Investigación en Gobierno Digital

(Estructura, patrones y su cuantificación en interacciones de sitios Internet y de correo electrónico)

Por Carlos Javier Flores Saracho

Documento de Examen Predoctoral

En*tregado como requisito parcial para la aprobación del Examen Predoctoral del*
Doctorado en Ciencias en Desarrollo Científico y Tecnológico para la Sociedad
Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional
Noviembre de 2011

Resumen

En este trabajo se presenta un proyecto de investigación cuyos objetivos son visualizar la estructura e interacciones de los usuarios en los portales Internet y a través del correo electrónico, así como su cuantificación. El enfoque es una mezcla de herramientas de graficación por computadora, técnicas de análisis de redes sociales y aplicación de algoritmos matemáticos de minería de datos.

Con el incremento masivo de las comunicaciones sobre redes electrónicas, es imposible para una mente humana enfrentar la complejidad de la información sin auxiliarse de técnicas y herramientas que puedan proporcionar una "visualización" de las estructuras, que puedan servir para "descubrir" patrones, y a su vez puedan ayudar a cuantificar las grandes cantidades de "hechos electrónicos" en la Internet o en las computadoras a nuestro alcance.

Se explican en este documento los antecedentes y situación actual del problema; éste incluye los principales frentes de investigación activos tanto en el campo específico de investigación en el que se inscribe el tema de trabajo como en áreas afines (grupos líderes y sus publicaciones recientes); los principales puntos de controversia a la fecha y la argumentación científica de cada parte; y aquellos aspectos en donde incide el proyecto que se propone.

Asimismo este documento refiere el plan de trabajo de Investigación que consiste en describir el objetivo principal y metas parciales para alcanzarlo; la hipótesis de trabajo y su justificación; la metodología prevista y logística para desarrollarla (dónde, cómo, cuándo, con quién); el calendario de trabajo; los riesgos de viabilidad y por último se describen brevemente los resultados preliminares alcanzados a la fecha.

Se anexan la Bibliografía y Referencias citadas así como un listado de las revistas seleccionadas para la publicación de artículos.

Directores y Asesores

Dr. Gerardo Herrera Corral
Co-director del trabajo de tesis

Dr. Gerardo Hernández García
Co-director del trabajo de tesis

Dr. Miguel Ángel Pérez Angón
Asesor del trabajo de tesis

Dr. José Antonio Moreno Cadenas
Asesor del trabajo de tesis

Dr. Jesús Manuel Olivares Ceja
Asesor externo del trabajo de tesis (CIC-IPN)

Investigación en Gobierno Digital

(Estructura, patrones y su cuantificación en interacciones de sitios Internet y de correo electrónico)

Contenido

Archivo: PROYECTO de tesis de doctorado 27-nov-2011 v1.odt

Introducción

¿Qué relaciones se establecen y son más frecuentadas entre las personas y el Gobierno Digital por medios electrónicos? ¿Qué recursos de información son los más utilizados? ¿Cómo se relaciona una persona con un cierto grupo de personas?¿Qué grupo de personas pueden tener mayor influencia en una decisión por medios electrónicos?

Esta investigación va dirigida a resolver este tipo de preguntas.

Con el incremento masivo de las comunicaciones electrónicas se hace cada vez más complejo tratar de visualizar *qué usuario* se relaciona con *qué recurso de información*, qué *patrones* de comportamiento reflejan los usuarios, y qué medidas o métricas son aplicables a las interacciones en el Gobierno Digital.

El Análisis de Redes Sociales posee técnicas y herramientas para examinar las relaciones entre las personas, más precisamente, para analizar aquellos aspectos de la gente susceptibles de observarse y medirse. Los estudios de redes sociales típicos se enfocan en aspectos como *centralidad* (qué individuos están más conectados con otros o poseen mayor "influencia") y *conectividad* (si están conectados y cómo están conectados los individuos entre sí a través de la red social). Extenderemos este paradigma a las interacciones por Internet y por correo electrónico en el Gobierno Digital.

Específicamente, aplicaremos estas técnicas de análisis a las interacciones de los ciudadanos con el Gobierno Digital para tratar de determinar si existe en ellas *estructura*, si con la perspectiva de análisis y herramientas computacionales que utilizamos se pueden detectar *patrones* en las interrelaciones, y finalmente, si estos dos aspectos son susceptibles de *evaluarse* en base a un sistema propuesto de referencia, a fin de tener indicadores verificables y comparables.

Para demostrar si existen tanto *estructura* como *patrones*, utilizaremos técnicas de visualización por computadora en forma de grafos, y técnicas matemáticas plasmadas en algoritmos por computadora aplicados al conjunto de datos (análisis de agrupamientos -clusters-, análisis multivariado). Para la evaluación utilizaremos un esquema de referencia propuesto asignando medidas y ponderaciones a los datos.

Serán objeto de esta investigación dos casos de estudio: las estrategias de Gobierno Digital (Internet y Correo Electrónico) del Centro de Investigación en Computación del I. P. N. (CIC-IPN) y del Centro de Investigación y Estudios Avanzados del I. P. N. (CINVESTAV-IPN) ya que los datos necesarios para el análisis son accesibles. Otras instancias de gobierno digital pueden estudiarse con la perspectiva de análisis que proponemos.

Antecedentes y situación actual del problema

Cada vez más a todos los niveles de gobierno se utilizan las nuevas tecnologías de información y comunicaciones (TIC). Estas tecnologías han sido demostradas claramente en el sector privado con la tecnología del comercio electrónico ("eCommerce"). Ahora, con la disponibilidad de comunicaciones inalámbricas ubicuas (tecnología celular y "smartphones") y el uso creciente de las redes sociales (Facebook, Twitter, etc.), los gobiernos en todo el mundo se han montado en esta revolución y hacen esfuerzos para desarrollar en el sector público, cada vez más, el denominado Gobierno Digital (Gobierno Electrónico; "eGovernment", "Digital Agenda"). El gobierno digital será entendido en esta investigación como aquella parte del gobierno de una nación, estado o municipio que hace uso de las TIC para servir y comunicarse con los ciudadanos o usuarios.

Según la ONU en su reporte de 2010 [UNO, 2010], pp. 115, ciento ochenta y cuatro (184) de los ciento noventa y dos (192) países miembros han implantado estrategias de Gobierno Digital. Este crecimiento dinámico de la oferta de servicios y de información vía Internet y por tecnologías digitales asociadas, puede deberse a que el gobierno digital tiene el potencial para cambiar el gobierno tradicional "de ventanillas" hacia uno de mejor acceso y mayor disponibilidad en los servicios de gobierno. En este ambiente nuevo de intercambio electrónico, los usuarios cuentan con mayores facilidades para interactuar con el Gobierno y buscar información sin limitaciones de espacio y tiempo.

A la fecha, prácticamente todos los niveles de gobierno en México (federal, estatal y municipal) son accesibles vía Internet y proporcionan una variedad de información y servicios en línea.

Este fenómeno socio-tecnológico del Gobierno Digital ha cambiado nuestras costumbres y hábitos, nuestra forma de relacionarnos y nuestras actitudes. Tiene impacto creciente en todas las esferas de la actividad humana pudiéndose afirmar que estas tecnologías llegaron para quedarse.

En años recientes, prácticamente de los años noventas del siglo XX a la fecha, se han realizado innumerables investigaciones de sistemas interconectados en el campo de la Internet, las redes sociales, y las redes biológicas, por citar algunas áreas y, concurrentemente, se han desarrollado una variedad de técnicas para tratar de entender y predecir la estructura y el comportamiento de dichos sistemas o redes.

Entre las múltiples técnicas que destacan, algunas de las cuales utilizaremos en esta investigación, se encuentran las siguientes:

(1) Visualización de Información por Computadora (por mapas, gráficas, y modelos tridimensionales).
(2) Minería de la Red Internet ("Minería Web"; Web Mining).
(3) Minería de Datos (Data Mining).
(4) Análisis de Redes Sociales por computadora.

Una representación común de información es en la forma de grafos o gráficas, mediante recursos como puntos, líneas, flechas, colores, superficies y símbolos, para que se manifieste visualmente la relación que guardan entre sí; sirven para analizar el comportamiento de un proceso, o la interpretación de un fenómeno. La representación gráfica, particularmente cuando el conjunto de datos es grande, permite establecer valores que no han sido obtenidos experimentalmente, por interpolación o extrapolación, o por agrupamientos.

<Los orígenes tempranos de la "visualización de información" pueden remontarse a

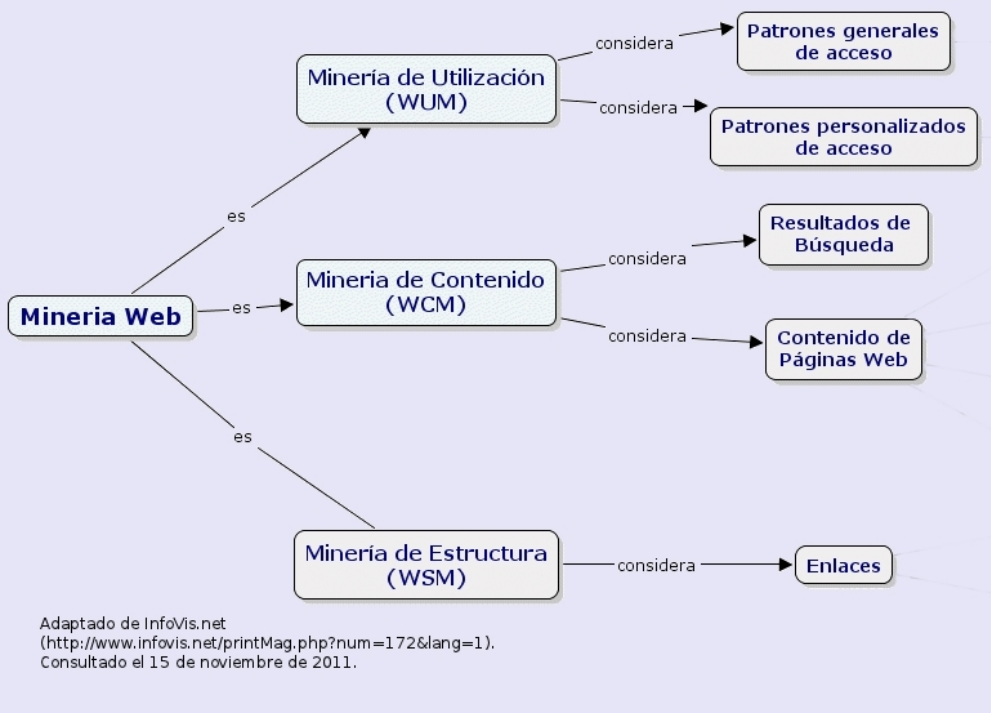
la construción de diagramas geométricos, mapas de las posiciones de las estrellas u otros cuerpos celestiales, y mapas de ayuda a la navegación y a la exploración. Actualmente, la *visualización de información auxiliada por computadora* se entiende como el estudio interdisciplinario de la "representación visual" de colecciones de grandes volúmenes de información como redes y relaciones en la Internet, bases de datos bibliográficas, programas y líneas de código en sistemas de software>[Friendly, 2008], así como datos médicos obtenidos en pacientes con aparatos de barrido digital, por citar algunas fuentes de información.

"En trabajos recientes, los enfoques más utilizados presentan metodologías para poder extraer información de los sistemas [de Internet] en línea [Matsuo, 2006], [Mika, 2005] para posteriormente obtener una red que represente lo más posible a los individuos y su forma de interactuar entre ellos"; [Mejía, 2010], p. 2.

Grandes volúmenes de información se prestan mayormente al análisis vía computadora. En esta investigación utilizaremos solamente información de las interacciones por Internet y por correo electrónico, como se señaló anteriormente por consideraciones de acceso a los datos.

La minería de la Red Internet es la extracción de información potencialmente útil en dicha red de redes. Se asocian generalmente tres diferentes enfoques [Bharanipriya, 2011]:

- Minería de Utilización de la Red (Web Usage Mining), que es el proceso de extraer patrones de interés en las bitácoras de acceso a los recursos de la Red (sitios Web, servidores de correos, etc.).
- Minería de Contenidos de la Red (Web Content Mining), que consiste en extraer información a partir del contenido de documentos en la Red o su descripción, y
- Minería de Estructuras de la Red (Web Structure Mining), que es el proceso de inferir conocimiento a partir de las relaciones que se establecen entre referencias (fuentes o recursos de información en la red) y referentes (ligas a dichos recursos).



En esta investigación utilizaremos los tres tipos de minería para el análisis de

interacciones de usuarios vía Internet y por correo electrónico.

Para la búsqueda de estructuras y patrones (posiblemente) presentes en el caudal de información por lo general se utilizan variantes de 10 algoritmos que son los mayormente utilizados [Wu, 2008], p. 6. Estos algoritmos son los siguientes: *C4.5, K-Means, SVM-Support Vector Machines, Apriori, EM, PageRank, AdaBoost, kNN: k-Nearest Neighbors, Naive Bayes.*

Cada uno de estos algoritmos permite extraer información no trivial. Para descubrir agrupamientos ("clusters") en la información utilizaremos, entre otros, el algoritmo "K-Means". En estadística y minería de datos el algoritmo para agrupamientos "K-Means", es un método de análisis que tiene como objetivo hacer particiones de *n* observaciones (datos) en *k* agrupamientos, donde cada observación (dato) pertenece al agrupamiento con la media más cercana. El algoritmo "K-Means" es un método iterativo simple para dividir un conjunto de datos en un número predeterminado de agrupamientos, *k*. Este algoritmo fue descubierto por diferentes investigadores en diversas disciplinas, destacadamente en el campo de las telecomunicaciones y la teoría de la información, la estadística, y la inteligencia artificial.

El Análisis de redes sociales, relacionado con la teoría matemática de redes, ha surgido como una metodología adiconal en las modernas ciencias sociales. También se ha utilizado para estudiar ecosistemas en Biología y diversos fenómenos en la Física y en la Química.

El Análisis de redes sociales ha pasado de ser sólo una metáfora sugerida por Jacob Levy Moreno [Moreno, 1978] en la fundamentación de la sociometría (en 1953), para constituirse en un enfoque analítico y un paradigma, con sus principios teóricos, métodos y líneas de investigación propios. Los analistas de redes sociales aducen que pueden estudiar el efecto producido por la acción selectiva de los individuos en la red; desde su estructura las relaciones, el individuo o el grupo, hasta diversos atributos supuestamente cuantificados como el comportamiento o la actitud.

Cuando las redes sociales son enormes el apoyo computacional no sólo es necesario sino indispensable. Aplicaremos en la investigación este enfoque de análisis a las redes sociales.

Principales frentes de investigación activos, tanto en el campo específico de investigación en el que se inscribe el tema de trabajo como en áreas afines (grupos líderes y sus publicaciones recientes)

En el campo específico de investigación de la *Visualización de Información por Computadora* aplicada a Internet y Correo Electrónico, el enfoque de esta investigación es hacia la *visualización de redes de interacción*, desde la perspectiva del *Análisis de Redes Sociales*, y donde los usuarios y y los *recursos de información* son las *entidades* de interés (representados como "nodos") y las *relaciones* son las interacciones que se producen entre los usuarios y/o los recursos de información en el contexto del gobierno digital (representados como "líneas").

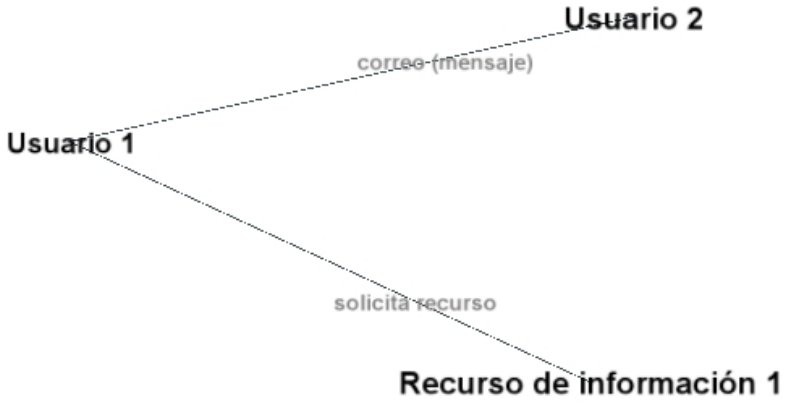


Figura. Entidades (nodos) y relaciones (líneas) para la visualización de interacciones entre usuarios (por correo) y entre usuarios y recursos de información (por Internet).

Esta área de investigación es especialmente activa con desarrollos de programas de visualización estimulados en particular por las grandes empresas en la Internet como Google Inc., y Yahoo Inc., con propósitos comerciales principalmente, y por los grandes laboratorios de investigación en telecomunicaciones, como HP Research Lab, con propósitos académicos, de investigación y también comerciales.

La siguiente tabla representa algunos grupos de investigación activos y los programas de visualización desarrollados por dichos grupos <Adaptada de [Huisman, 2011]>.

Tabla 1. Grupos de investigación y programas de Visualización de Información por Computadora y referencias bibliográficas recientes

Grupo de Investigación / Desarrollador	Nombre del programa	Utilización	Licencia	Referencia bibliográfica reciente
Vladimir Batagelj. Department of Mathematics, FMF University of Ljubljana, Slovenia Andrej Mrvar . Faculty of Social Sciences University of Ljubljana, Slovenia	**Pajek**	Redes en general. Redes Sociales	Académica / Gratuita	[de_Nooy, 2005]
Compañía privada alemana AbsInt Angewandte Informatik GmbH	**aiSee**	Visualización de gráficas	Comercial / Académica / Demostrativo Gratuito	[Douglas, 2005]
Cytoscape Consortium con fondos de: National Center for Research Resources (NCRR) , National Resource for Network Biology (NRNB) y el National Institute of General Medical Sciences (NIGMS) del National Institutes of Health (NIH), todos de E. U. A.	**Cytoscape**	Visualización de interacción en redes moleculares	Académica / Gratuita	[Lopes, 2010] [Yeung, 2008]
Consorcio Gephi (Francia)	**Gephi**	Plataforma de visualización y exploración	Académica / Gratuita	[Bastian, 2009] [Blondel, 2008]
ATT Research Laboratories	**Graphviz**	Visualización de gráficas	Académica / Gratuita	[Ellson, 2001]
Elbirt Technologies (compañía privada)	**Jacob's Ladder**	Animación de datos multidimensional, visualización	Académica / Gratuita	[Elbirt, 2007]
Carnegie Mellon - Heinz School	**KrackPlot**	Visualización de redes sociales	Académica / Gratuita	[Krackhardt, 1994b]
Duke University	**KineMage**	Despliegue de vectores tridimensionales con cinemática gráfica	Académica / Gratuita	[Chen, 2009]
Analytic Technologies (compañía privada)	**NetDraw**	Gráficación de redes	Académica / Gratuita	[Chen, 2005]

Grupo de Investigación / Desarrollador	Nombre del programa	Utilización	Licencia	Referencia bibliográfica reciente
Alemania: Universidad Técnica de Dormund; Universidad Friedrich-Schiller de Jena; Universidad de Colonia, y compañía privada oreas GmbH	**OGDF (sucesor del AGD)**	Plataforma abierta de dibujo de gráficas	Académica / Gratuita	[Chimani, 2007]
Cooperative Association for Internet Data Analysis, con base en el Centro de Super-Cómputo de la Universidad de California en San Diego (UCSD)	**Otter**	Herramienta de despliegue topológico	Académica / Gratuita	[Huffaker, 1999]
Dan McFarland y Skye Bender-deMoll, con fondos del Research Incentive Award de la Universidad de Stanford (Oficina de Licenciamiento de Tecnología)	**SoNIA**	Visualización longitudinal de datos de redes	Académica / Gratuita	[Bender-deMoll, 2006]
David Auber (autor original) y Patrick Mary LaBRI de la Universidad de Burdeos, Francia	**Tulip**	Visualización de gráficas grandes	Académica / Gratuita	[Archambault, 2007] [Auber, 2006]
Grupo Bernd Krieg-Brückner's de la Universidad de Bremen, Alemania	**uDrawGraph) (sucesor de daVinci)**	Dibujo de gráficas	Académica / Gratuita	[Würthinger2007]
Prith Banerjee, Vicepresidente Senior de los Laboratorios de Investigación Hewlett-Packard	**Zoomgraph**	Visualización de gráficas y herramienta de acercamiento (zoom)	Académica / Gratuita	[Hao, 2001]
TouchGraph, LLC (compañía privada) subsidiaria de Google Inc.	**TouchGraph**	Visualización de información en general	Comercial / No gratuita	[Eppstein, 2010]

Adicionalmente, por sus capacidades de capturar directamente de la Internet la red de nodos y relaciones en la estructura de un sitio Web se utilizará el software siguiente:

Tabla 2. Grupos de investigación y programas de Visualización de Información por Computadora (adicional)

Grupo de Investigación / Desarrollador	Nombre del programa	Utilización	Licencia
Dimitris V. Kalamaras, investigador independiente	**SocNetV**	Análisis y visualización de redes sociales	*Académica / Gratuita*

Se dispone de revistas especializadas para este campo, entre ellas, *Information Visualization*, de SAGE Publishing.

Para la Minería de la Red Internet algunos grupos de investigación activos y los programas desarrollados por dichos grupos se presentan en la siguiente tabla.

Tabla 3. Grupos de investigación y programas para la Minería de la Red Internet

Grupo de Investigación / Desarrollador	Nombre del programa	Utilización	Licencia
Google Inc.	**Google Analytics**	Estadísticas de utilización de sitios web (tráfico del sitio y efectividad de mercado)	Comercial / Académica / Demostrativo Gratuito
Compañía privada Alexa Internet, subsidiaria de Amazon.com	**Araña Web** **(sin nombre específico de alexa.com)**	Estadísticas de utilización de sitios web (tráfico del sitio y efectividad de mercado)	Comercial / Académica / Demostrativo Gratuito
Tor Egge, inicialmente en su trabajo de doctorado en la Universidad Noruega de la Ciencia.	**AllTheWeb**	Buscador Web	(Nota.- Buscador descontinuado por la empresa Yahoo al igual que muchos más que fueron descontinuados por el éxito del buscador Google)

La Minería de Contenidos de la Red [Feinerer, 2008] identifica información de interés a partir de los contenidos, datos o documentos en la Red Internet. Para el tratamiento y análisis de tal inormación se utilizan dos enfoques principales. (1) Minería de textos no-estructurados (textos en formato libre, "free texts") y (2) minería de textos estructurados o semi-estructurados (datos en formato de tablas, y datos semi-estructurados como documentos en formato de hipertexto).

En esta investigación utilizaremos la minería de textos al analizar el contenido de mensajes de correo electrónico y el contenido de las interacciones de un usuario con un sitio de Internet cuando el usuario específicamente proporciona información (no navega en las páginas únicamente, sino que envía datos asociados a su persona <"user-id", o identificación de usuario> o asociados al objeto de su interés dentro del sitio <"*queries"*, o consultas>).

La minería de textos [Feinerer, 2008], "abarca un campo de métodos y técnicas teóricos con una cosa en común: el texto como información de entrada... En general, la minería de textos es un campo interdisciplinario de actividad [donde se aplica la] lingüistica, la estadística computacional, y las ciencias de la computación. Las técnicas estándar incluyen clasificación, agrupamiento de textos ["text clustering"], creación de ontologías y taxonomías, [y] resumen de documentos..."

La siguiente tabla representa los programas de minería de textos de uso Académico / Gratuito más conocidos, que también tienen sus versiones comerciales:

Tabla 4. Grupos de investigación y programas para la Minería de Textos

Grupo de Investigación / Desarrollador	Nombre del programa	Utilización	Licencia
Hamish Cunningham, Kalina Bontcheva y una serie de colaboradores de la Universidad de Sheffield, Inglaterra	**GATE**	Minería de textos y un ambiente gráfico de desarrollo	*Académica / Gratuito / Comercial*
Alias-i, compañía privada fundada por el desarrollador Breck Baldwin	**LingPipe**	Procesamiento de texto usando linguistica computacional	*Académica / Gratuito / Comercial*
Calais, una compañía subsidiaria de Thompson-Reuters	**OpenCalais**	Programa para incluir funcionalidad semántica en blogs, sitios web, sistemas demanejo de contenidos y aplicaciones	*Académica / Gratuito / Comercial*
Rapid-I GmbH, compañía alemana.	**RapidMiner**	Minería de textos y minería de datos	*Académica / Gratuito / Comercial*
Anthony Fader y un grupo de colaboradores de la Universidad del Centro Turing de Washington, E. U. A.	**ReVerb**	Extracción de información de la Red Internet.	*Académica / Gratuito*
Bing Liu, Xiaoli Li y una serie de colaboradores independientes	**LPU**	Sistema de aprendizaje a partir de textos y clasificación	*Académica / Gratuito*
Un grupo de linguistas y científicos en computación del Colegio Midlebury, de Vermoont, E. U. A.	**The Semantic Indexing Project**	Técnicas para la búsqueda de patrones y para la organización de colecciones de datos	*Académica / Gratuito*
The R Project for Statistical Computing	**"tm" de R**	Minería de textos	*Académica / Gratuito*

Para la minería de datos los grupos de investigaciones y la oferta de programas es enorme. En la siguiente tabla se presenta el grupo de investigación y el programa desarrollado a utilizar en esta investigación.

Tabla 5. Grupo de investigación y programa para la Minería de Datos

Grupo de Investigación / Desarrollador	Nombre del programa	Utilización	Licencia
Desarrolladores de la Universidad Waikato de Australia	**Weka**	• Minería de Datos y Algoritmos de aprendizaje por máquina. • Preprocesamiento de datos, clasificación, regresión, agrupameinto, reglas de asociación y visualización	*Académica / Gratuito*

La siguiente tabla presenta algunos grupos de investigación y los programas para Minería de Estructuras de la Red (Web Structure Mining), que incluye arañas de Internet

(Web Spiders, Web Crawlers), con capacidades de capturar directamente de Internet la estructura de un sitio Web, entre otras:

Tabla 6. Grupos de investigación y programas para la Minería de Estructuras de la Red Internet

Grupo de Investigación / Desarrollador	Nombre del programa	Utilización	Licencia
Free Software Foundation	**GNU wget**	Creación de espejos de sitios web (copia de sitios)	*Académica / Gratuita*
Dirk Stoecker, analista independiente	**Pavuk (Linux)**	Capturador de sitios web multifuncional, con descargas recursivas basadas en los enlaces de documentos HTML	*Académica / Gratuita*
Mike Sutton de la compañía privada Solent Software	**PageNest offline browser**	Capturador de sitios web (para Windows)	*Comercial / Académica / Demostrativo Gratuito*
Xavier Roche y otros desarrolladores independientes	**HTTrack website copier (Windows)**	Capturador de sitios web	*Académica / Gratuita*
Jan Erik Samsonsen y otros desarrolladores noruegos de la compañía privada Calluna Software	**Webripper 2.0**	(para Windows plataforma .NET)	*Académica / Gratuita*

En esta sección se han identificado grupos de investigación y programas computacionales desarrollados para *Visualización de Información por Computadora, Minería de la Red Internet , Minería de Textos, Minería de Datos, y Minería de Estructuras de la Red Internet, respectivamente.*

Principales puntos de controversia a la fecha y argumentación científica de cada parte

*Tabla 7. Principales puntos de controversia: **Computación y visualización por computadora***

Principales puntos de controversia	Argumentación científica a favor	Argumentación científica en contra
Computación y visualización por computadora: "La ciencia computacional [y la computación aplicada a las ciencias] en general, hasta los años 90's del siglo XX, era percibida por la mayoría de los científicos y físicos solamente como meras herramientas y cálculos numéricos (number crunching), y no como sujetos de estudio por su propia cuenta." [Hjorth_Jensen, 2003]	[Hjorth_Jensen, 2003]: "[El uso de la computación] es una parte importante de la teoría y del experimento. La habilidad para hacer computación es ahora parte del repertorio esencial de los investigadores científicos... Campos nuevos han emergido y han reforzado sus posiciones en los últimos años, tales como ciencia de materiales computacional, bioinformática, mecánica y matemática computacional, química y física computacional, etc." [Lenoir, 1999]: "Los experimentos computacionales atraen juntos la teoría y el experimento de laboratorio de formas completamente nuevas. La nueva ciencia computacional ligada a la visualización tiene un gran efecto en los campos de la bioquímica, dinámica molecular, y farmacología molecular. Los enfoques computacionales han transformado y extendido sustancialmente el dominio de la teorización en estas áreas de maneras no disponibles antes de la era computacional."	[Harrison, 2010]: "Una de las críticas más comunes a los experimentos de laboratorio es que no aplican al mundo real. La misma crítica puede elevarse contra los experimentos en el mundo virtual. Es más, una crítica socorrida es que los resultados del mundo virtual no se pueden generalizar en el laboratorio."

*Tabla 8. Principales puntos de controversia: **Análisis de redes sociales***

Principales puntos de controversia	Argumentación científica a favor	Argumentación científica en contra
Análisis de redes sociales: la investigación de redes (sociales) no tiene un marco teórico.	[Borgatti, 2003]: "Diversos autores teorizan sobre las redes sociales extensamente, [Barnes, 1972]; [Granovetter, 1979]; [Rogers, 1987]."	-[Salancik, 1995], p. 348, argumenta que el análisis de redes (sociales) es poderosamente descriptivo pero que no es teórico.

Aspectos en donde incide el proyecto que se propone

Esta investigación es relativa al campo particular de la caracterización y evaluación del Gobierno Digital, específicamente en el estudio de las comunicaciones (interacciones) de los usuarios por medio de Internet (visitas a sitios Internet de gobierno) y por correo electrónico, cuando aquellos son atendidos por los funcionarios o autoridades.

En este trabajo de investigación se pondrán a prueba las herramientas de visualización de información, minería de datos y de textos, así como minería de la Red Internet, con el objetivo específico de obtener, a partir de la complejidad de los decenas de miles de datos de servidores de Internet y de correo electrónico, pautas de su estructura, descubrir patrones, si estos existen (o son asequibles por las herramientas de análisis utilizadas), y de cuantificar esas comunicaciones. Con el trabajo de investigación se pretende obtener los siguientes resultados tangibles:

- Una perspectiva de análisis que incluye herramientas computacionales dada la naturaleza de los datos bajo estudio.

- Un marco de referencia para avanzar en el estudio y cuantificación de las interacciones de usuarios con sitios de Internet y en redes sociales, éste último para el caso específico del correo electrónico.

Por estas razones, se afirma que el proyecto de investigación aquí descrito incide en los campos de visualización de información por computadora; análisis de redes sociales; minería de datos y de textos; y minería de la Red Internet, por lo que los artículos y los trabajos para Congresos, resultado de esta investigación, presentan favorecedoras perspectivas y diversidad de contenidos para su publicación.

Plan de trabajo

Objetivos principales y metas

Objetivo 1.- Investigación y experimentación sobre la estructura, patrones y cuantificación de las ***interacciones del sitio de Internet*** del Centro de Investigación en Computación del IPN (CIC-IPN): www.cic.ipn.mx

Meta 1.1: Obtención de los datos
Meta 1.2: Determinación de estructura
Meta 1.3: Determinación de patrones
Meta 1.4: Cuantificación de interacciones
Meta 1.5 Interpretación de resultados

Objetivo 2.- Investigación y experimentación sobre la estructura, patrones y cuantificación de las ***interacciones de CORREO*** en el Centro de Investigación en Computación del IPN (CIC-IPN)

Meta 2.1: Obtención de los datos
Meta 2.2: Determinación de estructura
Meta 2.3: Determinación de patrones
Meta 2.4: Cuantificación de interacciones
Meta 2.5 Interpretación de resultados

Objetivo 3.- Investigación y experimentación sobre la estructura, patrones y cuantificación de las ***interacciones del sitio de Internet*** del Centro de Investigación y Estudios Avanzados del IPN: www.cinvestav.mx

Meta 3.1: Obtención de los datos
Meta 3.2: Determinación de estructura
Meta 3.3: Determinación de patrones
Meta 3.4: Cuantificación de interacciones
Meta 3.5 Interpretación de resultados

Objetivo 4.- Investigación y experimentación sobre la estructura, patrones y cuantificación de las **interacciones de CORREO** en el Centro de Investigación y Estudios Avanzados del IPN

Meta 4.1: Obtención de los datos
Meta 4.2: Determinación de estructura
Meta 4.3: Determinación de patrones
Meta 4.4: Cuantificación de interacciones
Meta 4.5 Interpretación de resultados

Objetivo 5.- Validación de los resultados por un panel de expertos del CINVESTAV Y DEL CIC-IPN.

Meta 5.1: Validación de resultados del CIC-IPN
Meta 5.2: Validación de resultados del CINVESTAV-IPN

Objetivo 6.- Publicación de investigación y resultados en revistas de corriente principal y en Congresos.

Meta 6.1: Elaboración de artículo para revista
Meta 6.2: Autorización de publicación de artículo
Meta 6.3: Elaboración de ponencia para Congreso
Meta 6.4: Envío de ponencia

Hipótesis de trabajo y justificación

Hipótesis de trabajo

La hipótesis principal de trabajo es la siguiente:

"En las interacciones de los usuarios o ciudadanos con el gobierno digital por medios electrónicos, existen estructuras y patrones que se pueden caracterizar y cuantificar"

Justificación de la investigación.

La justificación del trabajo de investigación se basa en los elementos de valoración siguientes:

(1) Las nuevas formas de gobierno digital o electrónico tienen implicaciones sociales, económicas, políticas y de gobernabilidad.

(2) La tecnología del gobierno digital (redes sociales; redes de investigación; portales de gobierno electrónico; sistemas de transacciones electrónicas), afecta e influye en prácticamente todos los ámbitos de la sociedad.

(3) Es importante caracterizar y evaluar un sistema de gobierno digital, entre otras formas como se describe en esta investigación.

(4) Estudiar un fenómeno tecnológico-social como es el gobierno electrónico o digital deriva en conocimiento práctico y teórico de utilidad.

Metodología prevista

1.- Investigación de Bibliografía.

2.- Elaboración y conclusión del Marco Conceptual.

3.- Obtención de datos.

4.-Experimentacion con el uso de herramientas informáticas y algoritmos matemáticos para comprobar o refutar la hipótesis (en dos casos de estudio).

5.- Interpretacion de resultados.

6.- Generacion de conclusiones.

7. Validación de conclusiones por expertos.

Logística para el desarrollo de la investigación

Tabla 9. Logística principal para el desarrollo de la investigación

Actividad	Tipo de actividad	Lugar	Apoyos	Consultas
INVESTIGACION DE BIBLIOGRAFIA	De gabinete	CINVESTAV-IPN CIC-IPN	BIBLIOTECAS Y ACCESO A REVISTAS ESPECIALIZADAS	-
ELABORACION DEL MARCO CONCEPTUAL.	De gabinete	CINVESTAV-IPN CIC-IPN	BIBLIOTECAS Y ACCESO A REVISTAS ESPECIALIZADAS	A Co-Directores y Asesores de Tesis
OBTENCION DE DATOS	De campo	CINVESTAV-IPN CIC-IPN	Archivos electrónicos con los datos de interacciones de Internet y con correos del servidor. Firma de autorizaciones.	Autoridades y administradores de CINVESTAV-IPN y CIC-IPN
EXPERIMENTACION CON EL USO DE HERRAMIENTAS INFORMATICAS Y ALGORITMOS MATEMÁTICOS PARA COMPROBAR O REFUTAR LA HIPOTESIS (EN LOS DOS CASOS DE ESTUDIO).	De gabinete	CINVESTAV-IPN CIC-IPN	HERRAMIENTAS YA DISPONIBLES (OPEN SOURCE). ALGORITMOS YA DISPONIBLES (SOFTWARE WEKA)	-
	De campo	CINVESTAV-IPN CIC-IPN		A Co-Directores y Asesores de Tesis para verificar datos vs. resultados A expertos del CIC y del CINVESTAV para verificar datos vs. resultados y la utilización correcta de las herramientas
INTERPRETACION DE RESULTADOS.	De gabinete	CINVESTAV-IPN CIC-IPN	-	A Co-Directores y Asesores de Tesis para verificar interpretación corecta
	De campo	CINVESTAV-IPN CIC-IPN	-	A expertos del CIC y del CINVESTAV para verificar interpretación corecta
GENERACION DE CONCLUSIONES.	De gabinete	CINVESTAV-IPN CIC-IPN	-	A Co-Directores y Asesores de Tesis para verificar conclusiones
	De campo	CINVESTAV-IPN CIC-IPN	-	A expertos del CIC y del CINVESTAV para verificar conclusiones

Calendario de Trabajo.

(VER ANEXO 1)

Riesgos de viabilidad

La siguiente tabla resume los riesgos de viabilidad del proyecto.

Tabla 10. Resumen de riesgos de viabilidad del proyecto

Actividad	Riesgo	Observaciones
Obtención de datos de sitio web (estructura del SITIO WEB CIC-IPN)	-	No existe riesgo. Los datos son públicamente accesibles.
Obtención de datos de sitio web (estructura del SITIO WEB CINVESTAV-IPN)	-	No existe riesgo. Los datos son públicamente accesibles.
Obtención de datos de sitio web (interacciones de usuarios en el portal CINVESTAV-IPN)	-	No existe riesgo. La entrega de los datos ha sido autorizada por escrito describiendo y delimitando la utilización de la información con fines exclusivos de investigación
Obtención de datos de sitio web (interacciones de usuarios en el portal CIC-IPN)	Los administradores de los sitios y las autoridades pueden argumentar "confidencialidad" de los datos.	Se hará una petición formal por escrito describiendo y delimitando la utilización de la información con fines exclusivos de investigación
Obtención de datos de interacciones de usuarios por correo electrónico en el CIC-IPN	El administrador del servidor de correo del puede argumentar "confidencialidad" de los datos.	Se hará una petición formal por escrito describiendo y delimitando la utilización de la información con fines exclusivos de investigación
Obtención de datos de interacciones de usuarios por correo electrónico en el CINVESTAV-IPN	El administrador del servidor de correo del puede argumentar "confidencialidad" de los datos.	Se hará una petición formal por escrito describiendo y delimitando la utilización de la información con fines exclusivos de investigación

Resultados preliminares

1.- Se han obtenido las herramientas computacionales siguientes:

WGet- para la copia remota del sitio de Internet del CIC y del CINVESTAV. (Copia en computadora de todo el sitio Web respectivo)

SocNetV - para la obtención local y remota del mapa de relaciones entre páginas de sitios de Internet del CIC y del CINVESTAV.

Gephi - para la edición y construcción de mapas de sitios Web.

Weka - para la minería de datos de Internet y correos.

R - para la minería de datos de Internet y correos.

2.- Se ha experimentado con las herramientas mencionadas haciendo mapas de interacciones Internet del CINVESTAV y mapas de correos con datos obtenidos de otras fuentes (Lista de correos pública de R-development).

3. Se han escrito las líneas generales para la publicación del primer artículo en Inglés.

4. Se ha solicitado un nuevo lote de datos de interacciones Internet a la Coordinación General de Tecnologías de Información CGSTIC del CINVESTAV.

Nota.- El avance de la investigación será mostrado en la presentación del examen PREDOCTORAL en forma de gráficas y tablas actualizadas.

Revistas seleccionadas para la publicación de artículos

The Information Society, An International Journal. ISSN: 1087-6537 (electronic) 0197-2243 (paper). *Publication Frequency: 5 issues per year. Publisher: Routledge.*

Electronic Government, an International Journal (EG). *ISSN (Online): 1740-7508 - ISSN (Print): 1740-7494.*

International Journal of Electronic Governance (IJEG). ISSN (Online): 1742-7517 - ISSN (Print): 1742-7509. Published in 4 issues per year.

Government Information Quarterly. An International Journal of Information Technology Management, Policies, and Practices. *Impact Factor: 2.098; 5-Year Impact Factor: 2.255. Issues per year: 4.*

Information Technology for Development. (Wiley Online Library). *Online ISSN: 1554-0170.*

International Journal of Electronic Government Research (IJEGR). *An Official Publication of the Information Resources Management Association. DOI: 10.4018/IJEGR; ISSN: 1548-3886; EISSN: 1548-3894.*

Transforming Government: People, Process and Policy. ISSN: 1750-6166.

The Journal of E-Government Studies and Best Practices (JEGSBP). ISSN 2155-4137. *The Journal of E-Government Studies and Best Practices is an international peer reviewed and applied research journal which accepts contributions that are based on original research, relevant studies, best practices, case studies, and real-world experiences.*

Policy Studies Journal. Published on behalf of the Policy Studies Organization and the Public Policy Section of the *American Political Science Association. Edited by: Peter deLeon School of Public Affairs, University of Colorado Denver and Chris Weible School of Public Affairs, University of Colorado Denver. Impact Factor: 0.574.*
Information Polity. International Journal of Government & Democracy in the *Information Age. (Página:* http://www.iospress.nl/loadtop/load.php?isbn=15701255*).*

International Journal of Business and Management, peer-reviewed journal, published by Canadian Center of Science and Education. The journal publishes research papers in the fields of business, management, marketing, finance, economics, human resource management and relevant subjects. The journal is published in both printed and online versions. The online version is free access and download. http://www.ccsenet.org/journal/index.php/ijbm

The Electronic Journal of e-Government (EJEG), publishes research on topics relevant to the design, evaluation, implementation and management of e-Government/e-Governance, e-Democracy, e-Participation and other dimension of this field of study. Sitio http://www.ejeg.com.

Communications of the Association for Information Systems, (http://aisel.aisnet.org/cais/). Editor-in-Chief: Ilze Zigurs, University of Nebraska at Omaha. ISSN: 1529-3181

Information Visualization, international, peer-reviewed journal, SAGE JOURNALS, http://ivi.sagepub.com/. Print ISSN: 1473-8716. Online ISSN: 1473-8724.

Knowledge and Information Systems. An International Journal. ISSN: 0219-1377 (printed version). ISSN: 0219-3116 (electronic version). Publicado por Springer. Factor de impacto 2.0.

Redes. Revista Hispana para el Análisis de Redes Sociales. ISSN 1579-0185. Disponible en línea: http://revista-redes.rediris.es/

Bibliografía y Referencias

Archambault, D;Munzner, T;David, D (2007). *Topolayout: multi-level graph layout by topological features*. IEEE Transactions on Visualization and Computer Graphics. Vol. 13. 2. pp. 305-317.

Auber, D;Delest, M;Domenger, J P;Dulucq, S (2006). *Efficient drawing and comparison of rna secondary structure*. Journal of Graph Algorithms and Applications. Vol. n/d. n/d. pp. 329-351.

Barnes, J A (1972). *Social networks*. New York: Addison-Wesley. pp.

Bastian, M;Heymann, S;Jacomy, M (2009). *Gephi: an open source software for exploring and manipulating networks*. In International AAAI Conference on Weblogs and Social Media (2009). pp.

Bender-deMoll, Skye;McFarland, Daniel A (2006). *The art and science of dynamic network visualization*. Journal of Social Structure. Vol. 7. 2. pp. n/d.

Bharanipriya, V;Kamakshi, P (2011). *Web content mining tools: a comparative study*. International Journal of Information Technology and Knowledge Management. Vol. 4. 1. pp. 211-215.

Blondel, V D;Guillaume, J_L;Lambiotte, R;Lefe, E (2008). *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics: Theory and Experiment. Vol. 2008. 10. pp. .

Borgatti, S (2003). *The state of organizational social network research today*. Department of Organization Studies, Boston University, Manuscript.

Chen, H;Fuller, S S;Friedman, C;Hersh, W (2005). *Medical informatics knowledge management and data mining in biomedicine*. XLIV. pp. 647

Chen, V B;Davis, I W;Richardson, D C (2009). *King (kinemage, next generation): a versatile interactive molecular and scientific visualization program*. Protein science : a publication of the Protein Society. Vol. 18. 11. pp. 2403-2409.

Chimani, M;Gutwenger, C;Jünger, M;Klein, K;Mutzel, P;Schulz, M (2007). *The open graph drawing framework*. 15th International Symposium on Graph Drawing 2007, Sydney (GD07). pp.

de_Nooy, W;Mrvar, A;Batagelj, V (2005). *Exploratory social network analysis with pajek*. 2nd Edition. Series: Structural Analysis in the Social Sciences (No. 34). pp. 442 pages

Douglas, S M;Montelione, G T;Gerstein, M (2005). *Pubnet: a flexible system for visualizing literature derived networks*. Genome Biology. Vol. 6. . pp. .

Elbirt, B (2007). *Jacob's ladder 11.0 multidimensional data animation, visualization and intonation application for creating virtual reality displays*. Sunbelt XXVII Conference; Vancouver, CA.. pp. 10.

Ellson, J;Gansner, E R;Koutsofios, E;North, S C;Woodhull, G (2001). *Graphviz - open source graph drawing tools*. Graph Drawing . Vol. n/d. n/d. pp. 483-484.

Eppstein, David (2010). *Graph drawing. revised papers. series: lecture notes in computer science, vol. 5849.*. Subseries: Theoretical Computer Science and General Issues, 1st Edition. pp. 426

Feinerer, I;Hornik, K;Meyer, D (2008). *Text mining infraestructure in r*. Journal of Statistical Software - published by the American Statistical Association. Vol. 25. Issue 5. pp. 1-54.

Friendly, Michael (2008). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Trabajo documental apoyado por una beca de la "National Sciences and Engineering Research Council of Canada", Grant OGP0138748. Manuscrito.. Vol. N/D. N/D. pp. 79.

Granovetter, M (1979). *The theory-gap in social network analysis*. In P. Holland and S. Leinhardt, eds., Perspectives on Social Network Research. pp. 501-518

Hao, M C;Dayal, U;Hsu, M;Sprenger, T;Gross, M H (2001). *Visualization of directed associations in e-commerce transaction data*. Proceedings of VisSym'01, n/d.

Harrison, G W;Haruvy, E;Rutström, E (2010). *Remarks on virtual world and virtual reality experiments*. Center for the Economic Analysis of Risk. Georgia State University, Southern Economic Journal. Manuscript.

Hjorth_Jensen, M (2003). *Lecture notes on computational physics*. University of Oslo. pp.

Huffaker, B;Nemeth, E;claffy, k (1999). *Otter: a general-purpose network visualization tool*. , The Internet Society.

Huisman, M;van Duijn, M A (2011). *A reader's guide to sna software*. In J. Scott and P.J. Carrington (Eds.). The SAGE Handbook of Social Network Analysis, pp. 578-600. London: SAGE. pp.

Krackhardt, D;Blythe, J;Mcgrath, C (1994). *Krackplot 3.0: an improved network drawing program*. Connections. Vol. 17. 2. pp. 53-55.

Lenoir, T (1999). *Shaping biomedicine as an information science*. in Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems, edited by Mary Ellen Bowden, Trudi Bellardo Hahn, and Robert V. Williams. ASIS Monograph Series. (Medford, NJ: Information Today, Inc., 1999), pp. 27-45, by Timothy Lenoir, Program in History & Philosophy of Science, Stanford University .

Lopes, C T;Franz, M;Kazi, F;Donaldson, S L;Morris, Q;Bader, G D (2010). *Cytoscape web: an interactive web-based network browser*. Bioinformatics. Vol. 26. 18. pp. 2347-2348.

Matsuo, Yutaka;Mori, Junichiro;Hamasaki, Masahiro;Ishida, Keisuke;Nishimura, Takuichi;Takeda, Hideaki;Hasida, Koiti;Ishizuka, Mitsuru (2006). *Polyphonet: an advanced social network extraction system from the web, in proceedings of the 15th international conference on world wide web (www '06)*. Proceedings of the 15th international conference on World Wide Web (WWW '06). Vol. N/D. N/D. pp. 397-406.

Mejía, C P (2010). *Análisis de redes sociales a gran escala*. Tesis de Maestría-Sección de Computación-CINVESTAV. pp. i-xiii; 102

Mika, Peter (2005). *Flink: semantic web technology for the extraction and analysis of social networks*. Web Semant. Vol. 3. . pp. 211-223.

Moreno, J L (1978). *Who shall survive? foundations of sociometry, group psychotherapy and sociodrama. reedición del original de 1953*. Beacon House Inc. (1978 Third Edition. Copyright 1953 Ed.). pp. cxiv; 763

Rogers, E (1987). *Progress, problems and prospects for network research*. Social Networks. Vol. 9. n/d. pp. 285-310.

Salancik, G R (1995). *Wanted: a good network theory of organization*. Administrative Science Quarterly. Vol. 40. . pp. 40: 345-349.

UNO, (2010). *United nations e-government survey 2010. leveraging e-government at a time of financial and economic crisis*. . pp. 140

Wu, Xindong;Kumar, Vipin;Ross Quinlan, J;Ghosh, Joydeep;Yang, Qiang;Motoda, Hiroshi;McLachlan, Geoffrey;Ng, Angus;Liu, Bing;Yu, Philip;Zhou, Zhi-Hua;Steinbach, Michael;Hand, David;Steinberg, Dan (2008). *Top 10 algorithms in data mining*. Knowledge and Information Systems. Vol. 14. N/D. pp. 1-37.

Würthinger, T (2007). *Visualization of program dependence graphs*. Thesis M. Science. Institute for System Software. Johannes Kepler University in Linz

Yeung, N;Cline, M S;Kuchinsky, A;Smoot, M E;Bader, G D (2008). *Exploring biological networks with cytoscape software*. Current protocols in bioinformatics, Vol. 8. pp.

Anexo 1. Calendario de Trabajo

	2011		2012											
ACTIVIDAD	NOVIEMBRE	DICIEMBRE	ENERO	FEBRERO	MARZO	ABRIL	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE	OCTUBRE	NOVIEMBRE	DICIEMBRE
EXAMEN PREDOCTORAL														
INVESTIGACION DE BIBLIOGRAFIA														
ELABORACION Y CONCLUSION DEL MARCO CONCEPTUAL														
OBTENCION DE DATOS														
INVESTIGACIÓN Y EXPERIMENTACIÓN SOBRE LA ESTRUCTURA, PATRONES Y CUANTIFICACIÓN DE LAS ***INTERACCIONES DEL SITIO DE INTERNET*** (CIC-IPN)														
INVESTIGACIÓN Y EXPERIMENTACIÓN SOBRE LA ESTRUCTURA, PATRONES Y CUANTIFICACIÓN DE LAS ***INTERACCIONES DE CORREO*** (CIC-IPN)														
INVESTIGACIÓN Y EXPERIMENTACIÓN SOBRE LA ESTRUCTURA, PATRONES Y CUANTIFICACIÓN DE LAS ***INTERACCIONES DEL SITIO DE INTERNET*** (CINVESTAV-IPN)														
INVESTIGACIÓN Y EXPERIMENTACIÓN SOBRE LA ESTRUCTURA, PATRONES Y CUANTIFICACIÓN DE LAS ***INTERACCIONES DE CORREO*** (CINVESTAV-IPN)														
INTERPRETACIÓN DE RESULTADOS														
VALIDACIÓN DE LOS RESULTADOS														
ELABORACIÓN Y ENVÍO DE ARTÍCULO PARA REVISTA														
AUTORIZACION DE PUBLICACIÓN DE ARTÍCULO														
ELABORACIÓN DE PONENCIA PARA CONGRESO														
ENVÍO DE PONENCIA														
EXAMEN DOCTORAL														