Una metodología para mejorar el contenido de un sitio web a partir de la identificación de sus web site keywords

José I. Fernández*
Juan D. Velásquez*

Resumen

Presentamos una metodología para identificar aproximadamente qué palabras atraen la atención del usuario cuando se encuentra visitando páginas de un sitio web. Estas palabras son llamadas "web site keywords" y pueden ser usadas para la creación de contenidos relacionados a un tópico específico con el que se pretende atraer la atención del usuario.

A través de la utilización de las palabras correctas, se puede mejorar gradualmente el contenido de un sitio web, ayudando de esta forma a los usuarios a encontrar la información que buscan, lo cual se considera clave para el éxito y continuidad del sitio.

Aplicando algoritmos de clustering, y asumiendo que existe una correlación entre el tiempo invertido en una página y el interés del usuario, se realiza una segmentación de los usuarios según comportamiento de navegación y preferencias de contenidos. A continuación, se identifican las palabras clave del sitio web. Esta metodología fue aplicada en datos originada desde un sitio web real, mostrando su efectividad.

Palabras Clave: Web site keywords, Clustering, Comportamiento del usuario web.

*Departamento de Ingeniería Industrial, Universidad de Chile

1. Introducción

Para muchas compañías y/o instituciones, ya no es suficiente el desarrollo de un sitio web para ofrecer sus productos y servicios en el mercado digital. Lo que a menudo hace la diferencia entre un éxito o fracaso en un e-business es el potencial del sitio web para atraer o retener usuarios. Este potencial depende del contenido del sitio, diseño y aspectos técnicos como tiempo de descarga de páginas del sitio hacia navegador del usuario, entre otros. En términos de contenido, las palabras usadas en el texto libre en las páginas de un sitio web son muy importantes, por cuanto dicen relación con la información que los usuarios buscan. En efecto, la gran mayoría de lo usuarios recurre a motores de búsqueda, tales como Yahoo! y Google, para realizar consultas respecto de un contenido de su interés, a través de consultas basadas en términos en motores de búsqueda para encontrar información en la Web. Estas consultas son realizadas usando palabras clave, es decir, una palabra o grupo de palabras [14] que caracterizan el contenido de un página web dada o un sitio web.

El correcto uso de las palabras con que se crea el contenido textual de una página web, mejora la información presentada a los usuarios, ayuda a la búsqueda efectiva de información, mientras atrae a nuevos usuarios y retiene a los actuales, mediante actualizaciones continuas del contenido textual de la página. El desafío, entonces, es identificar qué palabras son importantes para los usuarios. Lo anterior tiende a relacionarse con cual es "palabra más frecuentemente usada". Algunas herramientas comerciales[1] ayudan a identificar palabras clave objetivo que los consumidores son más propensos a utilizar mientras realizan sus búsquedas en la Web [6].

Mediante la identificación de las palabras más relevantes en las páginas de los sitios, desde el punto de vista del usuario, las mejoras pueden ser realizadas en el sitio web completo. Por ejemplo, el sitio puede ser reestructurado colocando un nuevo hyperlink relacionado con la palabra clave y por supuesto el contenido textual podría ser modificado utilizando las palabras clave relacionadas con un tópico específico para enriquecer el texto libre en una página Web.

En este trabajo se presenta una metodología para analizar el comportamiento de navegación del usuario y sus preferencias de contenido a través de la aplicación de algoritmos de web mining en datos originados en la web, también llamados web data, específicamente registros de un sitio web (web logs) y su contenido textual.

La metodología apunta a identificar aproximadamente cuales palabras atraen

[1] Ver por ejemplo http://www.goodkeywords.com

la atención del usuario cuando esta visitando páginas en un sitio web. Estas palabras son denominadas "palabras clave de un sitio web" [31] y pueden ser utilizadas para la creación de contenidos de texto mejoradas relacionadas con tópicos específicos.

Este paper esta organizado de la siguiente forma: La sección 2 introduce una revisión breve acerca del trabajo relacionado. El proceso de preparación para transformar la web data en vectores de características para ser utilizados como entrada en los algoritmos de web mining es mostrado en la sección 3. En la sección 4, la metodología para identificar las palabras clave de un sitio web es explicada y aplicada en la sección 5. Finalmente, la sección 6 muestra las conclusiones principales de este paper.

2. Trabajos Previos

Cuando un usuario visita un sitio web, datos respecto de qué página visitó son almacenados en archivos de registro llamados web logs. Entonces es directo conocer cuáles páginas son visitadas y cuáles no, e inclusive el tiempo gastado por el usuario en cada una de ellas. Debido a que usualmente las páginas contienen datos acerca de un tópico específico, es posible conocer aproximadamente las preferencias de información de los usuarios. En ese sentido la interacción entre el usuario y el sitio es como una indagación electrónica, entregando los datos necesarios para analizar las preferencias de contenido del usuario en un sitio web particular.

El desafío para analizar las preferencias de texto del usuario en el sitio web es doble. Primero la cantidad de registros en el archivo web log usualmente es enorme, y una parte son datos irrelevantes acerca del comportamiento de navegación del usuario en el sitio. Segundo, el texto libre dentro de las páginas web es comúnmente plano, es decir, sin información adicional que permita conocer directamente cuáles son las palabras que atraen la atención del usuario.

En esta sección se revisan las principales aproximaciones para analizar las web data para extraer patrones significativos relacionados con las preferencias de texto de los usuarios en el sitio web.

2.1. Minando los web data

Las técnicas de web mining emergieron como resultado de la aplicación de teoría de data mining al descubrimiento de patrones desde los web data [8, 16, 25]. El web mining no es una tarea trivial considerando que la Web es una enorme colección de información heterogénea, no clasificada, distribuida, variante en el tiempo, semi estructurada y altamente dimensional. El web mining debe considerar tres importantes pasos: Preprocesamiento, descubrimiento de

patrones y análisis de patrones [27].

Las siguientes terminologías comunes son utilizadas para definir los diferentes tipos de web data.

- Contenido. El contenido de la página web, es decir, imágenes, texto libre, sonidos, etc.
- Estructura. Información que muestra la estructura interna de una página web. En general, tienen etiquetas HTML o XML, alguna de las cuales contienen información acerca de hipervínculos con otras páginas web.
- Uso. Información que describe las preferencias del visitante mientras navega en un sitio web. Es posible encontrar esta información dentro de los archivos web log.
- Perfil del usuario. Colección de información acerca del usuario: Información personal (nombre, edad, etc.), información de uso (por ejemplo, páginas visitadas) e intereses.

Con las definiciones anteriores, y dependiendo de los web data a procesar, las técnicas de web mining pueden ser agrupadas en tres áreas: Minado de contenido web (WCM o Web Content Mining), Minado de la estructura web (WSM o Web Structure Mining), y Minado de la utilización de la web (WUM o Web Usage Mining).

2.1.1. Identificando palabras para la creación de un resumen automático de texto de una página web

La meta es construir automáticamente resúmenes de lenguaje natural de documento [11]. En este caso, una semi estructura relativa es creada por la aplicación de etiquetas HTML desde el contenido textual de una página web, la cual examina temas sin restricción de dominio. En muchos casos, las páginas pueden solamente contener pocas palabras sin elementos textuales (por ejemplo video, imágenes, audio, etc.) [1].

En la investigación de resumen de texto, tres importantes aproximaciones son [18]: basadas en parágrafos, basadas en oraciones y utilización de señales de lenguaje natural en texto.

La primera aproximación consiste en seleccionar un párrafo de un segmento de texto [19] que apunta a un tema en el documento, bajo la suposición que hay varios temas en el texto. La aplicación de esta técnica en una página web no es obvia; los diseñadores web tienen la tendencia de estructurar el texto en párrafos por página. Por lo tanto un documento contiene un solo tema, lo cual hace la aplicación de esta técnica dificultosa.

En la segunda aproximación, las frases más interesantes o frases clave son extraídas y ensambladas en un texto individual [9,37]. Es claro que el texto

resultante puede no ser cohesivo, pero la meta de la técnica es proveer la máxima expresión de información en el documento. Esta técnica es aplicable para páginas web, dado que la entrada puede consistir de pequeñas piezas de texto [6]. La aproximación final es un modelo de discurso basado en la extracción y resumen [14,15] mediante la utilización de señales de lenguaje natural como identificación de nombres propios, sinónimos, frases claves, etc. Este método arma oraciones mediante la creación de una colección de texto con información del documento completo. Esta técnica es más apropiada para documentos dentro de un dominio específico y esto para la implementación en un sitio web es dificultoso.

2.2. Extracción de texto de páginas web y aplicaciones

Las componentes de texto clave son partes de un documento completo, por ejemplo un párrafo, frase y una palabra que contiene información significativa acerca de un tema particular, desde el punto de vista del usuario del sitio web. La identificación de estos componentes puede ser útil para mejorar el contenido textual de un sitio web.

Usualmente, las palabras clave en un sitio web están correlacionadas con las "palabras más frecuentemente utilizadas". En [6], se introduce un método para la extracción de las palabras clave desde un gran conjunto de páginas web. La técnica está basada en la asignación de importancia a las palabras, dependiendo de su frecuencia en todos los documentos. Seguidamente, los párrafos o frases que contienen las palabras clave son extraídos y su importancia es validada a través de pruebas con usuarios reales.

Otro método, en [2], recolecta palabras clave desde un motor de búsqueda. Esto muestra las preferencias globales de palabras de una comunidad web, pero no brinda detalles acerca de un sitio web particular.

Finalmente, en lugar de analizar palabras, en [17] se desarrolla una técnica para extraer conceptos desde el texto de una página web. Los conceptos describen objetos del mundo real, eventos, pensamientos, opiniones e ideas en una estructura simple, como términos descriptivos. Entonces, utilizando el modelo de vector espacial, los conceptos son transformados en vectores de características, permitiendo la aplicación de algoritmos de clustering o clasificación a páginas web.

3. Proceso de preparación de la Web Data

De toda la información web disponible, la más relevante para el análisis del comportamiento y preferencias de navegación del usuario, son los registros (web logs) y las paginas web [33]. Los web logs contienen información acerca

de la secuencia de navegación de páginas y el tiempo gastado en cada página visitada, aplicando el proceso de sesionización. La fuente de la página web es el sitio web en si mismo. Cada página web es definida por su contenido, en particular texto libre. Para estudiar el comportamiento del usuario ambas fuentes - web logs y páginas web - se preparan mediante la utilización de filtros y por la estimación de sesiones reales de usuario. La etapa de preprocesamiento implica, primero, un proceso de limpieza y, segundo, la creación de vectores de características como entrada a los algoritmos de web mining, dentro de la estructura definida por los patrones vistos.

3.1. El proceso de reconstrucción de sesiones

El proceso de segmentación de las actividades de usuarios en sesiones individuales es llamado ***sesionización*** [10] y está basado en los web logs del sitio web. En consideración de los inconvenientes mencionados anteriormente, el proceso no esta libre de errores [26]. La sesionización asume que la sesión tiene un tiempo de duración máximo y que no es posible saber si el visitante ha presionado el botón "volver" (back) en el navegador del sitio web. Si la página esta en el cache del navegador y el visitante vuelve a ella en la misma sesión, podría no quedar registrada en los logs del sitio web. Por esto han sido propuestos el uso de esquemas invasivos como el envío de otra aplicación al browser para capturar el comportamiento de navegación exacto del usuario [3, 10]. Si embargo, este esquema podría ser fácilmente evitado por el visitante.

Muchos autores [3, 10, 20] han propuesto la utilización de heurísticas para la reconstrucción de sesiones por los web logs. En esencia, la idea es crear subgrupos con las visitas de usuarios y aplicando mecanismos sobre los web logs generados para permitir la definición de una sesión como series de eventos entrelazados durante un cierto periodo de tiempo.

La reconstrucción de sesiones apunta a encontrar sesiones de usuarios reales, es decir, cuales páginas fueron visitadas por un ser humano. En ese sentido, cualquiera sea la estrategia utilizada para descubrir las sesiones reales, esta debe satisfacer dos criterios esenciales: las actividades realizadas por una persona real pueden ser agrupadas entre si y el conjunto en actividades que pertenecen a la misma visita (otros objetos requeridos por la página web visitada) también pertenecen al mismo grupo.

Hay varias técnicas para sesionización, las cuales pueden ser agrupadas en dos estrategias mayores: *proactiva y reactiva* [26].

Las **Estrategias Proactivas** intentan identificar el usuario utilizando métodos de identificación como cookies que consisten en una pieza de código asociado al sitio web. Cuando el visitante ingresa al sitio por primera vez, una cookie es enviada al navegador. Luego, cuando la página es revisitada, el navegador muestra el contenido de la cookie al servidor web y automática-

mente la identificación toma lugar. El método tiene problemas desde el punto de vista técnico y también con respecto a la privacidad del usuario. Primero, si el sitio es revisitado después de varias horas, la sesión será considerada muy larga, y será entonces una nueva sesión. En segundo lugar, algunos aspectos de las cookies parecen incompatibles con los principios de protección de datos de algunas comunidades, como la Unión Europea [26]. Finalmente, las cookies pueden ser fácilmente detectadas y desactivadas pro el visitante.

Las Estrategias Reactivas son no invasivas con respecto a la privacidad y hacen uso de la información contenida sólo en los web logs y consiste en el procesamiento de los registros para generar un grupo de sesiones reconstruidas.

En el análisis del sitio web, el escenario general es que los sitios web usualmente no implementan mecanismos de identificación. La utilización de estrategias reactivas puede llegar a ser más útil. Estas pueden ser clasificadas en dos grupos principales [4, 10]:

- Heurísticas orientadas a la navegación: asumen que el visitante llega a páginas a través de hyperlinks desde otras páginas. Si el requerimiento de una página es inalcanzable a través de las páginas previamente visitadas por el usuario, una nueva sesión es iniciada.

- Heurísticas Orientadas al tiempo: se coloca un tiempo máximo de duración, que es usualmente 30 minutos para la sesión completa [7]. Basado en este valor se pueden identificar las transacciones pertenecientes a una sesión específica utilizando filtros programados.

3.1.1. Procesando el contenido textual de una página web

Hay varios métodos para comparar el contenido de dos páginas web, aquí se considera el texto libre dentro de las páginas web. El proceso común es coincidir los términos que componen el texto libre, por ejemplo, mediante la aplicación de un proceso de comparación de palabras. Un análisis más complejo incluye información semántica contenida en el texto libre que involucra también una tarea de aproximación de términos comparados.

La información semántica es fácil de extraer cuando el documento incluye información adicional acerca del contenido del texto, por ejemplo, etiquetas de marcado. Algunas páginas web permiten la comparación de documentos mediante la información estructural contenida en las etiquetas HTML, incluso con restricciones. Este método es utilizado en [28] para comparar páginas escritas en lenguajes diferentes con una estructura HTML similar. La comparación es enriquecida por la aplicación de un proceso de equiparar el contenido textual [29], el cual considera una tarea inicial de traducción a ser completada. El método es altamente efectivo cuando el lenguaje utilizado es el mismo en las páginas que se encuentran en comparación. Una breve encuesta de algorit-

mos para comparar documentos por la utilización de estructuras similares es encontrada en [5].

Las comparaciones son realizadas por una función que retorna un valor numérico mostrando similitudes o diferencias entre dos páginas web. Esta función puede ser utilizada en algoritmos de web mining para procesar un conjunto de páginas web, las cuales pueden pertenecer a una comunidad web o un sitio web aislado. El método de comparación debe considerar un criterio de eficiencia en el procesamiento de contenido de páginas web [13]. Aquí el modelo de vector espacial [24], permite una representación vectorial simple de las páginas web y. mediante el uso de comparación de distancia entre vectores, provee de una medida de las diferencias y similitudes entre páginas.

Las páginas web deben ser limpiadas antes de transformarlas en vectores, tanto para reducir el número de palabras - no todas las palabras tienen el mismo peso - y hacer el proceso más eficiente. Por esto, el proceso debe considerar los siguientes tipos de palabras:

- Etiquetas HTML: En general, estas deben ser limpiadas. Sin embargo, la información contenida en cada etiqueta puede ser utilizada para identificar palabras importantes en el contexto de una página. Por ejemplo, la etiqueta "<titulo>" enmarca el tema central de la página web, es decir, de una noción aproximada del significado semántico de la palabra y, es incluida en la representación vectorial de la página.

- Palabras de detención. (por ejemplo pronombres, preposiciones, conjunciones, etc.).

- Stem de palabras. Después de aplicar el proceso de remoción del sufijo de la palabra (stemización de la palabras [22]), obtenemos la raíz de la palabra o stem.

Para el propósito de representación vectorial, sea R el número total de palabras diferentes y Q el número de páginas en el sitio web. Una representación vectorial del conjunto de páginas es una matriz M de tamaño R x Q.

$$M = (m_{ij}), \text{ con } i = 1, \ldots, R \text{ y } j = 1, \ldots, Q \tag{1}$$

Donde m_{ij} es el peso de la palabra i en la página j.
Basado en *tfxidf-weighting* introducido en [24] los pesos son estimados como:

$$m_{ij} = f_{ij}(1 + sw_i)\log(\frac{Q}{n_i}) \tag{2}$$

Aquí, f_{ij} es el número de ocurrencias de la palabra i en la página j y n_i es el número total de documentos del sitio web que contienen la palabra i.

Adicionalmente, la importancia de las palabras es incrementada por la identificación de palabras especiales, las cuales correspondiente a los términos en la página web que son más importantes que otras, por ejemplo, palabras destacadas (haciendo uso de etiquetas HTML), palabras utilizadas por el usuario en la búsqueda de información y, en general, palabras que implican los deseos y necesidades de los usuarios. La importancia de palabras especiales es almacenada en un arreglo sw de dimensión R, donde sw_i representa un peso adicional para la i-ésima palabra.

El arreglo sw permite al modelo de vector espacial incluir ideas acerca de información semántica contenida en el texto de la página web por la identificación de palabras especiales.

Las fuentes comunes de palabras especiales son:

1. E-Mails: El ofrecimiento de envío de emails por parte del usuario para la plataforma de call center. Este texto enviado es una fuente para identificar las palabras más recurrentes. Sea $ew_i = \frac{w^i_{email}}{TE}$ el arreglo de palabras contenidas en los e-mails, que también están presentes en el sitio web, donde wi email es la frecuencia de la i-ésima palabra y TE es la cantidad total de palabras en el grupo completo del arreglo de palabras de e-mail.

2. Palabras destacadas. En un sitio web, hay palabras con etiquetas especiales, como diferentes fuentes, por ejemplo, itálica, negrita, o palabras pertenecientes al título. Sea $nw_i = \frac{w^i_{marks}}{TM}$ el arreglo de palabras destacadas dentro de las páginas web, donde w^i_{marks} es la frecuencia de la i-ésima palabra y TM es la cantidad de palabras destacadas en el sitio web completo.

3. Palabras de consultas: Un banco, por ejemplo, tiene motores de búsqueda a través de las cuales los usuarios pueden preguntar por asuntos específicos, por la introducción de palabras clave. Sea $aw_i = \frac{w^i_{ask}}{TA}$ el arreglo de palabras usadas por el usuario en el motor de búsqueda y que esta contenida en el sitio web, donde w^i_{ask} es la frecuencia de la i-ésima palabra y TA es la cantidad total de palabras en el grupo completo.

4. Sitios web relacionados. Usualmente un sitio web pertenece a un segmento de mercado, en este caso el mercado de las instituciones bancarias. Luego, es posible recolectar páginas de sitios web que pertenecen a otros sitios en el mismo mercado. Sea $rw_i = \frac{w^i_{rws}}{RWS}$ el arreglo con palabras utilizadas en el mercado de sitios web incluyendo el sitio web bajo estudio,

donde w^i_{rws} es la frecuencia de la i-ésima palabra y RWS es el número total de palabras en todos los sitios web considerados.

La expresión final $sw_i = ew_i + mw_i + aw_i + rw_i$ es la suma simple de los pesos descritos anteriormente.

En la representación vectorial, cada columna de la matriz M es una página web. Por ejemplo, la késima columna mik con i = 1,. . . ,R es la k-ésima página en el grupo completo de páginas.

Definición 1 (Vector de Palabras por página) *es un vector* $WP^k = (wp^k_1, \ldots, wp^k_R) = (m_{1k}, \ldots, m_{rk})$*, con* $k = 1, \ldots, Q$*, es la representación vectorial de la k-ésima página en el grupo de páginas bajo análisis.*

Con las páginas web en representación vectorial, es posible utilizar la medida de distancia para comparar los contenidos de texto. La distancia común es el coseno del ángulo calculado como:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} WP^i_k \cdot WP^j_k}{\sqrt{\sum_{k=1}^{R}(WP^i_k)^2}\sqrt{\sum_{k=1}^{R}(WP^j_k)^2}} \quad (3)$$

La ecuación (3) permite comparar el contenido de dos páginas web, retornando un valor numérico entre [0, 1]. Cuando las páginas son totalmente diferentes, $dp = 0$, y cuando son las mismas, $dp = 1$. Otro aspecto importante es que la ecuación 3 cumple con el requerimiento de ser computacionalmente eficiente, lo cual la hace más apropiada para ser utilizada en algoritmos de web mining.

4. Extrayendo las preferencias de contenido del usuario de las páginas web

Diferentes técnicas son aplicadas para analizar el comportamiento del usuario en el sitio web, desde una simple estadística de uso de una página hasta complejos algoritmos de web mining. En el último caso, la investigación se concentra en predicciones acerca de cuales páginas el usuario visitará y la información que esta buscando. Principalmente por la utilización de la combinación de las aproximaciones de WUM y WCM, el propósito es analizar las preferencias de texto del usuario web y por esta vía, identificar cuales palabras atraen la atención del usuario durante su navegación en el sitio. Previamente a la aplicación de una herramienta de web mining, la data relacionada con el comportamiento del usuario ha sido procesada para crear vectores de características, cuyos componentes dependerán de la implementación particular del algoritmo de web mining a utilizar y la preferencia de patrones ha ser extraídos.

4.1. Modelando el comportamiento del usuario web

La mayoría de los modelos de comportamiento de usuario web examinan la secuencia de páginas visitadas para crear vectores de características que representan el perfil de navegación del usuario web [12, 21, 36]. Estos modelos analizan el comportamiento de navegación del usuario en un sitio web mediante la aplicación de algoritmos que extraen los patrones de navegación. El siguiente paso es examinar las preferencias del usuario, definido como el contenido preferido de la página web por el usuario; y este es el contenido de texto que captura la atención especial, dado que es utilizada para encontrar información interesante relacionada a un tema particular por un motor de búsqueda. Por lo tanto, es necesario incluir una nueva variable como parte de la información del vector de comportamiento del usuario web acerca del contenido y tiempo gastado en cada página web visitada.

Definición 2 (Vector de comportamiento del usuario (UBV)) *Es un vector* $\nu = [(p_1, t_1), \ldots, (p_m, t_n)]$ *, donde son los parámetros que representan la i-ésima página del visitante y el tiempo gastado en ella en la sesión, respectivamente. En esta expresión,* p_i *es el identificador de la página.*

En la definición 2, el comportamiento del usuario en un sitio web es caracterizado por:

1. Secuencia de páginas; la secuencia de páginas visitadas y registradas en los archivos log. Si el usuario retorna a una página almacenada en el caché del browser, esta acción puede no ser registrada.

2. Contenido de la página; representa el contenido que puede ser texto libre, imágenes, sonidos, etc. Para propósitos de este paper, el texto libre es el utilizado principalmente para representar una página web.

3. Tiempo gastado, tiempo utilizado por el usuario en cada página. Para la página, el porcentaje de tiempo gastado en cada página durante la sesión del usuario puede ser directamente calculado.

4.2. Analizando las preferencias de texto de los usuarios

El objetivo es determinar las palabras más importantes para un sitio web dado para los usuarios, mediante la comparación de las preferencias de texto libre, a través del análisis de páginas visitadas y de tiempo gastado en cada una de ellas [34]. Sin embargo, difiere de las propuestas mencionadas anteriormente, dado que el ejercicio es encontrar las palabras clave que atraen y retienen a los usuarios en el uso de data disponible en la web. La expectación esta en involucrar usuarios pasados y actuales en un proceso continuo de determinación de palabras clave.

Las preferencias del contenido web del usuario son identificadas por la comparación de contenido de las páginas visitadas, [34, 33, 35] por la aplicación

del modelo de vector espacial a las páginas web, con la variante propuesta en la sección 3.2, ecuación (2). Los temas principales de interés pueden ser encontrados por el uso de la medición de la distancia entre vectores (por ejemplo, distancia euclidiana).

Desde el vector de comportamiento del usuario (UBV), las páginas más importantes son seleccionadas asumiendo que el grado de importancia esta correlacionado al porcentaje de tiempo gastado en cada página. El UBV se ordena de acuerdo al porcentaje de tiempo total gastado en cada página. Las ι página más importantes, es decir, las primeras ι páginas, son seleccionadas.

Definición 3 (Vector de Páginas Importantes (IPV)). *Es un vector* $\vartheta_\iota(\nu) = [(\rho_1, \tau_1), \ldots, (\rho_i, \tau_i)]$ *, donde* (ρ_ι, τ_ι) *es el componente que representa la* ι*-ésima página más importante y el porcentaje de tiempo gastado en ella por la sesión.*

Sean α y β dos UBV. La medida de similitud propuesta entre los dos IPV es introducida en la ecuación 4 como:

$$st(\vartheta_\iota(\alpha), \vartheta_\iota(\beta)) = \frac{1}{\iota} \sum_{k=1}^{\iota} \text{mín}\{\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}\} dp(\rho_k^\alpha, \rho_k^\beta) \tag{4}$$

El primer elemento en (4) indica el interés del usuario en las páginas visitadas. Si el porcentaje de tiempo gastado por los usuarios α y β en la k-ésima página visitada es cercano a la otra, el valor de la expresión min.,. será cercano a 1. En el caso opuesto, será cercano a 0. El segundo elemento en (4) es dp, la distancia entre páginas representada en forma vectorial, introducida en (3). En (4) el contenido de las páginas más importantes es multiplicado por el porcentaje de tiempo total gastado en cada página. Esto permite a las páginas con contenidos similares ser distinguidas por intereses diferentes de usuarios.

4.3. Identificando palabras clave del sitio web

Una palabra clave de un sitio web (o web site keyword) es definido como "una palabra o posiblemente un grupo de palabras que hacen de una página web más atractiva para un usuario eventual durante su visita al sitio web" [32]. Es interesante notar que las mismas palabras clave del sitio web pueden ser utilizadas por el usuario en un motor de búsqueda, cuando este está en busca de contenido web.

Para encontrar palabras clave de un sitio web, es necesario seleccionar las páginas web con el contenido textual que es significativo para los usuarios. La suposición es que existe una relación entre el tiempo gastado por el usuario en una página web y su interés en el contenido [31]. La relación es almacenada por el vector de páginas importantes (IPV), dando la información necesaria para extraer las palabras clave de un sitio web a través de la utilización de una herramienta de web mining.

Entre estas técnicas de web mining, se debe colocar especial atención a los algoritmos de clustering. La suposición es, dado un grupo de clusters extraídos de la información generada durante la formación de las sesiones de los usuarios en el sitio en, es posible el extraer las preferencias de los usuarios mediante el análisis de los contenidos del cluster. Los patrones en cada cluster detectado podrían ser suficientes para extrapolar el contenido que él o la usuario esta buscando [20, 23, 30].

En cada IPV, el componente página tiene una representación vectorial presentada por la ecuación (2). En esta ecuación, un paso importante es el cálculo de pesos considerados en el arreglo de palabras especiales swi. Las palabras especiales son diferentes a las palabras normales en el sitio, dado que pertenecen a una fuente alternativa y relacionada o ellas tienen una información adicional mostrando su importancia en el sitio, por ejemplo, una etiqueta HTML que enfatiza una palabra.

El algoritmo de clustering es utilizado para agrupar IPV similares por comparación de la cada componente de tiempo y página del vector, siendo importante el uso de la medida de similitud presentada en la ecuación (4). El resultado debería ser un grupo de clusters cuya calidad debe ser chequeada mediante el criterio de aceptación / rechazo. Un camino simple es aceptar los clusters cuyas páginas comparte un tema principal similar, y en otro caso, rechazar el cluster. En este punto, es necesario conocer que páginas en el sitio son cercanas con los vectores del cluster. Debido a que conoceremos la representación vectorial de las páginas web del sitio y utilizando la ecuación (3) podemos identificar la página más cercana de un cluster dado y de esta forma obtener las páginas adecuadas a un cluster para revisar si las páginas comparten un tema principal en común.

Para cada cluster aceptado y recordando que los centroides contienen páginas donde los usuarios gastan más tiempo durante su sesión respectiva y en la representación vectorial tienen los pesos más altos, el procedimiento de identificación de palabras clave del sitio web es aplicar una medida, descrita en la ecuación (5) (miembro geométrico) para calcular la importancia de cada palabra

$$kw[i] = \sqrt{\prod_{p \epsilon \varsigma} m_{ip}}, \tag{5}$$

donde $i = 1, \ldots, R$ y kw es un arreglo que contiene los pesos para cada palabra relativa a un cluster dado y ς el grupo de páginas representando el cluster. Las palabras clave del sitio web son el resultado del ordenamiento de kw y de la detección de palabras con los pesos más altos, por ejemplo, las 10 palabras con mayor peso.

5. Extrayendo patrones de los datos originados en un sitio web real

Para propósitos experimentales, el sitio web seleccionado debe ser complejo con respecto a varias características: número de visitas, actualización periódica (preferiblemente mensual con el fin de estudiar la reacción de los usuarios a los cambios) y ser rico en contenido textual. La página web de un banco virtual Chileno (sin sucursales físicas, todas las transacciones realizadas electrónicamente) cumple con estos criterios. Cabe destacar que para efectos de privacidad de los datos usados en la investigación, se firmó un acuerdo con el banco, por lo cual so nombre no pude ser mencionado.

Las principales características del sitio web del banco son las siguientes; presentado en Español, con 217 páginas web estáticas y aproximadamente ocho millones de filas en los registros de web log para un periodo de estudio entre Enero y Marzo del 2003.

El comportamiento del usuario en el sitio web del banco es analizado en dos formas. Primero, mediante la utilización de los archivos de registro que contienen información acerca del visitante y del comportamiento de navegación del cliente. Esta información requiere de una reconstrucción previa y limpieza antes de que las herramientas de web mining sean aplicadas. Segundo, la web data en el sitio web en si mismo, específicamente el contenido textual de las páginas web - esto también necesita de un preprocesamiento y limpieza.

5.1. Proceso de reconstrucción de sesiones

La Fig. 5.1 muestra parte de los registros del sitio web bancario e incluye tanto a clientes identificados como visitantes anónimos.

Figura 1: Extracto de un archivo de web log generado en el sitio web de un banco

#	IP	Id	A	Time	Method/URL/Protocol	Statu	Byte	Referer	Agent
1	164.77.129.50	-	-	12/Apr/2003:23:47:44	GET /img/tab.gif HTTP/1.1	200	89	http://www.thebank.cl	MSIE 6.0; Windows 98
2	200.28.206.200	-	20	12/Apr/2003:23:48:31	GET transa/info.htm HTTP/1.1	200	144	/infoeco/info.html	MSIE 4.01; Windows 95
3	200.86.248.170	-	-	12/Apr/2003:23:48:37	GET /img/gen.gif HTTP/1.1	304	0	/ofert/wines/	MSIE 6.0; Windows 98
4	66.249.65.97	-	-	12/Apr/2003:23:48:41	GET /index.htm HTTP/1.1	200	88	-	Googlebot/2.1; google.com/bot.html
5	216.241.8.179	-	31	12/Apr/2003:23:50:03	GET /tx/infoeco/card.htm HTTP/1.1	200	210	/tx/infoeco/prom/	MSIE 6.0; Windows NT 5.1
6	164.77.129.50	-	-	12/Apr/2003:23:48:34	GET /tx/infoeco/ HTTP/1.1	200	186	/tx/infoeco/card.htm	MSIE 6.0; Windows 98
7	200.28.206.200	-	20	12/Apr/2003:23:51:13	GET transa/account.htm HTTP/1.1	200	180	/transa/info.htm	MSIE 4.01; Windows 95
8	216.241.8.179	-	31	12/Apr/2003:23:51:23	GET /tx/infoeco/ind.htm HTTP/1.1	200	300	/tx/infoeco/card.htm	MSIE 6.0; Windows NT 5.1
9	200.86.248.170	-	-	12/Apr/2003:23:51:41	GET /prom/wine.html HTTP/1.1	404	0	/ofert/wines/	MSIE 6.0; Windows 98
10	164.77.129.50	-	44	12/Apr/2003:23:52:04	GET /tx/infoeco/ind.htm HTTP/1.1	200	186	/tx/infoeco/	MSIE 6.0; Windows 98

El acceso de los clientes al sitio es a través de una conexión segura, utilizando un protocolo SSL que permite el almacenamiento de un valor de identi-

ficación en el parámetro de autentificación de usuario en el archivo de registros web. Otro modo de identificación de usuarios es mediante cookies, pero algunas veces estas son desactivadas por los usuarios en sus navegadores. En este caso sería necesario el reconstruir la sesión del visitante.

Durante el proceso de reconstrucción de sesiones, se aplican filtros a los registros del sitio web. En este caso particular, solo se utilizan registros de requerimiento de páginas web para analizar el comportamiento específico del usuario en el sitio. También es importante la limpieza de sesiones anormales, por ejemplo, web crawlers, como es mostrado en la Fig. 1, línea 4, donde un robot perteneciente a Google es detectado.

Las filas de los registros log del sitio web contienen cuatro meses de transacciones, con aproximadamente 8 millones de registros. Sólo se consideran los registros relacionados con páginas web para la reconstrucción de sesiones y análisis del comportamiento del usuario; la información que apunta a otros objetos como imágenes, sonidos, etc, son limpiadas.

5.2. Preprocesamiento del contenido de una página web

Mediante la aplicación de filtros a los textos de las páginas web, se ha encontrado que en el sitio completo contiene R=2034 palabras diferentes para ser utilizadas en el análisis.

Considerando los pesos de las palabras y la especificación de palabras especiales, fue utilizado el procedimiento presentado en la sección 3.2, con el fin de calcular sw_i, en la ecuación 2. Las fuentes de datos fueron:

1. Palabras destacadas. Dentro de las páginas web, se encontraron 743 palabras diferentes después de la aplicación del paso de preprocesamiento y limpieza.

2. Sitios web relacionados: Cuatro sitios web fueron considerados, cada un de ellos con aproximadamente 300 páginas.

El número total de palabras diferentes fue de 9253, con 1842 de ellas contenidas en el contenido del sitio web.

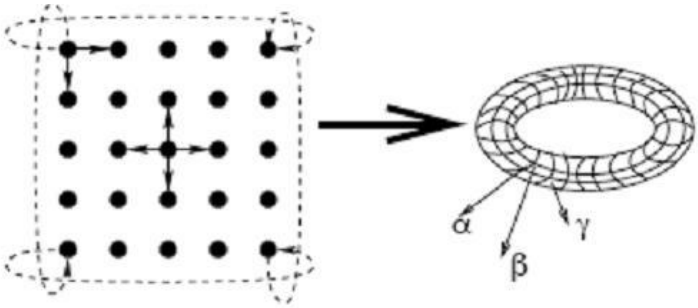
Después de la identificación de palabras especiales y sus respectivos pesos, es posible calcular el peso final para cada palabra en la totalidad del sitio web, por la aplicación de la ecuación (2). Luego, se obtiene la representación vectorial para todas las páginas del sitio

5.3. Analizando las preferencias de texto del usuario

Dos redes neuronales fueron aplicadas al web data para la identificación de clusters. La red neuronal artificial del tipo Kohonen (Self Organizing Feature Map; SOFM) y K-means.

Esquemáticamente, una red SOFM es una red neuronal artificial no supervisada, correspondiente a un arreglo de neuronas de dos dimensiones. Cada neurona esta constituida por un arreglo bidimensional de vectores de n dimensiones cuyos componentes son los pesos sinápticos. Por construcción, todas las neuronas reciben el mismo input en un momento determinado. La noción de vecindad entre neuronas define diversas topologías. Para el caso de este trabajo, se utilizó la topología toroidal [38] que significa que las neuronas localizadas de un borde, son cercanas al borde opuesto. La ventaja de la topología radica en que mantiene la continuidad de los clusters o cuando la data corresponde a secuencias de eventos.

Figura 2: de Vectores de Páginas Importantes en un SOFM con topología toroidal.

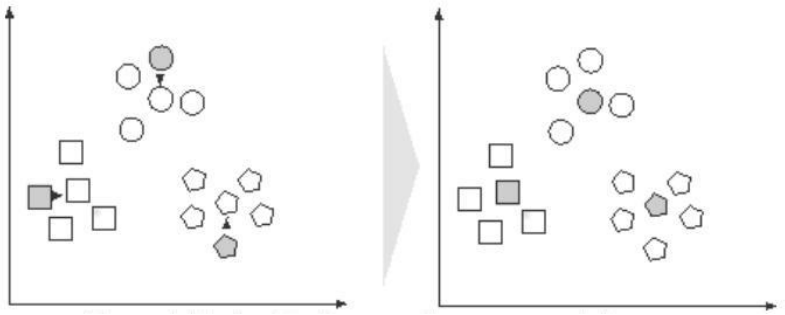


K-means es una red de aprendizaje supervisada y predefiniendo el número de centroides, genera las agrupaciones de vectores llamados miembros en torno a ellos. K-means para detectar las pertenencias a sus centroides tradicionalmente utiliza la distancia euclideana para discernir que centroide es más representativo para un vector. Puesto que la investigación se centra en un vector de comportamiento del usuario con una estructura diferente a la euclideana, se hace modificación de esta red de aprendizaje y se utiliza la medida de similitud presentada en la ecuación (4) para establecer las pertenencias a los centroides correspondientes. Para el caso de esta investigación, el principal input de este algoritmo - los K centroides - será originado por el resultado que entregue la red SOFM que será inicialmente utilizada para el análisis del comportamiento del usuario y que parte de los resultados que retorne serán los clusters detectados. La Fig. 3 muestra el comportamiento de los centroides a medida que se van encontrando mejores representantes.

5.4. Analizando las preferencias del usuario con una red SOFM

Se ha fijado en 3 el número máximo de dimensiones del vector. Luego, un SOFM con 3 neuronas de entrada y 32*32 neuronas de salida fue utilizado

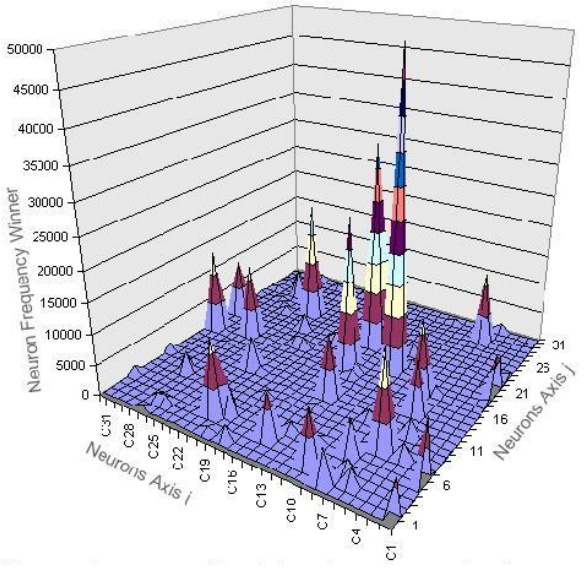
Figura 3: Evolución de centroides en una red K-means



para encontrar los clusters de vectores de páginas importantes.

La Fig. 4 muestra las posiciones de las neuronas en el SOFM en los ejes x e y. El eje z es la frecuencia normalizada de veces que una neurona gana durante el entrenamiento.

Figura 4: Clusters de vectores de páginas importantes desde una red SOFM.



La Fig. 4, muestra 8 cluster principales que contienen información acerca de las páginas más importantes del sitio web. Sin embargo, sólo 5 fueron aceptadas. El criterio de aceptación / rechazo es simple; si las páginas de un centroide de cluster tienen el mismo tema principal, entones el cluster es aceptado, de otra forma se rechaza.

Los centroides de los clusters son mostrados en el Cuadro 1. La segunda columna contiene las neuronas centrales (neuronas ganadoras) para cada cluster y representa las páginas visitadas más importantes.

Cuadro 1: Vectores de páginas importantes obtenidos con SOFM.

Cluster	Páginas Visitadas
1	(171, 130, 159)
2	(76, 58, 130)
3	(175, 78, 10)
4	(78, 32, 130)
5	(130, 171, 159)

5.5. Analizando las preferencias del usuario con una red K-means

Desde el resultado obtenido con la aplicación de SOFM, se inicializa el entrenamiento de la red K-means. La cantidad de centroides es de acuerdo al número final aceptado. Luego, el proceso se inicializó con 5 centroides. La detención de asignación de miembros al clusters se produce cuando los mejores representantes de cada centroide van variando de menor manera llegando en un punto a quedar prácticamente establecido el centroide que será el representante de los miembros.

Para cada centroide se obtiene las páginas visitadas representativas de los grupos. En el Cuadro 2 se muestran las páginas visitadas de los representantes de los clusters y la cantidad de miembros identificados en ellos.

Cuadro 2: Vectores de páginas importantes obtenidos con K-means.

Cluster	Páginas Visitadas
1	(117,192,19)
2	(21,10,179)
3	(205,128,210)
4	(55,18,41)
5	(24,104,95)

5.6. Identificación de web site keywords

Se requiere un paso final para obtener las palabras clave de un sitio web: analizar cuales son las palabras que tienen una mayor importancia relativa con respecto al sitio web completo.

Las palabras clave y su importancia relativa en cada cluster son obtenidas por la aplicación de la ecuación (5). Por ejemplo, si el cluster es ($\varsigma = \{171, 130, 159\}$), entonces $kw[i] = \sqrt[3]{m_{i171}, m_{i130}m_{i159}}$, con $i = 1, \ldots, R$.

Finalmente, ordenando las kw de forma descendente, podemos seleccionar las k palabras más importantes para cada cluster, por ejemplo $k = 5$.

No se nos permite mostrar las palabras clave específicas debido a la cláusula de confidencialidad con el banco, por esta razón las palabras son numeradas. El Cuadro 3 muestra las palabras encontradas con el método propuesto.

El Cuadro 4 muestra un grupo seleccionado de palabras clave de todos los clusters. Las palabras clave en si, sin embargo, no tienen mucho sentido. Estas necesitan un contexto de página web donde ellas podrían ser utilizadas como palabras especiales, por ejemplo, palabras destacadas para enfatizar un concepto o como palabras vinculadas a otras páginas.

Cuadro 3: Las 5 palabras más importantes por cluster

C	Palabras Clave	Peso ordenado
1	$(w_{2032}, w_{1233}, w_{287}, w_{1087}, w_{594})$	(2.35,1.93,1.56,1.32,1.03)
2	$(w_{1003}, w_{449}, w_{895}, w_{867}, w_{1567})$	(2.54,2.14,1.98,1.58,1.38)
3	$(w_{1005}, w_{948}, w_{505}, w_{1675}, w_{1545})$	(2.72,2.12,1.85,1.52,1.31)
4	$(w_{501}, w_{733}, w_{385}, w_{684}, w_{885})$	(2.84,2.32,2.14,1.85,1.58)
5	$(w_{200}, w_{1321}, w_{206}, w_{205}, w_{1757})$	(2.33,2.22,1.12,1.01,0.93)

Cuadro 4: Parte de las palabras descubiertas.

#	Palabras Clave
1	Cuenta
2	Fondo
3	Inversión
4	Tarjeta
5	Hipotecario
6	Seguro
7	Cheques
8	Crédito

La recomendación específica es utilizar las palabras clave como “palabras para escribir” en un sitio web, es decir, los párrafos escritos en la página deberían incluir algunas palabras clave y algunas podrían ser un enlace a otras páginas.

Además es posible sobre la base de este ejercicio el realizar recomendaciones de contenidos de texto. Sin embargo, para reiterar, las palabras clave no funcionan de forma separada sino que requieren de un contexto que las utilice. Revisando el Cuadro 2, para cada cluster, la palabra clave descubierta podría ser utilizada para reescribir un párrafo o una página web completa. Adicionalmente, es importante insertar palabras clave para destacar conceptos específicos.

Las palabras clave también son utilizadas como palabras índice para un motor de búsqueda, es decir, algunas podrían ser utilizadas para personalizar

el crawler que visita sitios web y carga páginas. Luego, cuando un usuario esta buscando por una página en específico en un motor de búsqueda, la probabilidad de obtener el sitio web se incrementa.

5.7. Mejorando el contenido textual el sitio web

Las palabras clave son conceptos para motivar los intereses de los usuarios y hacerlos visitar el sitio web. Están para ser jugadas dentro de su contexto como palabras aisladas que pueden tener un pequeño sentido , dado que los clusters representar contextos diferentes. La recomendación específica es utilizar palabras clave como "palabras para escribir" en el sitio web.

En cuanto cada página contiene un contenido de texto específico, es posible asociar las palabras clave de un sitio web a un contenido de la página; y desde esta sugerir la revisión o reconstrucción de un nuevo contenido en el sitio web. Por ejemplo, si la nueva versión de la página es relacionada con "tarjetas de crédito", entonces las palabras clave del sitio web "crédito, puntos y promociones" deben ser asignadas para la reescritura del contenido textual de la página.

5.8. Testeo de la efectividad de las recomendaciones de texto

La detección y aplicación de web site keywords no garantizan el éxito de aplicación en un contenido textual. Incluso, el riesgo de utilizarlas puede generar disgusto en un usuario habitual y por lo tanto abandonar o dejar de utilizar el sitio web. Como medida precautoria, se realizaron test de efectividad de las web site keywords. Sobre el contenido del sitio web se extrajeron 10 párrafos que contenían para el caso de 5 de ellos web site keywords y otros no las contenían. El resultado se realizó sobre un universo de 10 personas con el fin de conocer la recepción que ellos tenían respecto a párrafos que contenían las palabras detectadas, según el contexto de si entregaban información relevante en un sitio bancario. El resultado de este test es el que se muestra en el Cuadro 6.

Como se puede apreciar, aquellas palabras que contenían web site keywords eran para el usuario mucho más interesantes e importantes en el contexto de navegación en que estaban inmersos, versus aquellos párrafos en que no había presencia de dichos web site keywords. Las web site keywords atraen la atención del usuario y pueden ser una muy buena guía en el diseño de contenidos específicos de un sitio web. Esta combinación de elementos que se alinean a los que el usuario busca puede otorgar un mejor resultado en la satisfacción de los clientes.

Cuadro 5: Párrafos testeados para análisis de keywords.

#	Incluye web site keyword	Párrafo
1	Si	Orientado a empresas que deseen manejar excedentes de caja, así como a Personas que quieran mantener parte de sus recursos **invertidos** en un **fondo mutuo**, cuya cartera esté compuesta exclusivamente por instrumentos de **deuda** nacional, obteniendo rentabilidad y liquidez a corto plazo.
2	Si	Solicitándolos con un día hábil bancario de antelación, se pagarán mediante **cheques** nominativos, **vales** vista o **depósitos** en **cuentas** corrientes, de acuerdo a sus instrucciones.
3	Si	Este plan busca otorgar a tus **Ahorros Previsionales** Voluntarios acumulados a esta fecha y los futuros, una atractiva y segura **rentabilidad** que te permitirá poder mejorar considerablemente tus **ahorros** para una mejor **pensión.**
4	Si	Para obtener información de tu **Cuenta Corriente** y de tu **Línea** de **Sobregiro** debes seguir los siguientes pasos
5	Si	El Servicio de Mensajería es un servicio de entregas y **retiros** de dinero, especies valoradas y documentos que podrás utilizar siendo **cliente**
6	No	Para solicitar tu Plan debes completar la siguiente información y se contactarán contigo.

6. Conclusiones

Cuando un usuario visita un sitio web, hay una correlación entre el máximo de tiempo gastado por sesión en una página y su contenido de texto libre. Esto permite modelar las preferencias del usuario a través del "Vector de Páginas Importantes (IPV)", el cual es la estructura de datos básica de almacenamiento de páginas donde el usuario gasta más tiempo durante e su sesión. Mediante la utilización de IPV como entrada en un SOFM y K-means, se pueden identificar clusters que contienen la navegación del usuario e información de sus preferencias de contenido.

El criterio de aceptación / rechazo de un cluster es simple: si las páginas dentro de cada cluster están relacionadas con un tema principal similar, entonces el cluster es aceptado, en caso contrario, se rechaza. Aplicando este

Cuadro 6: párrafos testeados para análisis de keywords.

		Opinión de aceptabilidad				
#	Incluye web site Keyword	Irrelevante	Moder. irrelevante	Algo relevante	Moder. relevante	relevante
1	Si				8	2
2	Si			4	4	2
3	Si			4	2	4
4	Si				7	3
5	Si			1	2	7
6	No	1	3	5	1	
7	No	3	2	5		
8	No	6	4			
9	No	5	2	1	2	
10	No	7	2	1		

criterio, 5 clusters son aceptados y el patrón contenido en cada una de ellas fue utilizado para extraer las palabras clave del sitio web.

El texto contenido en las páginas web puede ser mejorado utilizando las palabras clave del sitio web, y por esta vía atraer la atención del usuario cuando están visitando un sitio web. Sin embargo, es necesario recordar que estas palabras no pueden ser utilizadas de forma individual, de hecho necesitan de un contexto, el cual es provisto por un ser humano.

Como validación de las palabras detectadas, se realizó un testeo de párrafos que contenían dichos web site keywords versus otros que no contenían. El resultado fue satisfactorio corroborando la importancia de las palabras pues aquellos contenidos con web site keywords parecían más relevantes que otras que no contenían dichas palabras, por lo tanto el interés del usuario en contenidos con los keywords se hace mayor y de ahí la importancia de dar uso a estas palabras en los párrafos del contenido.

Como trabajo futuro, se aplicará la metodología en otros web data, por ejemplo las imágenes y objetos no textuales, con el fin de identificar cuales elementos atraen la atención del usuario en el sitio web.

Agradecimientos: Este trabajo fue parcialmente financiado por el Instituto Milenio Sistemas Complejos de Ingeniería

Referencias

[1] Green, Paul E. and V. Srinivasan (1990), "Conjoint Analysis in Marketing Research: New Developments and Directions", Journal of Marketing 54, 4, 3-19.

[2] E. Amitay and C. Paris. "Automatically summarizing web sites: Is there

any way around it?" In Procs. of the 9th Int. Conf. on Information and Knowledge Management, pages 173-179, McLean, Virginia, USA, 2000.

[3] R. Baeza-Yates. "Web usage mining in search engines", chapter Web Mining: Applications and Techniques, pages 307-321. Idea Group, 2004.

[4] B. Berendt, A. Hotho, and G. Stumme. "Towards semantic web mining". In Proc. in First Int. Semantic Web Conference, pages 264-278, 2002.

[5] B. Berendt and M. Spiliopoulou. "Analysis of navigation behavior in web sites integrating multiple information systems". The VLDB Journal, 9:56-75, 2001.

[6] D. Buttler. "A short survey of document structure similarity algorithms". In Procs. Int. Conf. on Internet Computing, pages 3-9, 2004.

[7] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. "Focused web searching with pdas". Computer Networks, 33(1- 6):213-230, June 2000.

[8] L. D. Catledge and J. E. Pitkow. "Characterizing browsing behaviors on the world wide web". Computers Networks and ISDN System, 27:1065-1073, 1995.

[9] G. Chang, M. Healey, J. McHugh, and J. Wang. "Mining the World Wide Web". Kluwer Academic Publishers, 2003.

[10] W. Chuang and J. Yang. "Extracting sentence segment for text summarization? a machine learning approach". In Procs. Int. Conf. ACM SIGIR, pages 152-159, Athens, Greece, 2000.

[11] R. Cooley, B. Mobasher, and J. Srivastava. "Data preparation for mining world wide web browsing patterns". Journal of Knowlegde and Information Systems, 1:5-32, 1999.

[12] U. Hahn and I. Mani. "The challenges of automatic summarization". IEEE Computer, 33(11):29-36, 2000.

[13] A. Joshi and R. Krishnapuram. "On mining web access logs". In Proc. of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 63- 69, 2000.

[14] A. P. Jr and N. Ziviani. "Retrieving similar documents from the web". Journal of Web Engineering, 2(4):247-261, 2004.

[15] D. Lawrie, B. W. Croft, and A. Rosenberg. "Finding topic words for hierarchical summarization". In Proc. 24th Int SIGIR Conf. on Research and Development in Information Retrieval, pages 349-357, New Orleans, Louisiana, USA, 2001. ACM Press.

[16] E. Liddy, K. McVearry, W. Paik, E. Yu, and M. McKenna. "Development, implementation and testing of a discourse model for newspaper texts". In Procs. Int. Conf. on ARPA Workshop on Human Language Technology, pages 159-164, Princeton, NJ, USA, 1993.

[17] G. Linoff and M. Berry. "Mining the Web". Jon Wiley & Sons, New York, 2001.

[18] S. Loh, L. Wives, and J. P. M. de Oliveira. "Concept based knowledge discovery in texts extracted from the web". SIGKDD Explorations, 2(1):29-39, 2000.

[19] I. Mani and M. Maybury. "Advances in automatic text summarization". MIT Press, Cambridge, Mass., 1999.

[20] S. Mitra, S. K. Pal, and P. Mitra. "Data mining in soft computing framework: A survey". IEEE Transactions on Neural Networks, 13(1):3-14, 2002.

[21] B.Mobasher, R. Cooley, and J. Srivastava. "Creating adaptive web sites through usage-based clustering of urls". In Procs.Int Conf IEEE Knowledge and Data Engineering Exchange, November 1999.

[22] B. Mobasher, R. Cooley, and J. Srivastava. "Automatic personalization based on web usage mining". Communications of the ACM, 43(8):142-151, 2000.

[23] M. F. Porter. "An algorithm for suffix stripping". Program; automated library and information systems, 14(3):130-137, 1980.

[24] T. A. Runkler and J. Bezdek. "Web mining with relational clustering". International Journal of Approximate Reasoning, 32(2-3):217-236, Feb 2003.

[25] G. Salton, A. Wong, and C. S. Yang. "A vector space model for automatic indexing". Communications of the ACM archive, 18(11):613-620, November 1975.

[26] M. Spiliopoulou. "Data mining for the web". In Principles of Data Mining and Knowledge Discovery, pages 588-589, 1999.

[27] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. "A framework for the evaluation of session reconstruction heuristics in web-usage analysis". INFORMS Journal on Computing, 15:171-190, 2003.

[28] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. "Web usage mining: Discovery and applications of usage patterns from web data". SIGKDD Explorations, 1(2):12-23, 2000.

[29] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. "Recovering traceability links in multilingual web sites". In Procs. Int Conf. Web Site Evolution, pages 14-21. IEEE Press, 2001.

[30] P. Tonella, F. Ricca, E. Pianta, and C. Girardi. "Restructuring multilingual web sites". In Procs. Int. Conf. Software Maintenance, pages 290-299. IEEE Press, 2002.

[31] J. D. Velázquez and V. Palade. "A knowledge base for the maintenance of knowledge extracted from web data". Journal of Knowledge-Based Systems, 20(3):238-248, 2007.

[32] J. D. Velásquez, S. Ríos, A. Bassi, H. Yasuda, and T. Aoki. "Towards the identification of keywords in the web site text content: A methodological approach". International Journal of Web Information Systems, 1(1):11-15, March 2005.

[33] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki. "A methodology to find web site keywords". In Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service, pages 285-292, Taipei, Taiwan, March 2004.

[34] J. D. Velásquez, H. Yasuda, and T. Aoki. "Combining the web content and usage mining to understand the visitor behavior in a web site". In Procs. 3th IEEE Int. Conf. on Data Mining, pages 669-672, Melbourne, Florida, USA, November 2003.

[35] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. "Using the kdd process to support the web site reconFig.tion". In Procs. IEEE/WIC Int. Conf. on Web Intelligence, pages 511-515, Halifax, Canada, October 2003.

[36] J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. "A new similarity measure to understand visitor behavior in a web site". IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization, E87-D(2):389-396, February 2004.

[37] J. Xiao, Y. Zhang, X. Jia, and T. Li. "Measuring similarity of interests for clustering web-users". In ADC '01: Proceedings of the 12th Australasian conference on Database technologies, pages 107-114, Washington, DC, USA, 2001. IEEE Computer Society.

[38] K. Zechner. "Fast generation of abstracts from general domain text corpora by extracting relevant sentences". In Procs. Int. Conf. on Computational Linguistics, pages 986-989, 1996.

[39] J. D. Velásquez, H. Yasuda, T. Aoki, R. Weber and E. Vera (2003) "Using self organizing feature maps to acquire knowledge about visitor behavior in a web site". Lecture Notes in Artificial Intelligence, 2773(1): 951-958