y control de Threads (hilos de ejecución) para realizar las operaciones del sistema (GARCÍA and MONTERO 2013).

Algunos ejemplos de su utilización:

1. webscraping.search (html, "//h3 [@class="ejemplo"]/a/@href"): Se obtiene el atributo href de todos los vínculos (<a>) dentro de la etiqueta HTML <h3> con la clase = "ejemplo".

2. webscraping.remove_tags (texto): Se eliminan el conjunto de etiquetas HTML encontradas dentro del texto.

3. webscraping.threaded_get (listado, metodo1, metodo2): Crea un hilo de ejecución por cada una de las direcciones URL contenidas en el listado. Mediante la definición del metodo1 se obtiene el HTML de la página correspondiente a la dirección URL y mediante el método2 se realiza el procesamiento de la información contenida en la misma.

- PostgreSQL: Es un sistema de gestión de base de datos relacional orientada a objetos y libre, publicado bajo la licencia BSD. Como muchos otros proyectos de código abierto, el desarrollo de PostgreSQL no es manejado por una empresa y/o persona, sino que es dirigido por una comunidad de desarrolladores que trabajan de forma desinteresada, altruista, libre y/o apoyada por organizaciones comerciales. Sus principales características son estabilidad, potencia, robustez, facilidad de administración e implementación de estándares (SCRIBID 2014).

- Aptana Studio: IDE de software libre que posee la licencia pública general (GNU, por sus siglas en inglés) y la licencia pública Aptana basada en Eclipse y desarrollado por Aptana, Inc. Tiene plataformas para funcionar bajo Microsoft Windows, Mac OS X y GNU\Linux y provee soporte para lenguajes como: PHP, Ruby, CSS, HTML y Python. Presenta la posibilidad de incluir complementos para nuevos lenguajes y funcionalidades (MAR *et al*. 2013).

- PyQt: Binding (adaptación de una biblioteca para ser usada en un lenguaje de programación especifico) de la biblioteca de Qt para trabajar sobre el lenguaje de programación Python. Qt es una biblioteca multiplataforma ampliamente usada para desarrollar aplicaciones con una interfaz gráfica de usuario. PyQt, bajo las licencias GPL y LGPL, está desarrollada y disponible para Microsoft Windows, GNU/Linux y Mac OS X al igual que Qt, permitiendo simplificar el desarrollo de aplicaciones de escritorio.

- Psycopg2: Biblioteca para la conexión de Python con el SGBD PostgresQL.

Resultados y discusión

Previo a la implementación de la solución propuesta fueron definidos un total de 14 requisitos funcionales del sistema y 9 requisitos no funcionales. Todos estos requisitos fueron extraídos partiendo de los resultados del estudio previo del estado actual de las soluciones similares existentes y a entrevistas realizadas a los especialistas del centro CIDI, centro encargado del desarrollo de aplicaciones y herramientas que utilizan el análisis de la información existente en los sitios web de Internet en la universidad, país y el resto del mundo.

La herramienta obtenida cuenta con 3 funcionalidades principales: el módulo de recopilación de información, el módulo de actualización de la información ya extraída y la generación de reportes sobre las operaciones realizadas.

- Módulo de recopilación: Proceso de identificación de las nuevas publicaciones en los sitios web y la obtención de los datos, haciendo referencia a: título, dirección, autor, cantidad de visitas, fecha, fuente de origen, cuerpo de la publicación y sus comentarios; de los cuales se obtienen: el autor, la fecha de realización y el texto del mismo.

- Módulo de actualización: Corresponde a la actualización de los datos recopilados anteriormente de las publicaciones. Se actualiza la cantidad de visitas, la cantidad de comentarios, se agregan los comentarios nuevos y algún cambio en el cuerpo de la publicación.

- Módulo de reportes: Muestra un resumen sobre la información recopilada y/o actualizada.

Modelo de datos

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 10, No. 5, Mes Mayo, 2017
ISSN: 2306-2495 | RNPS: 2343
Pág. 14-24
http://publicaciones.uci.cu

El modelo de datos es un lenguaje utilizado para la descripción de una BD. Por lo general, un modelo de datos permite describir el tipo de datos que incluye la base y la forma en que se relacionan, así como las restricciones de integridad y las operaciones de manipulación de los datos.

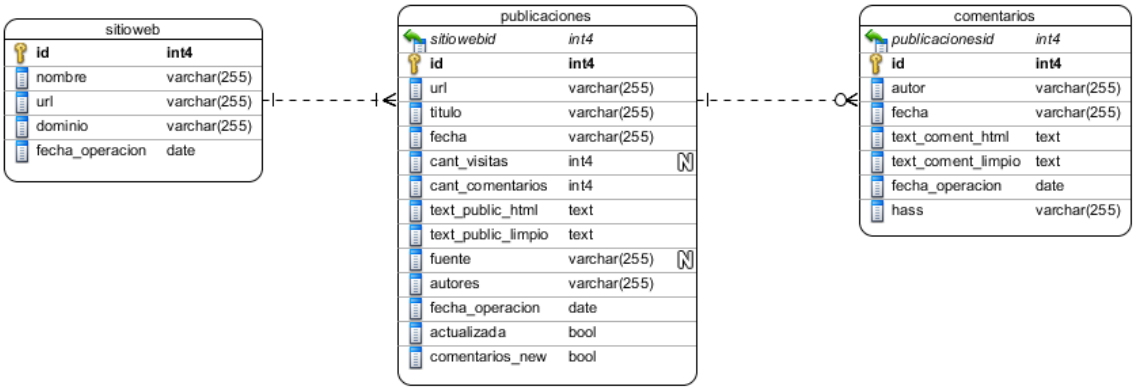

Fig.5 Diagrama entidad-relación del sistema

Aporte y Novedad

Con el empleo de la herramienta para recopilar la información en las noticias publicadas en los sitios web de Internet el proceso de obtener dicha información se optimiza respecto al tiempo utilizado y esfuerzo necesario en su realización. Con el objetivo de demostrar la afirmación anterior se realizaron una serie de experimentos, los cuales se detallan a continuación:

Experimento # 1: Sin utilizar la herramienta, realizar las operaciones por el método manual.

Observando y midiendo tiempo utilizado por el personal del centro (profesores y estudiantes) en la realización de las operaciones de extracción y actualización mediante el método manual, forma en que es realizado dicho proceso actualmente, se evidencia que:

- Para la obtención de los datos en las publicaciones (sin los datos de los comentarios) el tiempo promedio que se necesita es de 50 a 90 segundos por cada una.
- Para la obtención de los datos en los comentarios el tiempo promedio que se necesita es de 25 a 40 segundos por cada uno que contenga la publicación.

Estos indicadores se obtienen considerando el tiempo al colocar estos datos en el sistema de gestión y sin tener en cuenta el tiempo de respuesta del sitio web para acceder a la publicación.

Resultados obtenidos:

Con una muestra de 7 publicaciones y 144 comentarios pertenecientes al sitio web HumanOS, se obtuvo como resultado de la puesta en práctica del experimento que se requiere un tiempo 1,7 horas aproximadamente para obtener dicha información por el personal de la empresa o entidad.

Experimento # 2: Utilizando la herramienta de recopilación desarrollada.

Mediante la utilización de la herramienta en la realización de las operaciones de recopilación y actualización sobre las publicaciones y comentarios pertenecientes al sitio web HumanOS (http://humanos.uci.cu/), se obtuvieron los siguientes resultados:

- Para la obtención de los datos en las publicaciones (sin los datos de los comentarios) el tiempo que se necesita oscila entre 0.30 y los 2 segundos por cada una.

- Para la obtención de los datos en los comentarios el tiempo promedio que se necesita varía entre 0.15 a 1.5 segundos por cada uno que contenga la publicación.

Resultados obtenidos:

Al trabajar con el sitio web de HumanOS, se obtuvo que para las operaciones de:

Recopilación sobre 7 publicaciones y 144 comentarios, el sistema necesitó 0,87 minutos. (54 segundos)

b. Actualización sobre 142 publicaciones y 2744 comentarios (de los cuales 73 son comentarios nuevos, esto conlleva sumarle el tiempo necesario para efectuar la comparación para obtener los que no están registrados en la BD del sistema) el sistema necesitó 11, 8 minutos (709 segundos)

Conclusiones

Habiendo desarrollado la herramienta para recopilar la información en las noticias publicadas en los sitios web de Internet se puede afirmar que se logró el cumplimiento del objetivo.

Mediante la investigación se determinó que el método de recopilación de información web, que mejor se adapta y resuelve las necesidades del problema, es el método de Web Scraping o Raspado Web.

Se obtuvieron 9 requisitos no funcionales, 14 requisitos funcionales para la implementación de la herramienta cumpliendo con las funcionalidades definidas.

Se demostró la reducción del tiempo en la ejecución del proceso de monitoreo y seguimiento del centro CIDI mediante la realización de una serie de experimentos.

Referencias

ANGEL, G. Computación Evolutiva (CE), Programación Genética, Evolución Gramatical, Programación por Expresión Genética *Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle*, 2008: 20-45.

GARCÍA, L. and J. MONTERO Uso de Sistemas de Gestión de Contenidos de Aprendizaje para el desarrollo del Trabajo Independiente *Referencia Pedagógica*, 2013, Vol.2(No.2): 152-165.

GARZA-RÍOS, R.; C. GONZÁLEZ-SÁNCHEZ, *et al.* Concepción de un procedimiento utilizando herramientas cuantitativas para mejorar el desempeño empresarial *Ingeniería Industrial*, 2012, 33: 239-248.

GARZÓN, T. SISTEMAS GESTORES DE BASES DE DATOS *Innovación y Experiencia Educativa*, 2010, Vol.30.

GONZÁLEZ, J. Propuesta de algoritmo de clasificación genética *RCI*, 2013, Vol. 4 (No.2): 37-42.

MAR, O.; Y. PÉREZ, *et al.* Entorno Integral de desarrollo para lenguaje en ensamblador basado en los servicios de Linux *Sociedad de la Información*, 2013, 40.

MARIA, P.; G. C. CARMEN, *et al.* Los recursos educativos electrónicos: perspectivas y herramientas de evaluación *Perspectivas em Ciência da Informação*, 2012, Vol.17(No.3): 82-99.

RIQUELME, J.; R. RUIZ, *et al.* Minería de Datos: Conceptos y Tendencias *Revista Iberoamericana de Inteligencia Artificial*, 2006, Vol.10(No.29): 11.

SANTOS, I. Modelo de gestión de información digital agraria cubana *Ciencias de la Información*, 2013, Vol. 44(Nº 2).

SCRIBID. *Manual del usuario de PostgreSQL*, 2014. [Disponible en: http://es.scribd.com/doc/5703210/Manual-del-usuario-de-PostgreSQL.

ZALDÍVAR, Y. LA CULTURA ORGANIZACIONAL Y EL LIDERAZGO EN UNA EMPRESA ORIENTADA A LA EXCELENCIA *Alternativas cubanas en Psicología*, 2014, Vol.4(No.10).