Enl@ce: Revista Venezolana de Información,
Tecnología y Conocimiento
ISSN: 1690-7515
Depósito legal pp 200402ZU1624
Año 4: No. 2, Mayo-Agosto 2007, pp. 81-95

Sistemas transparentes para gobiernos electrónicos eficientes

Andrade Castro, Jesús Alberto[1]
Yedra Hernández, Yaskelly[2]

Resumen

Este trabajo hace una reflexión crítica del papel que desempeñan los sistemas transparentes en el gobierno electrónico. La misión de los sistemas transparentes computarizados es desarrollar aplicaciones confiables y robustas, con el propósito de sustituir la fiscalización y los controles jurídicos y contables del comportamiento administrativo, por verdaderas evaluaciones que incluyan la participación del ciudadano, en el ejercicio transparente de la acción gubernamental. Teniendo como base la necesidad de tener aplicaciones para gobierno electrónico, el Laboratorio de Investigación de Tecnologías y Sistemas de Información (LITSI) de la Facultad de Ciencias de la Universidad del Zulia desarrolla aplicaciones de minería de texto, para obtener datos que están envueltos en el metalenguaje de etiquetas (HTML) contenido en las páginas WEB. Con el prototipo que hemos desarrollado, se ha hecho un intento por convertir información desde documentos tipos texto no estructurados que están en la WEB, en información factible de ser analizada y contrastada con las acciones y políticas públicas. Se pretende así, desarrollar sistemas transparentes eficientes con aplicaciones computarizadas que permitan al ciudadano ejercer el control social de la gestión gubernamental.

Palabras clave: sistemas transparentes, gobierno electrónico, metaetiquetas, eficiencia, WEB

Recibido: 27-06-07 Aceptado: 12-07-07

[1] Economista. Master of Science in Management Information Systems. Magister en Computación Aplicada. Doctor en Ciencias Humanas. Profesor Titular (Emérito). Miembro del Programa de Promoción al Investigador; Nivel II.
Correo electrónico: jandrade01@gmail.com

[2] Ingeniero en Computación. Master en Telemática. Profesora Agregada del Departamento de Computación de la Facultad de Ciencias. Universidad del Zulia. Miembro del Programa de Promoción al Investigador; Nivel Candidato. Estudiante del Doctorado de Computación de la Universidad Central de Venezuela.
Correo electrónico: yaskelly@yahoo.es

Transparent Systems for Efficient e-Government

Abstract

This Works makes a critical reflection about the role placed by transparent systems in electronic government. Transparent systems' mission is to develop trustworthy and robust applications, with the purpose of substituting fiscal, accounting and judicial controls of administrative behaviour, by real evaluations that include citizen's participation, in the transparent exercise of government action. Taking as basis the need to have applications for electronic government, Laboratorio de Investigación de Tecnologías y Sistemas de Información (LITSI) of Facultad Experimental de Ciencias at Universidad del Zulia, develops applications of text mining, in order to obtain data that are engaged in label meta-language (HTML) of web pages. With the prototype that we have developed, an attempt has been made to convert information from text-type documents that are in the web, into information susceptible to analysis and contrast with actions and public policies. Thus, the pretension in to develop efficient transparent systems with computerized applications that may allow the citizen to exercise social control over government execution.

Key words: transparent systems, electronic government, meta-labels, efficiency, WEB.

Introducción

La idea de un Estado sustentado en el gobierno electrónico (GE) ha generado esperanzas de alto impacto tanto en el sector público como en el privado, porque su instauración se asocia, por un lado, a la transparencia y control que genera confianza en los actos de la administración pública, y por el otro lado, los entes privados presuponen que para enfrentar a una administración pública ineficiente, la contribución de las tecnologías de información (TICs), particularmente la Internet, pudiera significar mejoras en los procesos internos que trasciendan hacia un Estado mucho más organizado.

Con esas dos visiones de la administración pública, el gobierno electrónico se ha manifestado en una multiplicidad de formas y a distintos niveles, con la esperanza de que su presencia contribuya a la gestión gubernamental que incluya la participación ciudadana.

Hasta ahora, en la administración pública latinoamericana, la ciudadanía no ha sido tan activa como muchos pudieran pensar, puesto que participa limitadamente en acciones que se asocian principalmente con la presentación pasiva de información, sin embargo, ella permanece latente en espera por ocupar un lugar privilegiado en el desarrollo de estrategias y políticas que permitan una mayor transparencia y control de la gestión de gobierno. Y ello ha sido así, porque los distintos niveles del gobierno que han usado la Internet, lo han hecho principalmente como medio para promocionar el reparto de los servicios, y en muchos casos, como medio propagandístico de gestión, limitando y desvirtuando el potencial que tiene la Internet como una red de alcance para el control social, donde el gran protagonista sea el ciudadano.

De manera que existe una brecha profunda entre las expectativas que tienen los ciudadanos por obtener beneficios asociados a los procesos del gobierno electrónico, y lo que hasta ahora se le ha entregado al ciudadano común.

El potencial que tienen las TICs para integrar al ciudadano al sector público en la toma de decisiones existe; por tanto, el gobierno electrónico debería usarse intensivamente como medio para la diseminación de información y como herramienta para la participación ciudadana en la toma de decisiones de los asuntos que le competen. Se puede, así, promocionar y construir una nueva ciudadanía que esté más y mejor informada acerca de los mecanismos del gobierno y el alcance de la gestión. Para ello, se requieren sistemas que agilicen la participación ciudadana en las tareas de control, y así alcanzar gestiones más transparentes. Se necesitan, por lo tanto, sistemas de información que brinden al ciudadano la posibilidad de ejercer directamente un control social de la gestión gubernamental. En ello, las tecnologías de información pueden aportar mecanismos que faciliten la participación ciudadana en el control de los recursos del Estado y en la construcción de una sociedad donde la rendición de cuentas sea exigida directamente por los ciudadanos, para incrementar y mejorar la calidad democrática.

Sistemas transparentes en el gobierno electrónico

Los valores de la acción pública en los sistemas democráticos son los de la transparencia en la actuación y el de participación social en las decisiones. Los resultados de la acción gubernamental son producto de las capacidades administrativas que el Estado desarrolla para favorecer a los diversos grupos de ciudadanos.

La participación es el valor democrático que promueve la colaboración ciudadana en la formulación y en la implantación de la acción pública (Bañón i Martínez, 2006, p xviii.). Con participación se aumenta la eficiencia de la gestión pública, porque la actuación ciudadana es dinamizadora de la acción de gobierno y de los procesos democráticos.

Pero la participación no puede reducirse a mesas técnicas de trabajo o de observación, puesto que la participación es, sobre todo, toma de decisiones, y para ello tiene que haber necesariamente una buena información, donde el ciudadano sea protagonista en la ejecución de las decisiones políticas.

La transparencia de la acción pública corresponde al conjunto de mecanismos que aseguran la igualdad de los ciudadanos y el cumplimiento de sus actividades mediante el acceso y difusión de la información. Al proveer a los usuarios con información de políticas públicas y con resultados de la gestión gubernamental, se pueden establecer vínculos entre las acciones públicas y los intereses de los ciudadanos, con el fin de regular la acción gubernamental.

El gobierno electrónico es un sistema, que como cualquier otro, genera información organizada, pero que además, se caracteriza por estar orientada a apoyar la transparencia de la gestión gubernamental, potenciando la gobernabilidad democrática al legitimar las acciones asociadas a

la transparencia y al control, que a la postre robustecen las acciones del Estado. Para Gascó (2004, p.87), el gobierno electrónico incluye todas aquellas actividades basadas en las modernas tecnologías de información y la comunicación que el Estado desarrolla para aumentar la eficiencia de la gestión pública.

La gobernabilidad se incrementa si los sistemas y procedimientos que incentivan la participación en la vigilancia y el control de la gestión pública, son abiertamente ofrecidos a los ciudadanos. Para ello, los mecanismos que aumentan la eficacia y la eficiencia se deben ofrecer abiertamente a los ciudadanos, para que sean ellos mismos quienes vigilen el desarrollo y la aplicación de las políticas públicas. Se debe por lo tanto, desarrollar programas, procedimientos y sistemas adecuados, de fácil acceso y operatividad, a fin de aumentar la transparencia que se refleje en la vigilancia y el control de los actos que conducen a las prácticas de corrupción. Y es que el GE tiene el potencial de reducir la corrupción porque puede hacer a la administración pública mucho más transparente y participativa, al trasladar parte del control de la gestión gubernamental a los ciudadanos, y ejercer, así, plenamente la transparencia.

Si se desarrollaran sistemas que permitieran al ciudadano ejercer el control social, se estaría incentivando la participación de aquellos que se ven afectados directamente por las decisiones políticas. Por lo tanto, se requiere ofrecer sistemas de información transparentes con propósitos regulatorios orientados a minimizar la corrupción e incentivar la participación ciudadana en el ejercicio de la contraloría social.

De manera que la misión de los sistemas transparentes es desarrollar aplicaciones confiables y robustas, con el propósito de sustituir la fiscalización y los controles jurídicos y contables del comportamiento administrativo, por verdaderas evaluaciones que incluyan la participación del elemento humano, en el ejercicio transparente de la acción gubernamental.

Los sistemas transparentes ponen en el ciudadano un instrumento para la acción social, que debería ser el centro de la interacción entre el ciudadano que tiene acceso a los mecanismos y sistemas digitales, con los ejecutores y responsables de la políticas públicas.

Los sistemas transparentes obligan a cumplir con las responsabilidades y a ser coherentes —y consecuentes— con nuestros compromisos y objetivos como agentes de participación social. No se trata sólo de un modelo técnico de gestión o dirección (Sarasqueta, 2004, p. 71), sino que además, existe toda una carga de compromiso personal en el sistema de transparencia informativa, que en colectivo significa el accionar de políticas públicas con racionalidad y eficiencia social. Por lo tanto, el GE debe concebirse como un sistema transparente que sirva de instrumento *regulatorio* de la acción pública.

Eficiencia de los sistemas transparentes

Los sistemas transparentes (como pueden ser los del GE) aumentan la racionalidad y la posibilidad de *controlabilidad* del ciudadano sobre las estrategias, líneas de acción y procesos de las administraciones del Estado.

El gobierno electrónico se caracteriza por generar asimetrías de información que reflejan las prioridades de las acciones públicas. Y como consecuencia de los compromisos políticos, los sistemas transparentes pudieran ser construidos en formas que fallen en el avance de las metas políticas. Tales asimetrías pueden generarse debido a que las agencias de gobierno deben manejar con discrecionalidad determinados tipos de información.

Los gobiernos siempre tienen acceso exclusivo a información que tratan en forma confidencial, generando —a veces— desconfianza en el resto de la población. Y aunque, muchos tipos de información no son (ni deberían ser) secretos, buena parte de ellos se vuelven inaccesible para la ciudadanía, sino se discrimina en forma beneficiosa para la sociedad.

Los sistemas transparentes permiten que información nueva se ajuste fácilmente en las rutinas que alteran las opciones ciudadanas. Es allí donde el GE apunta a complementar y corregir la información que es socialmente relevante. De manera que disminuyendo las asimetrías de información se abona el camino para una mejor gestión de los asuntos públicos, y en ello, el GE es una herramienta adecuada para la transparencia, porque puede disminuir las asimetrías de información que desvían el interés colectivo.

De allí que debería estar en el interés de los gobiernos incentivar el uso de sistemas transparentes que sirvan de mecanismos de control de la acción pública gubernamental.

Cuando los sistemas proveen información relevante y de fácil acceso, y los ciudadanos la incorporan en sus acciones, se produce entonces un proceso de asimilación de la información en su toma de decisiones. Si los sistemas generan y responden a las políticas de transparencias, entonces son eficientes, y ello sólo ocurre cuando la información que producen se vuelve parte o se "incrusta" en rutinas de todos los días, particularmente en aquellas asociadas a la toma de decisiones donde participan los ciudadanos.

Los sistemas de información son eficientes sólo si ellos alteran las selecciones de los usuarios en una forma que es significativa a los objetivos de la política planteada. Cuando los sistemas generan respuestas positivas a los ciudadanos, se produce la más importante condición de transparencia que es su eficiencia. Para que eso ocurra, hacen falta sistemas de información transparentes, que estén ajustados y debidamente diseñados a la medida de las necesidades y objetivos de la acción del Estado.

Así, los sistemas de transparentes tienen efectos cuando alteran la selección de información de los usuarios y se manifiestan en conductas observables que son de beneficio social. Esto quiere decir, que un sistema transparente de GE sólo es eficiente, si la conducta ciudadana es modificada en términos de los objetivos planteados por la agencia de gobierno que lo implementa. Si esas conductas responden a los objetivos planteados, entonces se está en presencia de un sistema transparente para los fines de la política pública. Los sistemas transparentes están asociados al efecto y alcance de la política sobre el propio sistema y a distintos niveles de efectividad que se pueden generar.

En el GE, los sistemas eficientes generan confianza en la ciudadanía y motivan su participa-

ción, a la vez que brindan credibilidad al abrir el abanico de opciones que las tecnologías de información pueden generar.

La acción de gobierno puede crear procesos democráticos deliberativos, a través del uso de sistemas de información transparentes que se sometan a las métricas y permitan la comparación en formatos de fácil distribución. Por ello, los sistemas transparentes introducen información nueva en patrones de tomas de decisiones complejas existentes que trascienden al beneficio colectivo. Y ello es así, porque un sistema transparente obliga a la participación ciudadana y colectivamente a dar cuentas de los actos que afectan al público, y por tanto, a estar sometidos al juicio de los demás.

Sin embargo, la necesidad de implantar sistemas transparentes en las administraciones públicas no significa que cualquier información tenga un valor en sí misma. Aunque la información esté disponible, el público pudiera desconfiar de tales sistemas, porque más información, no es necesariamente mejor, ni garantiza su distribución equitativa, y pudiera terminar por confundir a los ciudadanos y hacerlos sentir frustrados, aislados o simplemente ignorados.

Los sistemas transparentes prometen políticas socialmente innovativas, pero crean retos difíciles para los gobiernos, el sector privado y los ciudadanos. Tales sistemas tienen importancia en la política pública, porque revelan información que de no organizarse y estructurase sería difícil de difundir. Por eso, independientemente de lo relevante que resulte la información, ella no puede proveer los fundamentos para un sistema transparente a menos que esté disponible en el tiempo, en el espacio y en un formato adecuado, de manera que se ajuste a la forma que a los ciudadanos les sea útil en el proceso de toma de decisión en el conjunto de opciones que se les puede ofrecer.

La ausencia de una cultura de la evaluación y de la transparencia de la acción pública deja el camino expedito a evaluaciones arbitrarias y aleatorias. Hacen falta sistemas que sirvan de instrumentos de medida en materia de gestión y control, para alcanzar una concepción instrumental de la acción pública a medida que se produzcan resultados que sean susceptibles de medición. Afortunadamente, lo sustantivo de los sistemas transparentes es la utilidad que tienen como instrumentos de evaluación del impacto social de la acción pública.

Sistemas transparentes y extracción de información (EI) desde portales WEB

Los objetivos funcionales del gobierno electrónico usualmente impulsan el uso de la tecnología en forma desconectada de las actividades relacionadas con políticas públicas y la participación de los actores sociales. Específicamente se le asocia al aumento de eficiencia a través de la mejora de la gestión interna, a una mayor oferta de servicios y una presencia más numerosa de las tecnologías de información. Por lo tanto, se le considera un modelo "tecnológico" porque se fundamenta en el uso de tecnologías como factor determinante en el desarrollo de las prácticas organizacionales públicas. Al final, termina imperando un modelo que obstaculiza la oportunidad de incorporar a las TICs como factor de desarrollo de una sociedad cuyos cimientos son las políticas públicas. Se tra-

ta de un modelo que, bajo el manto de una visión tecnológica y tecnocrática, pretende despolitizar aspectos inherentes al comportamiento político de las acciones públicas.

Hace falta, por lo tanto, una visión del Estado distinta a aquella basada en la eficiencia técnica, que conduce al indefectible camino de construir un gobierno electrónico basado en la tecnológico, para proponer un modelo menos consumista de tecnología que tienda a resolver problemas básicos de la sociedad, sustentado en el desarrollo de sistemas de información más eficientes en términos de la participación social y la calidad de los datos.

Un sistema transparente de GE debe ser principalmente político, para que el ciudadano sea el actor social en donde las acciones públicas se centran. El GE debe entonces ser visto como la plataforma para construir un modelo de sociedad mucho más participativo en términos de políticas públicas, que refleje además, la agilidad y la transparencia de sistemas que sirvan al ciudadano como ser político, y no como un ser pasivo que es resultado de la acción técnica. El GE se extiende a lo político, no por razones asociadas a su capacidad técnica, sino porque el fundamento técnico debe expresar las razones de la política pública.

Es así como podemos entender que en los términos de eficiencia en que se entiende el GE, debe prevalecer el factor social como centro de desarrollo de cualquier expresión tecnológica. Es necesario entonces concebir sistemas transparentes que permitan al ciudadano interactuar en forma activa como controladores de las gestiones de gobierno. Para ello, hace falta construir modelos con sistemas transparentes donde la variable fundamental de trabajo sea el dato como unidad mínima de información y la expresión de su uso sea el resultado de la acción del Estado, que le es entregada al ciudadano como insumo relevante para la toma de decisiones.

Kaufman y Sebastián (2007) proponen alejarse del gobierno electrónico que se centra en la dimensión tecnológica que sólo sirve para consumir tecnología, para proponer la construcción de *un modelo de GE mínimo incremental* que permita levantar los cimientos para resolver problemas básicos en función del desarrollo de sistemas de información compartidos (con garantía de calidad de los datos). En ese modelo incremental lo principal es el dato y, por lo tanto, la información se ve como un producto y no como un subproducto del sistema, así se pone énfasis en la calidad de la información y no en los aspectos tecnológicos. Ello permite que el gobierno electrónico valide los datos como fuente de sustento de la participación ciudadana.

Teniendo como norte que el dato es la fuente primaria en la construcción del modelo incremental, se podría comenzar por enlazar los sitios WEB con la actividad de los ciudadanos. De manera que se hace necesario desarrollar y construir aspectos donde se apoye la participación ciudadana en forma mucho más operativa.

La World Wide Web (WEB) consiste principalmente de texto envuelto en un metalenguaje que por lo general corresponde a los formatos HTML o XHTML, que se despliegan en páginas en la Internet. Obtener información desde ese tipo de páginas se ha hecho vital para el manejo de datos

públicos. De manera que extraer información es una actividad central en cualquier esfuerzo que se haga para descubrir conocimiento contenido (o generado) en la WEB. Sin embargo, debido a la alta variabilidad de código HTML es muy limitante definir vínculos entre los patrones del código HTML y los conceptos que como seres humanos nos formamos.

Extraer información desde páginas WEB es un paso crucial para el desarrollo de aplicaciones bajo la técnica de minería de texto en páginas con formato HTML. Al hacer análisis de las funciones del GE encontramos que, debido a la naturaleza abierta de los datos que están contenidos en los sitios electrónicos, las páginas WEB (portales) contienen información en formato texto que no está estructurada de la forma como se conciben otros tipos de información, como son bases de datos o archivos. Y es que la información en la WEB se encuentra en forma semi-estructurada o no estructurada, y por lo tanto, se encuentra distribuida en un formato que dificulta su accesibilidad.

Extraer información de páginas WEB no puede hacerse por los procedimientos sistemáticos tradicionales de captura de datos, y ello es debido a que los sitios WEB contienen información adicional a la que es relevante a los usuarios. De manera que en los portales y demás sitios WEB existen dos tipos de datos; por un lado, aquellos que conforman la estructura de la página, que forman parte de un metalenguaje conocido como lenguaje de marcado o de marcas, que se expresa a través del uso de *metaetiquetas* y que corresponde a la manera como se codifica un documento en el lenguaje de hipertexto (*Hypertext Markup Language)* característico en la Internet; y el otro, la parte que constituye la información que sí le es relevante al usuario.

Cuando decimos extraer información (EI), nos referimos a un proceso automatizado que como entrada toma texto, que no se ve a simple vista, y produce salidas de datos estructuradas. EI se usa para localizar información en un documento que, por lo general, contiene datos expresados en un lenguaje natural, por lo tanto en forma no estructurada.

La idea detrás de la extracción de datos (ED) es desarrollar procesos que tomen como entrada, texto no visto en los navegadores (browsers), pero que están contenidos en la páginas WEB en forma de código fuente, con el fin de generar salidas con formatos fijos y no ambiguos.

Un problema que se presenta cuando se navega con browsers en la internet es que, la forma de recolectar datos relevantes, está basada en métodos poco automatizados y eso en grandes volúmenes de datos es inadecuado. La automatización es deseable para altos volúmenes de datos y para casos donde las personas no están capacitadas en actividades de computación. Para ese tipo de situaciones es deseable acceder a datos desde programas computarizados que les permita interactuar y recolectar datos desde estructuras menos conocidas.

Hay, ciertamente, muchos documentos en la WEB que son dirigidos principalmente para presentar algunos datos estructurados, tales como listas (precios, artículos, etc.), tablas (horarios, cruce de datos asociados, etc.) y otras formas es-

tructuradas. Tales documentos se denominan *datos intensivos*, y son generados automáticamente desde el *back-end* de un sistema de base de datos. La información, en este tipo de documentos, usualmente es presentada en una forma clara y estructurada, de manera que el usuario puede encontrar la información deseada con poco esfuerzo (Burget, 2004).

Usualmente, este tipo de documento contiene una estructura jerárquica de encabezamiento y etiquetas de navegación que denotan el significado de cada parte del texto o el valor de los datos tratados, que permite al usuario ir desde el dato más general (ejemplo, desde el encabezamiento principal, que da una idea del tópico del documento) a una forma mucho más específica, a fin de alcanzar un valor del documento. Este tipo de jerarquía es llamada estructura lógica del documento.

Contrario a lo que ocurre con el código en HTML (metalenguaje) donde el usuario de documentos es limitado por las capacidades del WEB browser, la información no estructurada —que es relevante— debe ser buscada e interpretada por el usuario. Esto tiene diversos problemas comunes que son causados por la gran variabilidad de HTML y porque las construcciones no tienen relación directa con los datos semánticos.

Hacen falta métodos, técnicas y herramientas adecuadas para manejar sitios WEB, con el propósito de generar tipos de información a partir de datos que se encuentran contenidos en los documentos con formato HTML. Sin estas técnicas, asociadas a la extracción de datos, sería muy difícil obtener información que se despliega en la Internet.

Un modelo de este tipo de jerarquía es denominado "estructura lógica de documentos" (Summers, 1995). Diversos enfoques se han propuesto con el fin de descubrir estructuras lógicas en documentos de tipo HTML (Gu, Chen, Ma, Chen, 2002 y Kahn 2001). Con código HTML las limitaciones del manejo de datos estás dadas por las limitaciones del navegador, las cuales se someten a las jerarquías propias del lenguaje.

Estos procesos automatizados están basados en algoritmos y programas que son desarrollados bajo un enfoque de minería de texto, una variante de la minería de datos. En particular, la minería de textos permite explorar datos en la WEB para descubrir patrones desconocidos o para generar información con significado para algún tipo de usuario en particular. La accesibilidad y abundancia de información en los portales WEB hace del uso y desarrollo de la minería de datos un asunto de considerable importancia y necesidad. Los beneficios del uso de esta metodología incluye el mejoramiento en el manejo de grandes volúmenes de datos y la obtención de resultados más claros para propósitos definidos.

Un elemento que le añade dificultad, a la generación de información a partir de las páginas WEB, está relacionado con el diseño. Las páginas WEB usualmente tienen una inmensa variedad de diseños, de manera que de antemano no está claro si sería posible realizar una extracción de datos en forma sistemática y si los datos extraídos serán de utilidad para el procesamiento y generación de información, que sea útil para los propósitos y requerimientos funcionales que se definen.

Estudios relacionados con la extracción automática han sido realizados por Liu, Grossman,y Zhai (2003) quienes han propuesto un método de extracción de registros de datos en las páginas WEB. Reis, Golgher, Silva, y Laender (2004) investigaron acerca de la extracción de artículos de noticias. Craven (2003) propuso un método de extracción de resúmenes desde las páginas WEB.

Un desafío crucial en la extracción de información, como tecnología de aplicación en la WEB, es la adquisición de experticia. Las técnicas para extraer información del dominio son todavía muy débiles y en particular en la WEB estas técnicas están afectadas por la forma cómo se organizan los distintos tipos de documentos y por el tiempo que los expertos deben involucrarse en aportar su conocimiento. Jung; Yi; Kim y Lee (2005) propusieron estrategias para extraer información de los expertos y generar conocimiento automático a partir de documentos estructurados de la WEB. Su enfoque está dirigido a documentos estructurados, por lo tanto deja por fuera una vasta cantidad de documentos WEB.

El trabajo de Xue, Hu, Xin, Song, Shi, Cao, Lin y Li (2007) extrae datos en forma automática desde los títulos contenidos en el cuerpo de los documentos HTML publicados en la WEB. Ellos desarrollaron un método para extraer automáticamente los títulos bajo ciertas condiciones de tamaño de la letra, color, estilo, alineación, número de títulos contenidos, líneas de texto consecutivas, etc. En todas esas situaciones, el problema se centra en el diseño de la página, porque los títulos pueden estar distribuidos en diversos lugares de la página o incluso en diverso sentido (horizontal o vertical).

Otros estudios refieren a la extracción en los niveles de estructuras de datos, por ejemplo Breuel (2003) propuso un análisis sintáctico (*parsing*) de la página WEB, formando árboles de etiquetas en HTML. Song, Liu, Wen y Ma (2004) han propuesto dividir las páginas WEB en bloques, para luego extraer información desde esos bloques.

Lee, Seo, Lee, Jung, Cho, Lee, Kwak, Cha, Kim, Ahn, Kim y Kim (2001) desarrollaron la idea de recuperar respuestas en lugar de documentos, a través de procedimientos centrados en un sistema de preguntas y respuestas, a partir de respuestas "tipos" y así seleccionar la respuesta por cada respuesta "tipo". El enfoque en este caso consistió en clasificar posibles respuestas y diseñar un método para determinar los tipos de respuestas.

Una técnica desarrollada por Shim, Kim, Cha, Lee, y Seo (2002) consistió en hacer un análisis de pre procesamiento morfológico, en donde un pre-procesador remueve la mayoría de las etiquetas pertenecientes al metalenguaje HTML en un documento en página WEB, con excepción de las etiquetas **<title>** y **<keyword>** que son usadas posteriormente para propósitos específicos. El pre procesador mantiene el diseño de las tablas y determina las fronteras del cuerpo del documento. Todos los procesos después de este pre procesamiento son ejecutados en documentos HTML con sus etiquetas removidas, constituyéndose en un documento casi en formato texto simple. Luego, un analizador morfológico analiza las

sentencias en el lenguaje Koreano. Cada *eojeol*[3] en una sentencia, produce pares de morfemas con cierta parte de las etiquetas correspondientes al nuevo texto. El analizador morfológico hace una post edición del análisis y recupera el morfema de la secuencia incorrecta usando una base de datos de errores.

Propuesta de aplicación de minería de texto en el GE

En todos los trabajos anteriormente referenciados, el análisis circunda en torno a la extracción de información desde páginas en la WEB para transformar el contenido de la entrada de documentos en datos estructurados.

El uso generalizado de la Web ha convertido a HTML en un estándar de *facto* para intercambiar documentos. HTML es una simplificación de SGML, un lenguaje de especificación de texto estructurado diseñado originalmente con el objetivo de que fuera un lenguaje universal para intercambiar y manipular texto estructurado. Es bastante posible que XML reemplace a HTML en el futuro, y se hacen esfuerzos para estandarizarlo. La estructura que se puede derivar de un texto en ningún caso es similar a la relacional (como las de las bases de datos relacionales), que se puede separar en campos y registros fijos y tabulada.

Para los propósitos del gobierno electrónico, las técnicas de extracción de información brinda al ciudadano con sistemas transparentes que le permiten ejercer un mayor control de la ejecución gubernamental. Para que las tecnologías de información sean utilizadas sistemáticamente, hace falta la institucionalización de lineamientos básicos generales que sirvan de referencia para la adopción de sistemas transparentes. Debido a que cada instancia de gobierno aspira insertar o modificar el uso de TICs para incrementar su eficiencia, hacen falta lineamientos rectores que permitan la organicidad de las políticas del Estado.

En el Laboratorio de Investigación de Tecnologías y Sistemas de Información (LITSI) de la Facultad de Ciencias de la Universidad del Zulia estamos trabajando en el desarrollo de sistemas transparentes, que permitan obtener datos desde la WEB que cumplan patrones pre establecidos. Un prototipo ya desarrollado, permite buscar, dentro del metalenguaje, información que le es relevante a los usuarios (nuestro interés es a los ciudadanos).

Con el prototipo que hemos desarrollado, se ha hecho un intento por convertir información desde documentos tipos texto que están en la WEB (portales de GE) en información que es vertida como entradas en bases de datos relacionales (poblar la base de datos), para ser luego analizadas y contrastada con las acciones y políticas públicas.

Uno de los principales retos que se nos presentó cuando desarrollábamos aplicaciones para extraer información, fue el de ser consistente con el contenido de la página y el manejo del metalenguaje incrustado. A nivel del prototipo, desarrolla-

[3] Frases segmentadas y palabras en Koreano que se transforman en un espacio.

mos varios métodos empíricos relacionados con el proceso de extracción de información; por ejemplo, un enfoque predominante consistió en anotar manualmente una recopilación grande de datos extraídos en forma indirecta (en nuestro caso, portales WEB de gobierno electrónico de Venezuela) que sirvieron de pista para formalizar futuras búsquedas. Esto se hizo a través de un procedimiento que sirvió de aprendizaje para construir patrones de extracción desde el corpus de texto anotado. Este procedimiento fue anteriormente propuesto por Nahm, y Mooney (2000). También, considerando la propuesta de Yangarber y Grishman (2000) redujimos las anotaciones manuales aplicando directamente técnicas de aprendizaje con datos no anotados, luego de obtener indicaciones de lo que interesaba para la captura de patrones regulares de información.

Nuestra meta es desarrollar un sistema WEB de extracción de información altamente portable, sin anotación manual, proponiendo al menos dos ideas claves. Primero, desarrollar el sistema en un software universal no privativo y de libre acceso, XML y SGML para aplicar extracciones en documentos WEB basados en el estándar HTML, ello a través de una sintaxis expresada en una definición tipo documento modificado (Document Type Definition —mDTD—), desarrollado por Kim, Jung, Lee (2003) el cual depende de una interpretación analítica para identificar el objetivo de extracción desde el contenido del documento WEB. Y segundo, desde un documento DTD convencional pretendemos dos cosas, a) introducir un modelo con palabras clave y operadores que corresponden a los datos objetivos, y b) construir una información relevante para el desarrollo de una controlaría social a partir de una métrica predefinida que sirva para interpretar los datos. El procedimiento involucra el desarrollo de reglas en mDTD que permiten establecer un dominio de trabajo. Esto se logra a partir la extracción de documentos estructurados de la WEB sin ayuda manual.

Nosotros, hasta ahora, hemos contribuido al área de búsqueda en texto no estructurado a fin de poblar bases de datos (información estructurada). Planeamos continuar trabajando en lenguajes tipo XML, desarrollando prototipos para consultar datos, en páginas elaboradas en otros lenguajes como el XML. La capacidad de consultar eficientemente XML (y HTML como un caso simplificado) abrirá la puerta a mejoras de las máquinas de búsqueda en portales Web, tales como el de incorporar predicados sobre la estructura de los documentos. Así mismo, pretendemos hacer un análisis de grafo que nos indique el grado de profundidad que un sitio WEB genera.

Resultados preliminares

Los sitios WEB contienen etiquetas que dan significado a la manera cómo se despliega la información; adicionalmente, las páginas contienen datos que son considerados errores o información no deseada al corpus del texto tratado, ello dificulta la recuperación de información relevante.

Este alto número de errores existentes en los documentos desplegados de Internet, en la mayoría de los casos por no seguir los estándares, dificulta el tratamiento informático; por ello, estamos desarrollando técnicas de depuración de texto que previamente procesen y reparen las páginas web,

a fin de obtener documentos XHTML (eXtensible HTML) bien establecidos.

Las pruebas realizadas, aunque preliminares, muestran el gran potencial del método propuesto para encontrar información a partir de datos embebidos en metalenguajes, al usar la Web como corpus del texto, así como la viabilidad de la incorporación de nuevo conocimiento en repositorios y sistemas que intentan disminuir la ambigüedad del sentido de las palabras, que pueden a su vez ser usadas en sistemas de recuperación de información.

Para el ciudadano que interactúa a través de la internet, un sistema de información transparente, basada en calidad de los datos, le permitirá ejecutar políticas públicas a través del uso de métricas aplicadas a los portales de gobierno electrónico. Por ejemplo, cualquier ciudadano podría hacer seguimiento de la ejecución de obras, desde el momento en que se licite, vigilando el proceso de desarrollo que tiene la obra en un momento dado. Sin embargo, el sistema por sí mismo no garantiza que la información en el portal sea válida o que esté correctamente publicada. Lo que el sistema se limita a hacer es tomar la información publicada en el portal, para establecer la presencia y cumplimiento de ciertos indicadores previamente diseñados.

Conclusiones

Con el gobierno electrónico se pretende brindar a la ciudadanía un mayor y mejor acceso a los mecanismos de decisión, en temas que la afecten directamente. La idea es construir una administración pública más productiva, transparente y democrática, que facilite la eficiencia en el servicio público y modifique la relación tradicional entre el ciudadano y el Estado, a través de la existencia de sistemas transparentes.

Los resultados se reflejan en la modernización del Estado como entidad jurídico-administrativa, y se manifiesta en la incorporación de nuevos espacios de participación en la toma de decisiones y como apoyo para la contraloría social.

La participación ciudadana es vital para la el desarrollo del gobierno electrónico, pero hay que adaptarlo a la aplicación de políticas públicas; por lo tanto, estas deben prevalecer y reflejarse en el portal, para que el GE no sea visto como un elemento "adicional" de la gestión de gobierno, sino que se pueda considerar como un mecanismo útil y necesario para la gobernanza.

Desde un punto de vista práctico, la transparencia y la confianza en el GE se amarra al potencial que brinda el uso de las tecnologías de información y los sistemas transparentes, para que la ciudadanía ejerza su participación plena, con el fin de re estructurar la acción de la democracia.

La internet es el espacio donde el GE adquiere sentido, pero para los efectos de la participación y control hace falta que se desarrollen sistemas transparentes. En el Laboratorio de investigación de tecnologías y sistemas de información LITSI estamos desarrollando sistemas que permitan la operatividad de la acción política, a través del uso de sistemas computarizados dirigidos a ejercer el control social de la gestión pública.

Es necesario que el ciudadano común se instruya en el uso de herramientas mínimas para participar activamente en los procesos políticos, económicos y sociales que viven las sociedades. La ciudadanía debe impulsar la incorporación de sistemas transparentes en los asuntos públicos, de tal modo que se materialice el principio de la democracia participativa y protagónica, en donde sea la ciudadanía organizada quien desde sus niveles, cualidades y capacidades puedan decidir y ejecutar las acciones del gobierno electrónico.

Bibliografía

Bañón i Martínez, R. (2006). (Compilador). *La evaluación de la acción y de las políticas públicas*. España: Ediciones Díaz de Santos.

Breuel, T. (2003). Information extraction from HTML documents by structural matching. En P*roceedings of the second internacional workshop on web document analysis*.

Burget, R. (2004). Hierarchies in HTML Documents: Linking Text to Concepts. En *Proceedings of the Database and Expert Systems Applications, 15th international Workshop on (Dexa'04) - Volume 00* (August 30 - September 03, 2004). DEXA. IEEE Computer Society, Washington, DC, 186-190. DOI= http://dx.doi.org/10.1109/DEXA.2004.80. Recuperado el 12 de marzo de 2007 del sitio WEB: http://citeseer.ist.psu.edu/cache/papers/cs2/492/http:zSzzSzwww.fit.vutbr.czzSz~burgetrzSzpublicationszSzwebs2004.pdf/burget04hierarchies.pdf

Craven, T. (2003). HTML tags as extraction cues for web page description construction. *Informing Science Journal*, pp. 6, 1-12.

Gascó, M. (2004). E-gobienro en Bolivia y Paraguay. En *América Latina Puntogob. Casos y Tendencias en Gobierno Electrónico,* Coordinador Araya Dujisin; Porrúa Vigón.

Gu X.; Chen J; Ma W. y Chen G. (2002). Visual Based Content Understanding towards Web Adaptation, *Proc. Adaptive Hypermedia and Adaptive Web-Based Systems*, Malaga, Spain, pp. 164-173

Jung H.; Yi, E.; Kim, D. y Lee, G. (2005). Information extraction with automatic knowledge expansion. *Information Processing and Management* 41, pp. 217–242

Lee, G.; Seo, J.; Lee, S.; Jung, H.; Cho, B.; Lee, C.; Kwak, B.; Cha, J.; Kim, D.; Ahn, J.; Kim, H. y Kim, K. (2001). SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP. En *Proceedings of the 10th text retrieval conference.*

Liu, B.; Grossman, R. y Zhai, Y. (2003). Mining data records in web pages. En *Proceedings of the ninth ACM SIGKDD internacional conference on knowledge discovery and data mining* (pp. 601-606).

Nahm, U. y Mooney, R. (2000). Using information extraction to aid the discovery of prediction rules from text. En *Proceedings of the ACM SIGKDD-2000 workshop on text mining*.

Sarasqueta, A. (2004). *Una visión global de la globalización*. Espaýa: EUNSA.

Reis, D.; Golgher, P.; Silva, A. y Laender, A. (2004). Automatic web news extraction using tree edit distance. En *Proceedings of international WWW conference (*pp. 502–511).

Shim, J.; Kim, D.; Cha, J.; Lee, G. y Seo, J. (2002). Multi-strategic integrated Web document pre-processing for sentence and word boundarydetection. *Information Processing and Management*, 38(4).

Song, R.; Liu, H.; Wen, J.-R. y Ma, W.-Y. (2004). Learning block importance models for web pages. En *Proceedings of internacional WWW conference* (pp. 203-211).

Summers, K. (1995). Toward a taxonomy of logical document structures. *Electronic Publishing and the Information Superhighway: Proceedings of the Dartmouth Institute for Advanced Graduate Studies* (DAGS '95). Boston, USA.

Xue, Y.; Hu, Y.; Xin, G.; Song, R.; Shi, S.; Cao, Y.; Lin, C. y Li, H. (2007). Web page title extraction and its application. *Information Processing and Management*. 43 (2007) pp. 1332-1347

Yangarber, R. y Grishman, R. (2000). Extraction pattern discovery through corpus analysis. En: *Proceedings of the conference on applied natural language processing ANLP-NAACL.*