Protocolo del proyecto

1. Antecedentes y metodología.

Hace unos siete años, un grupo de investigadores del Laboratorio de Dinámica no Lineal de la Universidad Nacional Autónoma de México y del Instituto Finlay de Cuba, se plantearon como tema de trabajo: explorar las bases digitales de información, aprovechando las nuevas tecnologías para el análisis inteligente de datos y el descubrimiento de conocimiento, salvando así las limitaciones que tiene la aplicación de otros métodos tradicionales a tan grandes volúmenes de datos.

Estas investigaciones han estado auspiciadas por el Consejo Nacional de Ciencia y Tecnología (CONACYT) de México, el cual aprobó en el 2003 un proyecto conjunto para desarrollar está línea de trabajo (Redes Neuronales para la Minería de Datos y Textos: Aplicación al Análisis Exploratorio y Descubrimiento de Conocimiento en Grandes Bases de Datos de Información Biomédica, No. Referencia: J200.554/2004). En esencia, el proyecto ha estado dirigido al estudio de algoritmos de visualización (como las Redes Neuronales Artificiales, RNA) y técnicas de minería de datos y textos. Durante este lapso de tiempo se ha profundizado en el desarrollo de la metodología ViBlioSOM y en la creación de diversos sistemas de software como el DataSOMinig, el cual fue registrado en el Instituto Nacional de Derecho de Autor de México (Anexo 1). Con ambas creaciones, se han realizado aplicaciones experimentales en diversas áreas del conocimiento y sobre diferentes temas. El último de ellos, relacionado con las vacunas contra la Tuberculosis (Hernández, J., Guzmán, MV., Cuello, O., et al. 2007). El proyecto generó además una herramienta computacional para el estudio de mapas auto-organizados (LabSOM: laboratorio computacional para el estudio del SOM).

Este marco de colaboración, sirvió además para formar estudiantes de licenciatura (4 graduados), de maestría en matemáticas (1 graduado) y de un doctorado. Elio Atenógenes Villaseñor (maestría) abordó el tema "La red neuronal SOM como herramienta para el análisis y visualización de grandes volúmenes de información digital" y María Victoria Guzmán (doctorado) trató el tema de "ViBlioSOM: Metodología para la Visualización de Información Métrica con Mapas Auto-organizados". Esta tesis fue predefendida satisfactoriamente en Julio del 2008 en la Universidad de La Habana.

Se han originado una serie de artículos de investigación (ver acápite 7, Publicaciones) que abordan tanto la metodología, los algoritmos, así como las aplicaciones realizadas con el ViBlioSOM. Existen además un conjunto de trabajos presentados en eventos científicos nacionales e internacionales que se suman a estos resultados.

Como muestra de la capacidad en la generación de resultados a partir de la colaboración entre ambas instituciones, en el 2001 la coordinadora por la parte cubana obtuvo el Premio Nacional de la Academia de Ciencias de Cuba con la colaboración del líder de este proyecto y de colegas de otros países (Anexo 2). El trabajo estuvo vinculado al desarrollo de "Herramientas para el análisis de oportunidades científico-tecnológicas", su aporte fundamental es el desarrollo de una metodología propia y la mejora de capacidades para el análisis de información, adaptando un software de propósito general para los fines específicos de la Bibliometría, Patentometría y temas afines. Así como por la introducción en Cuba de la aplicación de la redes neuronales artificiales en el análisis de información bibliográfica y la difusión de la importancia del mapeo de la ciencia y la tecnología. Utilizando esta metodología se llevaron a cabo las primeras aplicaciones, asociadas fundamentalmente, a la Patentometría para la inteligencia científico-tecnológica y a los sistemas de inteligencia empresarial (Guzmán MV., Sotolongo G.; 2002), (Salgado D.; Guzmán MV.; Carrillo, H.; 2003).

Este proyecto se propone como continuidad del anterior, teniendo como base la experiencia acumulada en el desarrollo de software de análisis y visualización de datos; así como la capacidad de realizar aplicaciones de estas tecnologías en el análisis métrico. En esta nueva etapa se pretende:

a) Incorporar al Software ViBlioSOM, un modulo para el análisis y visualización de información científica proveniente de cualquier base de datos bibliográfica. Un modulo que permita el análisis tecnológico a partir de la exploración de bases de datos de patentes y un modulo cuyo objetivo sea descubrir información en bases de datos biológicas no bibliográficas. Esto permitirá realizar aplicaciones en diferentes campos y hacer mucho más robusto el software. Se acota que la versión obtenida en el proyecto inicial solo permite la recuperación, procesamiento, análisis y visualización de datos provenientes de la Base de datos bibliográfica Medline (el contenido fundamental de esta BD es la Biomedicina).

b) Realizar aplicaciones, usando la tecnología desarrollada, a modo de evaluación y validación del software. Además de cumplir el ciclo de inteligencia empresarial de las diferentes organizaciones implicadas, estas herramientas estarán disponibles al público con el objetivo de lograr un mayor impacto en la comunidad biomédica.

c) Dada la adscripción de los coordinadores y la orientación profesional de los miembros del grupo, el ámbito de aplicación y experimentación inicial será el de la ciencia y tecnología biomédica (neurociencia, biotecnología en vacunas, bioinformática y ciencias agropecuarias).

Metodología y motivación

El proyecto tiene como antecedentes los conocimientos precedentes (ver Anexos y acápite 7 sección Publicaciones) y elementos de las disciplinas siguientes:

Metodología ViBlioSOM: Consiste en un sistema modular y abierto, basado en el uso secuencial de diferentes software propietarios. Esta metodología permite aplicar una serie importante de indicadores métricos al análisis de información y obtener resultados interesantes, incluyendo una representación visual en forma de mapas o cartografías. Desde el punto de vista metodológico el ViBlioSOM sigue un proceso estructurado que se instrumenta en las etapas siguientes: (1) búsquedas bibliográficas y salvas de ficheros resultantes, (2) tratamiento (conversión) de los ficheros resultantes, (3) creación de la base de datos, (4) normalización de la base de datos, (5) tratamiento matemático - estadístico (análisis métrico), (6) obtención de los indicadores, (6) generación de representaciones visuales (8) interpretación y conclusiones.
Esta plataforma garantizaba una secuencia lógica en la obtención de resultados, desde el primer dato salvado hasta la posterior representación en un mapa. Estos resultados pueden ser reproducidos por cualquier especialista que conozca el sistema descrito. Aspecto que ofrece credibilidad a la metodología. Además el empleo de redes neuronales artificiales reduce considerablemente el tiempo de procesamiento de los datos.

ViBlioSOM Software: Basado en la Metodología ViBlioSOM. Tiene implementados los pasos explicados dentro de los procesos de KDD. Para la visualización se implementó el algoritmo de las RNA tipo SOM (Mapas autor-organizados de Kohonen), así como otras técnicas de agrupamiento (clustering). Por otra parte, la suite cumple con el estándar de calidad ISO/IEC 9126-1.

Descubrimiento de Conocimientos en las Bases de Datos (Knowledge Discovery in Databases, KDD). Se entiende, de forma muy genérica, que el KDD es el proceso de extracción de conocimiento no trivial, como la identificación de patrones, que sean válidos, novedosos, potencialmente útiles y entendibles, a partir del procesamiento semi-automáticos de grandes bases de datos (Norton, MJ.; 1999), (Fayyad, U.; Piatetsky-Shapiro, G., Smyth, P; 1996), (Fayyad, U.; 1996).

Minería de datos y textos: La minería de datos (MD) está asociada con la recuperación y análisis de información en condiciones adversas (ruido, búsquedas incompletas, etc.). Mientras que la minería de textos (MT) se deriva de la Minería de datos y es considerada como la extracción o recuperación automatizada de conocimiento proveniente de elementos textuales. La MD y la MT son consideradas por los expertos, como técnicas del proceso KDD. Con las nuevas tecnologías de la información, la minería de datos ha tenido un importante impacto en la industria y en el mundo de los negocios, para resolver una gran variedad de problemas como: planeación económica, inteligencia empresarial, finanzas, análisis de mercados y análisis de perfiles de clientes.

Análisis y visualización de información: la visualización de datos multidimensionales permite identificar relaciones entre una gran cantidad de variables, representaciones de grandes matrices o datos, que hayan resultado de algún procesamiento o análisis previo. Dada la alta dimensionalidad que suelen presentar los datos, los elementos o relaciones importantes, que aportan un conocimiento nuevo, pueden escapar del ojo humano, sin una representación gráfica adecuada. Por tanto, la utilidad de la visualización de datos es:

- Revelar los patrones que subyacen a los datos: Descubrir el conocimiento.
- Hacer comprensible el conocimiento: Transmitir un mensaje.

La Visualización de Información (VI) es el proceso que transforma datos, información, y conocimiento en una forma que permite al sistema visual humano percibir la información de forma integrada. Su meta es permitir al usuario observar, para entender, y pueda encontrarle sentido a la información. Su objetivo es crear interfaces visuales ricas que permitan ayudar a los usuarios a entender y a navegar a través de los espacios complejos de la información.

Técnicas de visualización de información: Diferentes dominios de aplicación requieren técnicas de representaciones visuales distintas como son: Eigenvalue (Valor propio o descomposición del valor (DVP)), el LSA (Latent Semantic Analysis o Análisis semántico latente), PFNet Pathfinder Networks (PFnets) y Triangulación (Triagulation) y más usadas como Análisis de Componentes Principales (PCA), Redes Neuronales Artificiales (ANN), Escalado Multidimensional (MDS), etc.

Por su importancia económico – social:

Las ventajas estratégicas y económicas que representan las investigaciones asociadas a este proyecto pueden estar contenidas en las propias aplicaciones que se pueden obtener con la utilización de estos métodos de análisis en diferentes ámbitos de la vida. El análisis y visualización métrica, al permitir la representación de grandes conjuntos de datos en forma de mapas, permite identificar situaciones estratégicas que no han sido divulgadas, como son las líneas tecnológicas en las que trabajan los competidores, alianzas entre empresas, tecnologías emergentes y en declive, etc. Pueden permitir la solución a un determinado problema tecnológico, por ejemplo identificando procesos para mejorar el petróleo pesado (Sotolongo G., Guzmán MV., Carrillo H; 2002). También han permitido evaluar la situación científico-tecnológica de aspectos importantes dentro de la investigación o la producción, así como medir la relación entre la investigación y la innovación (Guzmán MV., Milanés Y., Martínez T., Carrillo H.; 2006).

Además, resulta de gran utilidad en la gestión de los activos intelectuales de una empresa, y en su forma más general para gestionar el conocimiento existente dentro y fuera de la propia empresa. Se tiene la experiencia concreta de las empresas farmacéuticas, que necesitan contar con un dispositivo orgánicamente estructurado que le permita entre otras cosas: justipreciar su capital intelectual (conocimiento científico-tecnológico); así como hacer una mejor gestión del mismo con la competencia como horizonte. Mientras que en caso de la Bioinformática, se espera realizar diferentes exploraciones a las bases de datos biológicas. Aplicaciones consideradas de importancia fundamental por haberse convertido la bioinformática en eje de desarrollo de la Biotecnología. Los resultados de estas aplicaciones están considerados de importancia fundamental por haberse convertido la bioinformática en eje de desarrollo de la Biotecnología y

por el impacto socio-económico que implica. Afirmación avalada por su capacidad de reducir los tiempos de respuestas en la solución de un problema de salud determinado o de investigación, pero sobre todo porque puede marcar la diferencia en un modo de hacer ciencia.

Por otra parte, una correcta representación visual de los datos, resultados de un análisis métrico, con su consiguiente correcta interpretación permitirá re-direccionar proyectos de investigación, disminuir el tiempo en la obtención de los resultado de una investigación científica, adelantarse a los competidores y conocer el mercado donde se especula invertir, etc. En el campo de las ciencias sociales es útil para identificar grupos con problemas, zonas geográficas con incidencias sociales importantes, entre otros.

Otro enfoque económico y estratégico de este proyecto, es que puede insertarse como herramienta o parte de los sistemas de vigilancia científico – tecnológica de empresas comerciales u organismos del estado como los Observatorios de Ciencia y Tecnología, Ministerios y Programas de apoyo gubernamentales.

Así mismo, los resultados aquí obtenidos pueden constituir material para su generalización, material de estudio de alumnos de las carreras de ciencias y tecnologías de la Información y base para estudios teóricos posteriores.

2. Programa de trabajo a desarrollar.

I Fase: Diseño, desarrollo e implementación de herramientas para el análisis y visualización de datos métricos.

1. Estudio de los tipos de fuentes de información (Bibliográficas Científicas, Patentes y Biológicas) que se van a incorporar al ViBlioSOM.
    - Identificar en cada caso los principales descriptores u etiquetas de los campos del documento bajo estudio.
    - Definir estructura de datos por fuentes de información.
    - Normalizar las estructuras de datos en función de su integrabilidad para el procesamiento y análisis.

2. Análisis de los diferentes sistemas de recuperación de información online por tipología de fuentes de datos para desarrollar:
    - Filtrado de la información.
    - Descargas automatizadas
    - Conversión de ficheros
    - Validación de registros

3. Desarrollar una aplicación informática integrada al sistema que permita la normalización asistida de la información según su procedencia.

4. Revisión de los principales tesauros, listas de materias, ontologías, etc. para:
    - Identificar los principales descriptores de interés, en función con los objetivos del proyecto.
    - Construcción de un fichero u índice invertido para el software en desarrollo.

5. Proponer una Batería de Indicadores Métricos, visualizados en diferentes interfaces según la:
    - Naturaleza de la Información.
    - Niveles de Agregación.
    - Tratamiento Inteligente de los Datos

6. Trabajar en la modularidad o integralidad del sistema con una alta escalabilidad.

7. Desarrollar una metodología de evaluación sistema de software sustentada en los tipos de datos, batería de indicadores, algoritmos de visualización, etc.

II Fase: Evaluación y Validación funcional del software. Validar los resultados del sistema con aplicaciones prácticas para cada modulo.

8. Selección de herramientas de software (diferentes al ViBlioSOM) que utilicen algoritmos similares, con el objetivo de comparar las salidas visuales.

9. Seleccionar un método y estándar para elaborar el protocolo de validación y evaluación del sistema.

10. Diseñar y aplicar el protocolo de validación.

11. Aplicar un modelo de evaluación funcional, para cada área de aplicación:

    a) La investigación y el desarrollo de Vacunas a nivel internacional: evaluación del impacto científico de la investigación latinoamericana. (Bases de datos bibliográficas–científicas).

    b) Identificación de las áreas de impacto de la ciencia y la tecnología latinoamericana: el caso de Cuba y México. (Bases de datos científicas (artículos) y tecnológicas (Patentes)).

    c) Identificación de frentes de investigación especializados en el área de la Neuroinformática.(Bases de datos bibliográficas–científicas).

    d) Predicción de Epítopes T de *Mycobacterium Tuberculosis* (herramienta para el desarrollo de nuevas vacunas, BD biológicas).

    e) La investigación y el desarrollo de las ciencias agropecuarias en Latinoamérica: Evaluación científico-técnica de la producción de alimentos y su impacto en la economía de la región. Identificación de áreas clave en la producción de alimentos (Bases de datos bibliográficas - científicas).

3. Objetivo principal

Desarrollar herramientas de visualización y metodologías para el análisis métrico y su aplicación en esferas de gran impacto social como la Biomedicina.

4. Metas

1. Desarrollar, evaluar y distribuir en la comunidad científica de ambos países una herramienta informática (sistema de software) para el descubrimiento de conocimientos en bases de datos; con el objetivo de incidir en la toma de decisiones y mejorar eficiencia de la I+D+I+C[1] en el área de biomedicina.

[1] *I+D+I+C: Investigación + Desarrollo + Innovación + Comercialización.*

2. Intercambiar conocimientos y esfuerzos, partir de la integración entre ambos países en la rama de las tecnologías de la información. Homogenizar y analizar diversas metodologías, así como estándares novedosos que permitan la minería de datos y textos.

3. Formación de recursos humanos en técnicas de minería de datos y textos aplicados a la biomedicina.

4. Evaluar científicamente los resultados de investigación de la comunidad científica cubano-mexicana, a través de la construcción de indicadores métricos, en el dominio biomédico

5. Producir artículos de investigación por cada ítem del Programa de Trabajo, presentaciones de trabajos en eventos y 5 tesis de grado.

5. Justificación de las acciones internacionales solicitadas.

Existen varios antecedentes de colaboración internacional con el equipo cubano. Dicha colaboración ha sido muy fructífera, prueba de ello son los resultados obtenidos. Esta integración ha hecho que muchos intereses académicos y de investigación se hayan gestando en el transcurso de los últimos años. Las acciones llevadas a cabo durante este tiempo, propician la consolidación de un equipo multidisciplinario con personas altamente calificadas, que además cuentan con la infraestructura necesaria, para generar trabajos de investigación con un alto impacto social y desarrollar tecnologías con un gran valor agregado.

La parte cubana de nuestro grupo tiene experiencia en el campo de la investigación métrica y los aspectos específicos del trabajo con BD de información científica, de patentes y biológicas. La parte mexicana tienen experiencia en diversas aplicaciones de modelos no lineales y métodos matemáticos adaptivos (e.g.: redes neuronales) y al desarrollo de software científico. Es en estos algoritmos matemáticos en los cuales está basada la programación del software ViBlioSOM.

Las estancias en el extranjero (y las visitas recíprocas de los participantes cubanos) están programadas para transferirnos mutuamente conocimiento y experiencias, así como planear y discutir los avances del trabajo y afinar la escritura de los reportes de las investigaciones en curso.

Se han planeado estancias de 15 días para aprovechar al máximo cada viaje.