Mediciones y mapeo de la World Wide Web mexicana: hacia la generación y la incorporación de metadatos de origen telemático en los sistemas de información espaciales

Dr. Djamel Toudert
Instituto de Investigaciones Sociales de la UABC
Edificio del Postgrado e Investigaciones
Blvd, Benito Juárez S/N
Mexicali, Baja California, México.
Tel y Fax: (52)-(686)-66-29-85
E-mail: toudert@faro.ens.uabc.mx

Introducción:

La red mundial ó la WWW (por sus siglas en inglés: *World Wide Web*), constitye actualmente una nueva red emblemática de información global para cuasi la mayoría de los usuarios de la telemática[1].
En México, La aventura de la WWW inició en 1994 con el "Home page" del ITESM y la presentación de la UDG de su sección de Arte y Cultura Mexicana[2] con el "Mosaic". De las 150 direcciones IP[3] en México en 1994, el acceso a la Internet solamente para Telmex alcanzó a finales de 1999 el número de 402,754 cuentas en 117 ciudades del país, con un crecimiento de 175% anual (Telmex, 2000). Durante el período que va de enero de 1996 a enero del 2000 el incremento medio anual del número nacional de *hosts*[4], fue del orden del 46.7% anual. (NIC- México, 2000).
Mas allá de la problemática de la repartición espacial de la infraestructura de comunicación y la cuantificación de flujos de telefonía que han cautivado la atención de la telegeografia, la WWW ofrece también la posibilidad de acceso a los discursos inherentes al proceso mismo de comunicación. Lo que podemos llamar con todavía mucho cuidado la "cibergeografia", se esta haciendo camino hacia la producción electrónica del espacio enfocándose de una manera especial al ordenamiento de esta producción y sus posibles impactos sociales y territoriales.
Partiendo del paradigma identificando el discurso de la WWW como una producción de actores espacialmente localizados, es posible rastrear los contenidos para estudiar su lógica de organización y operar una cuantificación localizada del tipo y la naturaleza de los flujos telemáticos. La integración de los metadatos resultantes en un sistema de informacion espacial, puede contribuir con el conjunto de datos socioeconómicos habitualmente levantados a una mejor comprensión de los procesos de reorganización espacial y el papel de la telemática en las políticas de desarrollo y el ordenamiento territorial.
En esta contribución, vamos a desplegar una metodología de investigación de la WWW, que nos permitió realizar una encuesta nacional alrededor del uso de la Web mexicana en los principios del año 2000.

[1] La convergencia entre las comunicaciones y la informática.
[2] La información "histórica" del desarrollo de la Internet está disponible en la página electrónica de la sociedad mexicana de Internet.
[3] Direcciones que identifican cada máquina conectada a la red.
[4] En esta caso las computadoras conectadas a la Internet.

1.- Conceptualización de los marcos generales

La WWW para nosotros encarna una materialización del concepto del ciberespacio accesible a través de los URLs (*Uniform Resource Locators*) en un ambiente comunicativo de hipertexto en su mayoría todavía escrito en HTML (*HyperText Markup Language*).
Adoptando la representación en gráfico propuesta por Wood et al, 1995, se puede ver el ciberespacio como un espacio finito de sitios $A=\{a_1, a_i,...a_n\}$ de $|A|$ recursos con una relación $\alpha \subseteq A$ x A entre las parejas $(a_i, a_j) \in \alpha$ de los recursos ligado. En este ambiente, se puede modelar este sistema con los gráficos conectados $G = (A, \alpha)$ en donde $A = A(G)$ son los vértices y $\alpha = \alpha(G)$ son los arcos non direccionales enlazando los vértices del gráfico. En esta arquitectura, el tamaño de G esta determinado por él numero de los arcos:

$$n = |G| = |A(G)| = |A|$$

La informacion en el gráfico G esta expresada por una matriz adyacente $S = S(G)$ del gráfico G con un tamaño n x n. En el caso que el vértice a_i es adyacente a su homologo a_j: s_{ij} =1, y en el caso contrario, s_{ij} = 0. La diagonal de la matriz en esta representación es indefinida del momento que no se admitan autolazos en el gráfico.
Para la extracción de los objetos y componentes del gráfico G, se usan los elementos siguientes:

- Nodos adyacentes al vértice a

  $\Gamma: (G, A) \rightarrow \wp(A), \quad (G, a_i) \rightarrow \Gamma(G, a_i) = \Gamma_{ai}^{G}$

$\wp(A)$ nos da la posición de las componentes de A

- El grado del vértice a toma el numero de los arcos incidentes al vértice. En la matriz de adyacencia los grados nodales son iguales a la suma de la columna o de la línea, estos últimos representan la característica del recurso.

$$\Gamma : (G, A) \rightarrow \wp(A), (G, a_i) = \left|\Gamma_{ai}^{G}\right| = \sum_{j=1}^{n} s_{ij} = \gamma_{ai}^{G}$$

La desimilaridad entre dos recursos, puede ser definida por el ancho mas reducido de la secuencia de arcos:

$$d : AxA \rightarrow \mathrm{N}.(a_i, a_j) \rightarrow d(a_i, a_j) = d_{ij}$$

Con:

$$d_{ij} = 0.$$
$$d_{ij} \geq 0.$$
$$d_{ij} = d_{ji}.$$
$$d_{ik} \leq d_{ij} + d_{jk}.\forall a_i, a_j, a_k \in A$$

Y con:

$$d_{ij} = 1.\forall(a_i, a_j) \in \alpha, a_i \neq a_j$$

El calculo de la distancia geodésica se realiza con la construcción de una matriz exponencial empezando con p = 1. Cuando p =1, la matriz exponencial es la matriz de adyacencia: los recursos son adyacentes y las distancias entre ellos es igual a 1. en caso s_{ij} = 0 y s_{ij} > 0, la distancia más cercana tiene un ancho de 2. Así, el primer

$$d_{ij} = \min_p (s_{ij}^{|p|} > 0).$$

exponente p para quien s_{ij} es diferente de 0 nos da el ancho de la secuencia de arcos igual a d_{ij}:

Con esta definición métrica, la construcción de la matriz de distancia es posible: $D(G)=D=[d_{ij}]$ del grafico G compuesto por los vectores $di= (d_{i1}, d_{ij}, ..., d_{in})$ de los n-dimenciones. Así, cada vector d_i dado como único representante de un recurso, constituye un punto de n-dimenciones en el espacio.

2.- Conceptos y reglas del mapeo:

Siguiendo las indicaciones anteriores el mapeo del gráfico *G* a través del gráfico *H* como una aproximación de la noción del ciberespacio en un medio de baja dimensión es posible. Para poder concretizarlo, se utilizan dos tipos de aproximaciones. La primera consiste en buscar el mapeo del gráfico *G* a través del gráfico *H* con el apoyo de una informacion local conocida por cada recurso. En la segunda aproximación, se otorga una representación del gráfico *G* a través de un mapa de puntos a dentro de un espacio de baja dimensión. Este ultimo mapeo es basado en la posibilidad de búsqueda con el uso de la matriz de distancias y la correspondencia de los nodos en el gráfico *H*.
Para cumplir con las condiciones de visualización hay que buscar una representación propia del ciberespacio a través de la conservación de los objetos topologicos en un medio compuesto por componentes discretos.
Si supongamos que el mapeo puede definirse como una equivalencia de clases entre *A* y *B*:

$$\Phi : A \to /- = b. \qquad a_i \to \Phi(a_i) = \Phi_i = |a_i| = b_i$$

En esta sentido, la configuración de una imagen previa es la siguiente:

$$\Phi^{-1} : B \to A. \qquad b_r \to \Phi^{-1}(b_r) = \Phi_r^{-1} = a_i \subseteq AI\Phi(a_i \in A_i) = b_{r.}$$

Para optimizar el grado de la representación se debe tomar en cuenta el morfismo definido aquí en sus varios esquemas:

- Homomorfismo:
  $(\Phi.\alpha) \subseteq \beta.$ $\quad \forall(a_i, a_j) \in \alpha \Rightarrow (\Phi_i, \Phi j) \in \beta.$

- Monomorfismo:
  $(\Phi.\alpha) \subseteq \beta$ $\quad \Phi es 1-1 \quad \forall(a_i, a_j) \in \alpha \Leftrightarrow (\Phi_i, \Phi_j) \in \beta$

- Isomorfismo:
  $(\Phi.\alpha) \subseteq \beta'$ $\quad (\Phi^{-1}.\beta') \subseteq \alpha$

La resolución de estos problemas de morfismo no se puede dar por una forma polinomial y en estas condiciones, suponemos que el resultado de la transformación tiene un objeto central derivado de una relación monótona entre las distancias, lo que lleva a preservar una clasificación por orden de las desimilaridades en la transformación. Así la medida es abandonada durante el mapeo y la transformación debe conformarse a las condiciones de monotonía definidas por:

$$d_{ij} \leq d_{kl} \Leftrightarrow \delta(\Phi_i, \Phi_j) \leq \delta(\Phi_k, \Phi_l). \quad \forall a_i, a_j, a_k, a_l \in A.$$

Si se toma en consideración la medición en el mapeo, la configuración necesita satisfacer el siguiente:

$$d_{ij} = f(\delta(\Phi_i, \Phi j)). \qquad \forall a_i, a_j \in A$$

En este esquema, *f* es una función monótona de la distancia que puede tomar la siguiente forma:

$$f(\delta(\Phi_i, \Phi_j)) = E\delta(\Phi_i, \Phi_j) + \varepsilon$$

La condición para llevar el mapeo en este ultimo esquema es la isometria a fines de preservar la topología definida por los términos:

$$\delta(\Phi_i, \Phi_j) = d_{ij} \qquad \forall a_i, a_j \in A$$

Bajo estos condiciones de mapeo, las características de cualquier recurso pueden ser calculadas para un pixel *br* con la suma de los grados de los recursos correspondientes:

$$K_r = \sum_{a_i \in \Phi_r^{-1}} \gamma^G a_i$$

3.-La resolución practica de los conceptos teóricos

La resolución practica del modelo conceptual en términos de búsqueda, es llevada a cabo con el apoyo de los algoritmos transversales apropiados para puntear la informacion especifica de una estructura conectiva del gráfico. La búsqueda de profundidad y la búsqueda de amplitud, son aproximadas con una lista adyacente cumpliendo con un requerimiento de tiempo proporcional a $|A|+|\alpha|$.
Para la construcción de la matriz de adyacencia, los recursos son identificados con enteros comprendidos entre 1 y $|A|$. La construcción en esta sentido, cumple con $|\alpha|$ pasos y el resultado cabe en una matriz de $|A|^2$ bits.
La matriz de distancia es deducida de la matriz de adyacencia a partir de la relación identificando la longitud mas corta en la secuencia de los arcos desde y hacia cada recurso. La aplicación del algoritmo de la trayectoria corta con un numero de iteraciones igual a $|A|$ y una complejidad del resultado de $0((|A|+|\alpha|)|A|\log|A|)$. El algoritmo de Floyd, puede emplearse también para la resolución de conflictos generados por todas las parejas de trayectorias cortas en $(0|A|^3)$.
El mapeo propiamente dicho, puede generarse desde dos perspectivas de medición:

- Medición de la escala multidimencional:

La medición de escala multidimencional, consiste en buscar una configuración capaz de conservar la característica medible de la transformación o una reducción linear de la dimencionalidad (Cox et al; 1994). Para cumplir con este propósito, se utiliza la técnica de los componentes principales de la expansión Kar-hunen Loéve (Jolliffe,1986).

- Sin-medicion de la escala multidimencional:

La sin-medicion de escala multidimencional, consiste en no tomar en consideración la medición de escala multidimencional. Operando en este ambiente, se contempla solamente un mapeo desde el enfoque de la condición monótona (Kruskal y Wish, 1981; Li et al., 1995). Para la realización de esta operación, se puede utilizar el método del mapeo lineal (Kohonen, 1995) o el algoritmo del mapeo auto-organizado (Li et al., 1995)
El concepto del mapeo auto-organizado es ideado por la primera vez por Kohonen y consiste en operar una clasificación no supervisada a una red de neuronas compuestas por una capa de entrada y otra capa de neuronas competidoras (Kohonen, 1995). En este calculo, se da continuidad a la topología según su similaridad.
La capa de entrada, representa las coordinadas de cada recurso como un punto en n-dimenciones del espacio: $d_i = (d_{i1}, d_{i2},\ldots, d_{in})$. Las neuronas en la capa competidora son los pixeles b_r y el peso del vector de referencia asociado con ellos es : $w_r = (w_{r1}, w_{r2}, \ldots, w_{rn}) \in \Re^n$ con $r = 1, 2, \ldots, m$.

La concretizaron de la distancia Euclidiana entre un vector de entrada d_i y el peso del vector w_r de una neurona b_r, toma la expresión siguiente:

$$\|d_i - w_r\| = \sqrt{\sum_{j=1}^{n} (d_{ij} - w_{rj})^2}$$

En estas condiciones el mapeo es procesado a través del calculo de la neurona ganadora por cada vector de entrada, seguido por la aplicación de la función de mapeo por cada recurso a_i a fines de recibir de regreso el pixel b_r:

$$\Phi_i = \arg min_r (\|d_i - w_r\|).$$

4.- Aplicación y algunos resultados

El desarrollo de los conceptos anteriormente descritos, nos han permitido llevar acabo la encuesta nacional del uso de la WWW en los principios del año 2000.
Dentro del universo mexicano de la WWW estimado alrededor de 25000 hojas electrónicas[5], se selecciono al azar un 5% de las URLs totales sobre las cuales se opero una estandarización de los resultados a través de las proporciones totales de los dominios de bajo nivel existentes en todo el país (.net, .com, .edu, .gob, .org y otros genéricos).
En primer lugar, se bajo de la WWW el contenido textual de cada URLs de la muestra, seguido por un rastreo sistemático dentro de los recursos del "*Whos*" para poder localizar geográficamente cada URLs. En segundo lugar, se indexó el contenido textual de cada URLs en función de algunas variables nominales escogidas por nosotros: Estado, ciudad, idioma, dominio de bajo nivel, niveles de ordinacion de la URLs.
A este conjunto de URLs muéstrales, se practico una análisis neuronal utilizando el algoritmo de Kohonen que nos arrojo los contextos textuales mayoritarios con el grado de asociación entre ellos y con las variables nominales. Al conjunto variables resultantes, se opero una análisis de los componentes principales con la proyección sobre los ejes de mayor peso de los contextos en el ambiente de variables nominales para poder asociar de una manera más eficaz cada contexto a sus características nominales.
En el sentido de esta análisis, tenemos la elección entre analizar los contextos fuera de sus alcances espaciales a través de los resultados arrojados por la clasificación de Kohonen o llevar los tratamientos a un nivel capaz de tomar en cuenta el origen geográfico del discurso involucrado en la operación.

[5] Comunicación personal del NIC-Mexico.

- Análisis del contexto textual fuera del alcance geográfico

El análisis del contexto textual sin la localización del origen del discurso, nos permite elaborar una idea del contexto contenido en las paginas de la WWW y de determinar las relaciones en ellos. Por concepto de contexto, se entienda una aproximación del discurso en las paginas a través de palabras claves. En efecto, en caso que no sabemos de manera exacta que esta desciendo el autor de la pagina, con el apoyo de palabras claves, podemos saber de que esta hablando.
En la figura.1, podemos apreciar el diagrama relacional de los diferentes contextos resultado del tratamiento neuronal sobre una muestra de la WWW mexicana. En esta relacion, se da una importancia a la estructuracion de los contextos agrupados en clases a fines de identificacion de los diferentes discursos por la frecuencia de aparicion y de asociacion entre las diferentes clases. Las clases en esta ejemplo están definidas a través de la noción de territorios y territorialidad en el discurso de la Web y muestran que su contenido no permite aun identificar de una manera muy precisa la naturaleza de los contextos, lo que nos va llevar a desarrollar una segmentación mas elevada.
A este tipo de segmentación, se puede dar un entorno geográfico a través de la búsqueda en un ambiente definido alrededor de un contexto espacial como el caso de la figura.2. En esta ejemplo, se da énfasis a una organización de las palabras claves a la construcción de los contextos alrededor del nombre del estado "Chiapas". Según la cercanía de las palabras claves a la palabra "Chiapas" y la estructuración de estas entre ellas, se da una organización semántica al contenido de la paginas Web. Los contextos así definidos por el usuario según su interés, están punteados por las palabras claves hacia las URLs de interés para focalizar los temas con un índice de presencia relacional. Los índices de presencia relacionales, una vez incorporados en un sistema de informacion "espacial", nos puedan arrojar los temas y sus URLs de correspondencia por cada localización. Estos sistemas de ordenamiento semántico, puedan constituir una alternativa de búsqueda de la informacion geográfica en la red con mucho mas organización, precisión y rapidez que los motores actuales.

- Análisis del contexto textual territorial

El análisis del contexto textual territorial consiste en tomar en cuenta la localización del discurso para la extracción de los contextos y sus relaciones. Así por cada localización, podemos deducir las lógicas inherentes y incorporarlas a un sistema espacial de análisis integral. El rastreo sistemático de las URLs, nos arroja a demás de la localización, la latencia y la trayectoria seguida para alcanzar el objetivo de la consulta, lo que nos puede llevar generar estadísticas localizadas sobre los dominios de bajo nivel y deducir las jerarquías en las redes de comunicación (vea figura.3).
En la figura.4, podemos ver un rastreo de la URLs (del sitio Web de Todito.com) a través de todo los nodos intermediarios entre nuestra computadora y el servidor objectivo con una informacion relativa a cada nodo. Estos rastreos operados desde diferentes localizaciones, nos dan la posibilidad de trazar la red de infraestructura de comunicación electrónica presente en una región determinada, nos permitan también levantar el origen geográfico de los contenidos textuales y legarlos en una sistema de informacion espacial.

La posibilidad de armar estas rutinas en una cadena de ejecución, nos lleva a poner en marcha un sistema de informacion en tiempo real que nos permite tener una idea espacial sobre los contenidos y sus interacciones en la red.

Conclusión:

La WWW es muy lejos de ser solamente una fuente de informacion, de interacción y de comunicación, es también un medio de investigación de la producción telemática de los actores localizados. La posibilidad que ofrecen actualmente los algoritmos de búsqueda y de clasificación semántica, nos permiten alcanzar la dominación de un numero importante de discursos y de terminar sus lógicas frente a otros factores del desarrollo que prevalecen en cada territorio.

La estructuración de los contextos textuales en la WWW, mas allá de facilitarnos un mejor desempeño en la navegación de la red, abra para los estudios espaciales un nuevo campo para el entendimiento del papel de la telemática en la probable estructuración del espacio bajo el impulso de las tecnologías de comunicación.

Pero aquí vale mencionar que la interacción en una red telemática como la WWW, el contenido textual constituye solamente una parte del discurso, hasta el momento no se han encontrado algoritmos capaces de tomar de una manera integral textos, imágenes y sonidos para alcanzar todos los entornos del discurso.

Actualmente, algunas empresas del software ofrecen ambientes de trabajo parciales apoyándose en las bases de datos en red, pero hasta el momento no se ha diseñado aun un sistema integral tomando en cuenta la herencia de la análisis espacial de redes y la contribución actual consístante en la arquitectura semántica para el manejo de grandes bloques de informacion.

En nuestra aplicación, hamos probado la estabilidad de los algoritmos utilizados en cadenas parciales apoyándonos a veces en software exístante. Frente al interés de realizar la totalidad de las operaciones en un solo paquete, podemos afirmar que las posibilidades son reales aprovechando la mejoría de la capacidad de calculo y la aportación en la inteligencia artificial.

Para nuestros intereses de geógrafos, la metodología desarrollada en esta trabajo, nos ha permitido generar una estadística académicamente confiable afines de analizar el fenómeno de la polarización de los medios de comunicación en México.

Bibliografía

Cox, T. F; Cox, Michael A. A. 1994. Multidimensional scaling. Monographs on statistics and applied probability; 59. Chapman & Hall. London.

Jolliffe, I. T. 1986. Pincipal component analysis. University of Geneva. Springer. New York.

Kruskal, J.B; Wish, M. 1981. Multidimensional scaling. Quantitative applications in the social sciences; 07-011. Sage Publications. Beverly Hills. London.

Li, S; Vel, O; Coomans, D. 1995. Comparative performance analysis of non-linear dimensionality reduction methods. URL:http://www.cs.jcu.edu.au/ftp/pub/techreports/ 94-8.ps.gz.

Kohonen, T. 1984. Self-organization and associative memory. Springer-Verlag. Berlin.

Kohonen, T; Hynninen, J; Kangas, J; Laaksonen, J. 1995. The self-organizing map program package Version 3.1. URL:ftp://cochlea.hut.fi/pub/som_pak/som_pak-3.1.tar.Z.

Telmex. 2000. Resultados releventes 4 trimestre de 1999. Direccion de finanzas y administracion-relaciones con invescionistas.

Figuras y imágenes acompañando él articulo:

Figura.1: Diagrama relacional de los diferentes contextos

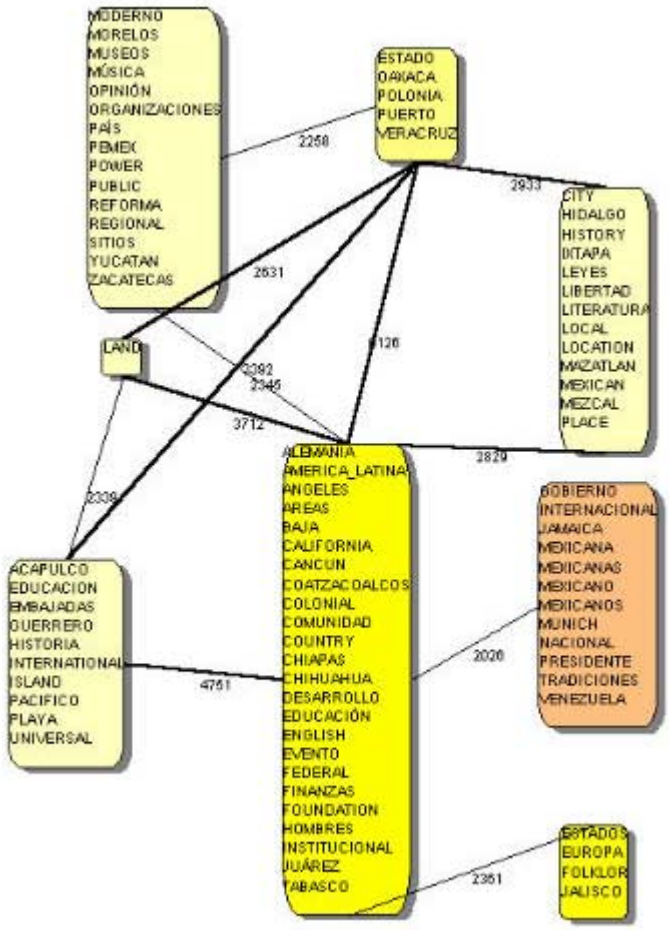

Figura.2: Mapeo de la organización de los contextos alrededor del nombre del estado de Chiapas.

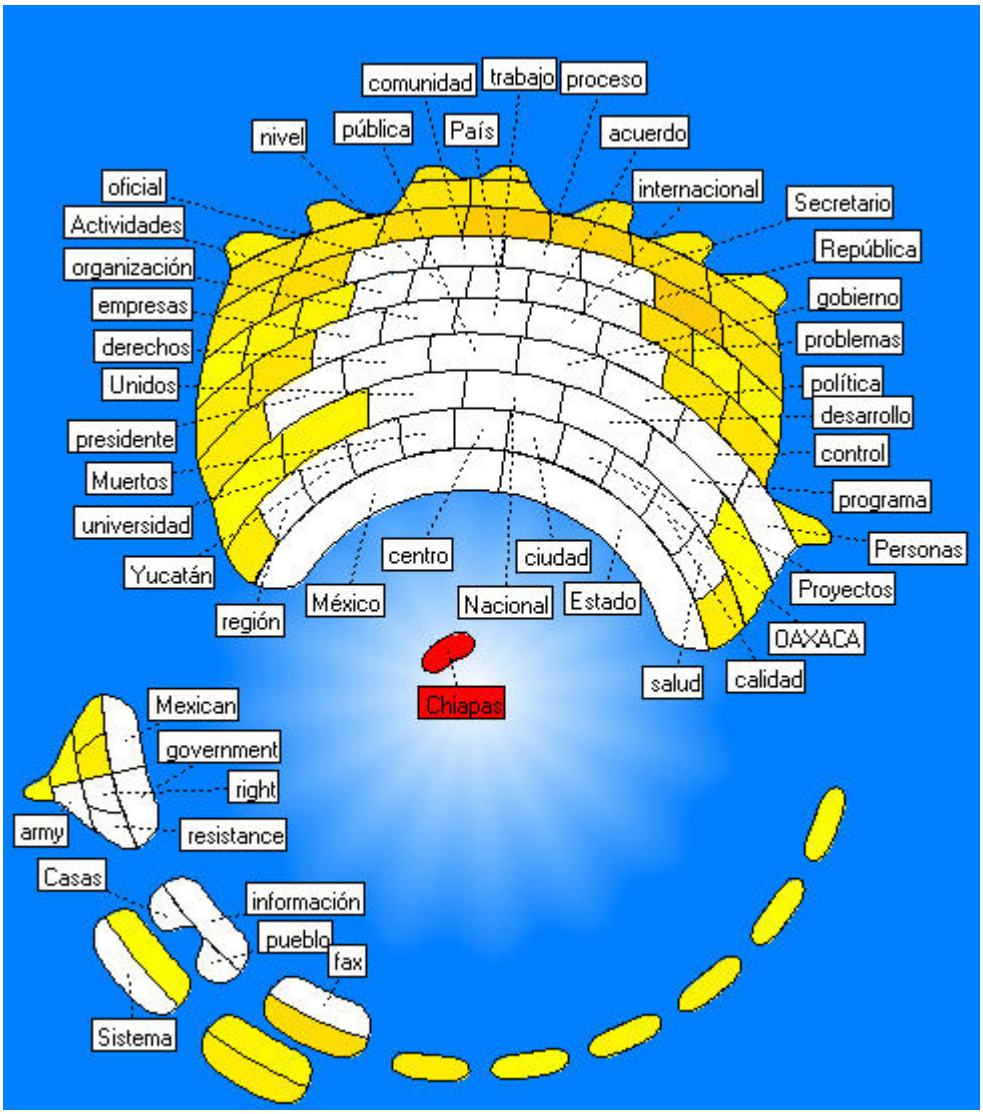

Figura.3: Estadísticas localizadas sobre los dominios de bajo nivel

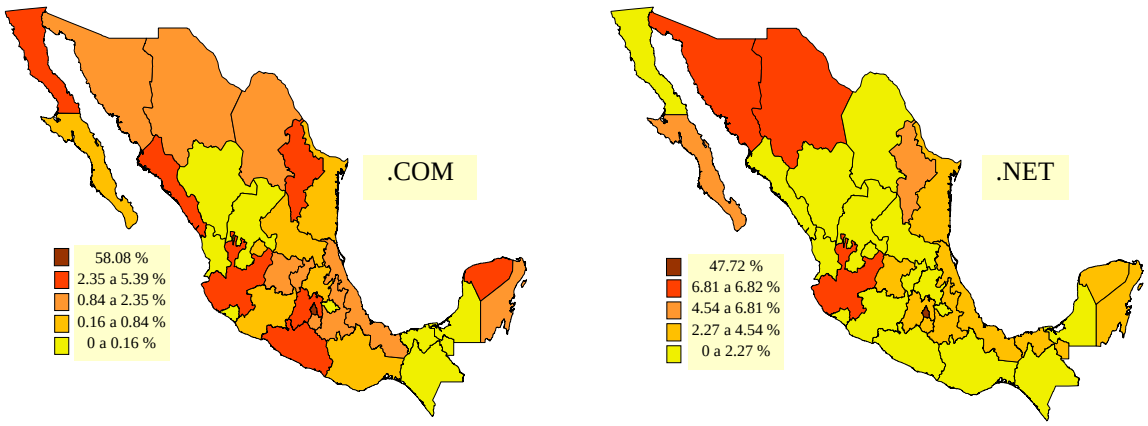

Figura.4: Rastreo de la URLs de Todito.com en un sistema de informacion en tiempo real