ISSN 1852-7094

Cuadernos de la

Facultad de Ingeniería e Informática

Número 5
Noviembre 2010

Universidad Católica de Salta

Publicación anual de la Facultad de Ingeniería e Informática
Universidad Católica de Salta

Campo Castañares
A4400 Salta
Tel. 0387 426 8539
ingenieria@ucasal.net
www.ucasal.net

AUTORIDADES

Gran Canciller:	S.E.R. Mons. Mario Antonio Cargnello
Rector:	Dr. Alfredo Gustavo Puig
Vice-Rector Administrativo:	Ing. Manuel Cornejo Torino
Vice-Rector Académico:	Dr. Gerardo Vides Almonacid
Secretaria General:	Prof. Lilian Constanza Diedrich de Duba
Decano de la Facultad de Ingeniería e Informática:	Ing. Ernesto Claudio Mondada
Secretario Académico:	Ing. Jorge Ricardo Guido
Jefe del Departamento de Ingeniería Civil:	Ing. Néstor Eugenio Lesser
Jefe del Departamento de Ingeniería Industrial:	Ing. Eduardo Cornejo Jovanovics
Jefa del Departamento de Ciencias Informáticas:	MBA Ing. H. Beatriz Parra de Gallo
Jefe del Departamento de Investigación:	Dr. Javier Moya
Jefe del Departamento de Extensión, Graduados y Bienestar:	Mg. Ing. Rodolfo Gallo Cornejo
Secretaria Técnica:	Sec. Ejecutiva Andrea A. Burgos
Coordinador de Laboratorios:	Ing. Manuel Zambrano Echenique

INDICE

Minería de texto para la categorización automática de documentos

M. Alicia Pérez Abelleira y Carolina A. Cardoso [*]

aperez@ucasal.net

Resumen

La clasificación de documentos de texto es una aplicación de la minería de textos que pretende extraer información de texto no estructurado. Su interés se justifica porque se estima que entre el 80% y el 90% de los datos de las organizaciones son no estructurados. Por otro lado, la búsqueda semántica permite al usuario especificar en una consulta no solamente términos que deben aparecer en el documento, sino conceptos y relaciones, que pueden detectarse mediante el análisis de texto. El objetivo de este trabajo es implementar un buscador semántico que aproveche el resultado de algoritmos de aprendizaje automático para la clasificación de documentos.

El dominio de aplicación es un corpus de más de 8000 documentos que contienen nueve años de resoluciones rectorales de la Universidad Católica de Salta en distintos formatos (Microsoft Word, texto plano, PDF). El sistema aprovecha las ventajas de la arquitectura UIMA sobre la que se han implementado analizadores que extraen meta-datos (fecha y número de resolución, unidad académica, personas, etc.) Asimismo se han explorado una variedad de algoritmos de aprendizaje semi-supervisado aplicados a la categorización de documentos, comparándolos experimentalmente entre sí y con algoritmos supervisados. Estos últimos precisan una gran cantidad de ejemplos etiquetados, algo generalmente costoso en la práctica en el caso de la clasificación de documentos. Los algoritmos semi-supervisados en cambio son capaces de aprovechar ejemplos no

[*] Alicia Pérez es Licenciada en Informática por la Universidad Politécnica de Madrid y PhD in Computer Science por Carnegie Mellon University. Actualmente se desempeña en la Facultad de Ingeniería e Informática de la UCS como docente de Sistemas Expertos y de Compiladores y como Jefa del Departamento de Investigación y en la Universidad Carlos III de Madrid (España) como docente de Inteligencia Artificial (2004, 2006) de la carrera de Ingeniería Informática – grupo bilingüe.
Carolina Cardoso es Licenciada en Ciencias de la Computación, se desempeña como Ayudante Docente de las asignaturas Estructuras de Datos y Algoritmos y Lenguaje I y Adjunta Interina de Compiladores.

etiquetados. En particular, en los experimentos en nuestro dominio el algoritmo de *co-training* ha demostrado tener buenas propiedades, incluso a pesar de la restricción teórica de que los atributos deben ser redundantes e independientes. No obstante el algoritmo supervisado SMO que entrena SVMs es superior.

Nuestro objetivo final es construir un buscador semántico que utilice los metadatos obtenidos automáticamente por los anotadores implementados en UIMA y las categorías asignadas automáticamente por los algoritmos de aprendizaje.

Palabras claves: buscador semántico, aprendizaje semisupervisado, categorización de documentos, minería de texto, UIMA

1. Introducción

El conocimiento es cada vez más un recurso de importancia estratégica para las organizaciones y su generación, codificación, gestión, divulgación aportan al proceso de innovación. Todos estos aspectos se incluyen en lo que se ha dado en llamar la *gestión del conocimiento*. La cantidad de documentos de diversos tipos disponibles en una organización es enorme y continúa creciendo cada día. Estos documentos, más que las bases de datos, son a menudo un repositorio fundamental del conocimiento de la organización, pero a diferencia de éstas la información no está estructurada. La minería de textos tiene como objetivo extraer información de texto no estructurado, tal como entidades (personas, organizaciones, fechas, cantidades) y las relaciones entre ellas. Por otro lado, la búsqueda semántica permite al usuario especificar en una consulta no solamente términos que deben aparecer en el documento, sino esas entidades y relaciones extraídas mediante el análisis de texto.

La categorización de documentos de texto es una aplicación de la minería de texto que asigna a los documentos una o más categorías, etiquetas o clases, basadas en el contenido. Es un componente importante de muchas tareas de organización y gestión de la información. El enfoque tradicional para la categorización de textos en que los expertos en el dominio de los textos definían manualmente las reglas de clasificación ha sido reemplazado por otro basado en técnicas de aprendizaje automático, o en combinaciones de éste con otras técnicas.

Nuestro trabajo se centra en desarrollar técnicas para la categorización automática de documentos según su contenido que

avancen el estado del arte en nuestro medio, aplicando el automático a la minería de texto. El objetivo final es implementar un buscador semántico que aproveche el resultado de algoritmos de aprendizaje automático para la clasificación de documentos. El dominio de aplicación es un corpus de más de 8000 documentos que contienen nueve años de resoluciones rectorales de la Universidad Católica de Salta en distintos formatos (Microsoft Word, texto plano, PDF).

Este artículo comienza describiendo la información no estructurada, repositorio fundamental del conocimiento de una organización, y las arquitecturas para la gestión de información no estructurada. La Sección 3 muestra nuestra instanciación del modelo general para el problema de la clasificación y búsqueda de resoluciones rectorales. En la Sección 4 se exploran diferentes algoritmos para la categorización de documentos y se describen los experimentos realizados para determinar su adecuación a nuestro dominio. Concluye el trabajo con el funcionamiento del motor de búsqueda semántica (Sección 5) y presentando algunas conclusiones.

2. Información estructurada y no estructurada

La información estructurada se caracteriza por tener un significado que pretende no tener ambigüedad y que está representado explícitamente en la estructura o formato de los datos. El ejemplo típico es una base de datos relacional. De la información no estructurada podría decirse que su significado no está implicado en su forma y por tanto precisa interpretación para aproximar o extraer el mismo. Los documentos en lenguaje natural o hasta de voz, audio, imágenes, etc. entran en esta categoría. El interés por extraer significado de la información no estructurada se debe a que se estima que entre el 80% y el 90% de los datos de las organizaciones son no estructurados (Moore, 2002). Aunque muchas organizaciones han invertido en tecnologías para minería de datos estructurados procedentes de sus bases de datos y sistemas transaccionales, en general no han intentado capitalizar sus datos no estructurados o semi-estructurados.

Una aplicación de gestión de la información no estructurada (UIM por sus siglas en inglés) típicamente es un sistema de software que analiza grandes volúmenes de información no estructurada con el fin de descubrir, organizar y entregar conocimiento relevante al usuario final. La información no estructurada puede ser mensajes de correo electrónico, páginas web o documentos generados con una variedad de procesadores de texto, como en el caso de las resoluciones rectorales de nuestra universidad. Estas aplicaciones utilizan para el análisis una

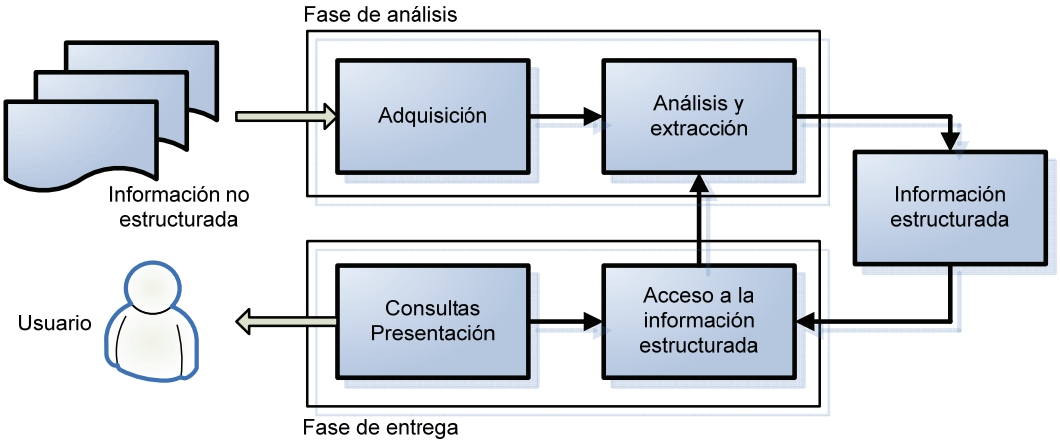


Figura 1: Esquema de una arquitectura UIM (Ferrucci & Lally, 2004)

variedad de tecnologías en las áreas del procesamiento del lenguaje natural, recuperación de la información, aprendizaje automático, ontologías y hasta razonamiento automático.

El resultado del análisis generalmente es información estructurada que representa el contenido de la entrada no estructurada y que se hace accesible al usuario mediante aplicaciones adecuadas. Un ejemplo puede ser la generación de un índice de búsqueda y la utilización de un buscador que facilita el acceso a documentos de texto por tema, ordenados según su relevancia a los términos o conceptos procedentes de la consulta del usuario.

Existen diversas arquitecturas para el desarrollo de aplicaciones UIM. Para nuestro trabajo hemos utilizado UIMA (Ferruci & Lally, 2004), una arquitectura software basada en componentes que surgió como proyecto de investigación de IBM y fue puesta a disposición de la comunidad como software libre.

3. Arquitectura del sistema

Conceptualmente suele verse a las aplicaciones de UIM con dos fases: una de análisis y otra de entrega de la información al usuario. En la fase de análisis se recogen y analizan colecciones de documentos. Los resultados del análisis se almacenan en algún lenguaje o depósito intermedio. La fase de entrega hace accesible al usuario el resultado del análisis, y posiblemente el documento original completo mediante una interfaz apropiada. La Figura 1 muestra ambas fases de manera esquemática.

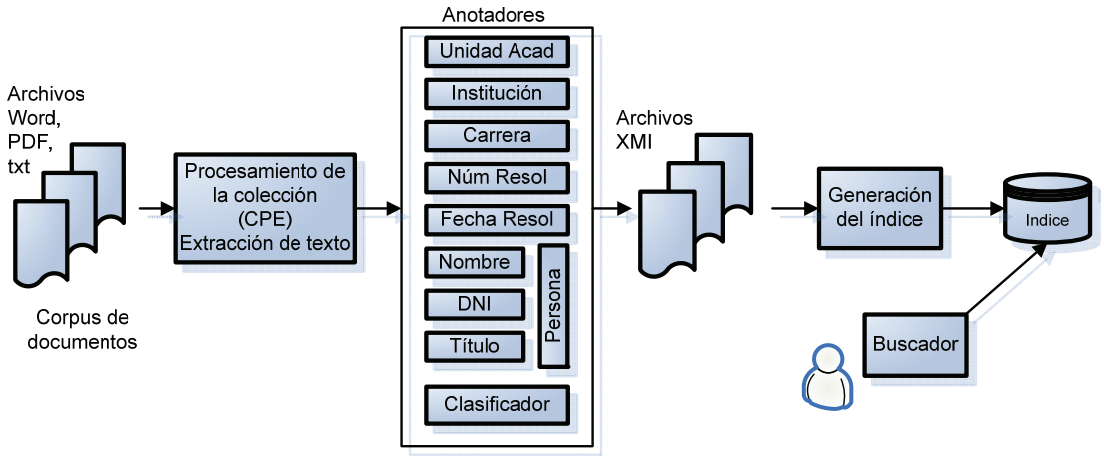


Figura 2: Arquitectura del sistema

La Figura 2 muestra la aplicación de este esquema a nuestro dominio, en el que partimos de más de 8000 resoluciones rectorales en archivos de texto de distinto tipo: Word, PDF, texto plano. Previo al análisis, se procede a la extracción del texto de cada archivo utilizando las herramientas de software libre POI (poi.apache.org) y tm-extractors (www.textmining.org). Las vocales acentuadas y otros caracteres especiales son reemplazados por los caracteres ASCII correspondientes (ej. *á* es reemplazado por *a*). También se divide en partes la resolución extrayendo el encabezado (texto que contiene el número y la fecha de la resolución) y el cuerpo con la mayor parte de la información, y descartando en lo posible el texto “de forma”.

La fase de análisis incluye tokenización y detección de entidades en documentos individuales tales como personas, fechas, organizaciones, unidades académicas y datos sobre la resolución (fecha y número). Además con la ayuda de un clasificador aprendido automáticamente del corpus de resoluciones, como se explica en la Sección 4, se anota cada documento con una categoría. Existen 21 categorías que fueron obtenidas del personal especializado en la elaboración de resoluciones. Algunos ejemplos son: designación de planta docente, convenio de pasantías, convenio de colaboración, llamado a concurso docente, o designación de tribunal de concurso.

El resultado de la fase de análisis es un conjunto de archivos en formato XMI (Sección 3.3). Estos archivos contienen, además de las partes relevantes del texto original, metadatos en forma de anotaciones correspondientes a las entidades y a la categoría de documentos. Estos archivos serán procesados para construir el índice de un motor de búsqueda que contiene los tokens (en nuestro caso, las palabras que

aparecen en el texto) así como las entidades y categorías extraídas automáticamente.

En la fase de entrega existe una interfaz para hacer consultas de búsqueda en el índice de forma que el usuario pueda buscar documentos que contengan combinaciones booleanas de tokens, entidades y categorías mediante un motor de búsqueda semántica.

3.1. Análisis a nivel de documento

UIMA es una arquitectura especialmente pensada para combinar los distintos componentes de la fase de análisis de texto y facilitar su disponibilidad para diversas aplicaciones en la fase de entrega. En UIMA, el componente que contiene la lógica del análisis se llama anotador. Cada anotador realiza una tarea específica de extracción de información de un documento y genera como resultado anotaciones, que son añadidas a una estructura de datos denominada CAS (*common analysis structure*). A su vez, esas anotaciones pueden ser utilizadas por otros anotadores. Los anotadores pueden ser agrupados en anotadores agregados.

La mayoría de los anotadores de nuestro sistema realizan reconocimiento de entidades con nombre (*named entity recognition* o NER), tarea principal de los sistemas de recuperación de información, que busca localizar y clasificar elementos atómicos del texto que pertenecen a categorías predefinidas. Las entidades extraídas por nuestro sistema son: personas, unidades académicas, carreras, instituciones, fechas, número y año de las resoluciones. Cada una de las entidades es extraída por un anotador. Además, para detectar entidades correspondientes a personas se utilizan entidades como nombres propios, DNIs y títulos, obtenidas por los anotadores correspondientes.

Las técnicas utilizadas para el reconocimiento de entidades son (Alias-i, 2008; Feldman & Sanger, 2007):

- Equiparación con expresiones regulares que capturan el patrón que siguen las entidades (ejemplos son la detección de DNIs, fecha y número de las resoluciones).
- Equiparación con diccionarios y *gazetteers* (ejemplos son las carreras, unidades académicas, instituciones, títulos y nombres propios). El diccionario de nombres propios consta de más de 1300 nombres y fue extraído automáticamente del sistema de gestión de alumnos. El enfoque basado en componentes de UIMA nos ha permitido adaptar el *Gazetteer Annotator* de Julie Lab (Tomanek &

Wermter, 2008), que está basado en la implementación que hace Lingpipe del algoritmo Aho-Corasick (Alias-i, 2008).
- Equiparación con plantillas: para detectar entidades correspondientes a personas se utiliza una plantilla que describe a la persona mediante los siguientes atributos: *nombre1, nombre2, apellido(s), DNI, título*. Sólo *nombre1* y *apellido(s)* son obligatorios. Todos los atributos son a su vez entidades detectadas por anotadores, excepto *apellido(s)* que se detecta mediante una expresión regular.

Además de los anotadores mencionados, se utiliza el anotador de UIMA que detecta tokens usando una sencilla segmentación basada en los espacios en blanco, y crea anotaciones con tokens y sentencias.

Aparte de todos estos anotadores, existe otro que asigna la categoría de documento en base al modelo aprendido automáticamente, como se describe en la Sección 4.

3.2. Análisis a nivel de colección

Mientras que el análisis descripto en la sección anterior se realiza en cada documento, existen casos en que el análisis debe ser realizado al nivel de una colección de documentos. Un caso particular es un bucle de realimentación, que produce recursos estructurados a partir del análisis de una colección de documentos y luego usa esos recursos para permitir el subsiguiente análisis de documentos (Ferrucci & Lally, 2004). Tal estructura de realimentación se ha utilizado para implementar el aprendizaje automático de categorías (Figura 3). Esta aplicación se puede dividir en dos fases: la fase de entrenamiento (parte superior de la figura) y la fase de clasificación (parte inferior).

En la fase de entrenamiento, un motor de procesamiento de colecciones (*Collection processing engine* o CPE) analiza una colección de documentos, denominada conjunto de entrenamiento. Este CPE invoca a un analizador que utilizando algoritmos de aprendizaje automático produce un modelo. Este modelo, basado en los atributos del documento, es capaz de asignarle una categoría. El modelo forma parte de un nuevo anotador, el clasificador de la Figura 2, que se encarga de aplicar el modelo. En la fase de clasificación, el modelo es consultado por el clasificador al analizar un nuevo documento y asignarle una categoría, que se convierte en una anotación más sobre el documento. La Sección 4 describe los algoritmos de aprendizaje utilizados para generar el modelo.

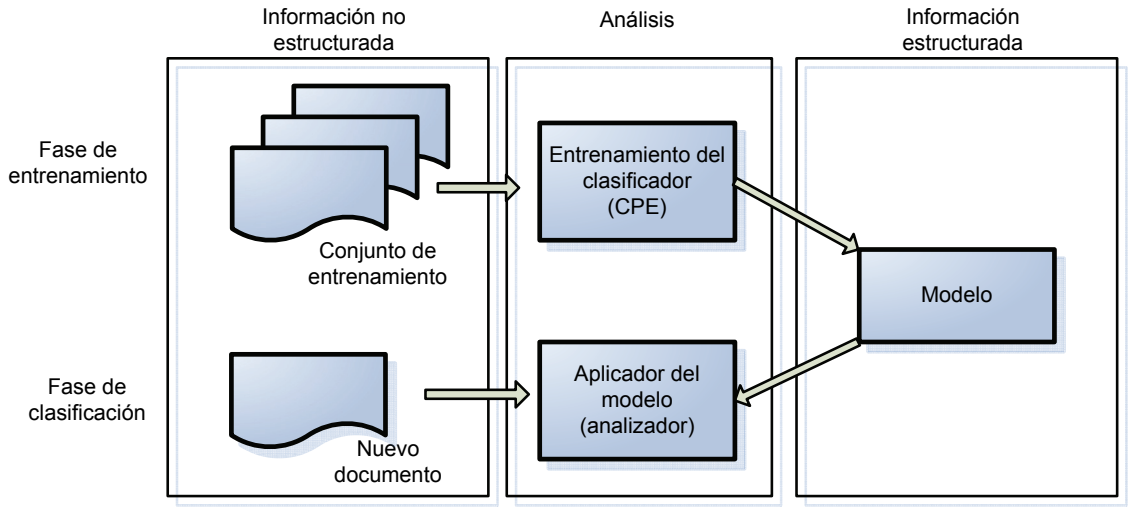


Figura 3: Implementación del aprendizaje automático en la arquitectura de UIMA

3.3. Formato de la información estructurada

Un importante principio de arquitectura en UIMA es que todos los analizadores operan sobre una estructura de datos estándar, el CAS. El CAS incluye el texto y las anotaciones. El CAS es el principal repositorio de información estructurada y utiliza como lenguaje UIMA el estándar XMI (XML Metadata Interchange) (OMG, 2007). XMI está basado en UML (Unified Modeling Language) y proporciona un formato en XML para el intercambio de modelos UML, lo que además hace posible la interoperabilidad con otras aplicaciones. En XMI el texto se incluye en el SofA (*Subject of Analysis*) del CAS como atributo sofaString y las anotaciones en el CAS tienen prefijos particulares en el espacio de nombres de XMI, como por ejemplo <*mio:Carrera*>. El espacio de nombres queda definido al declarar un sistema de tipos como parte del dominio en UIMA. La Figura 4 muestra un ejemplo del resultado del proceso de análisis. Dos *SofAs* recogen el encabezado, del que sólo se extrae la fecha y el número de resolución, y el cuerpo de la misma. La primera anotación del tipo *SourceDocument Information* guarda información sobre el archivo, para poder recuperarlo posteriormente, por ejemplo, para mostrarlo al usuario como resultado de una búsqueda. A continuación aparecen las anotaciones de Unidad Académica y de Carrera con su principio y fin en el texto del documento. La anotación Fecha de Resolución tiene varios campos: el texto original (*FechaResolCompleta*) y el día, mes y año extraídos del mismo.

Finalmente la anotación *Clase* contiene la categoría *DesigPlanta* asignada a esta resolución por el modelo aprendido.

```
<?xml version="1.0" encoding="UTF-8" ?>
<xmi:XMI … xmi:version="2.0">
…
  <cas:Sofa xmi:id="1" sofaNum="2" sofaID="encabezado"
mimeType="text" sofaString="ResoluciOn N 470/08 En el Campo
Castanares, sito en la Ciudad de Salta, Capital de la Provincia del
mismo nombre, Republica Argentina, sede de la Universidad Catolica
de Salta, a los veintisiete dias el mes de mayo del ano dos mil ocho:" />
  <cas:Sofa xmi:id="13" sofaNum="1" sofaID="_InitialView"
mimeType="text" sofaString="la presentacion efectuada por las
autoridades de la Escuela Universitaria de Educacion Fisica,
dependiente de la Facultad de Artes y Ciencias en virtud de la cual se
propone las modificaciones de designaciones docentes Planta
Transitoria, para la carrera Licenciatura en Educacion Fisica…" />
…
  <examples:SourceDocumentInformation xmi:id="25" sofa="13"
begin="0" end="0" uri="file:/D:/UIMA/docs/RES.%20%20Nº0470-08.txt"
offsetInSource="0" documentSize="2280" lastSegment="false" />
  <mio:UA xmi:id="48" sofa="13" begin="177" end="208"
confidence="30.0"
componentId="de.julielab.jules.lingpipegazetteer.GazetteerAnnotator"
specificType="UnidadAcademica" />
  <mio:Carrera xmi:id="72" sofa="13" begin="398" end="430"
confidence="0.0"
componentId="de.julielab.jules.lingpipegazetteer.GazetteerAnnotator"
specificType="Carrera" />
  <mio:NumeroResol xmi:id="33" sofa="1" begin="0" end="19"
nroResol="RESOLUCION N 470/08" numero="470" anio="2008" />
  <mio:FechaResol xmi:id="40" sofa="1" begin="196" end="248"
anio="2008" mes="MAYO" dia="27"
fechaResolCompleta="VEINTISIETE DE MAYO DE DOS MIL OCHO" />
   <mio:Clase xmi:id="28" sofa="1" begin="0" end="0" valor="
DesigDocPlanta" />
…
</xmi:XMI>
```

Figura 4: Ejemplo de texto anotado.

4. Aprendizaje automático para la categorización de documentos

Se define la categorización (o clasificación) de textos como la actividad de etiquetar textos en lenguaje natural con categorías temáticas de un conjunto predefinido (Sebastiani, 2005). Si a esto se añade la identificación automática de dichas categorías, el problema se denomina *clustering* de textos.

El enfoque dominante actualmente para el problema de categorización de textos se basa en técnicas de aprendizaje automático: se construye automáticamente un clasificador mediante aprendizaje inductivo de las características o atributos de las categorías a partir de un conjunto de documentos previamente clasificados que sirven como conjunto de entrenamiento. Se trata por tanto del llamado aprendizaje supervisado. Los algoritmos de aprendizaje generalmente utilizados para la (NaiveBayes, SMO, etc) pueden ser entrenados para clasificar documentos dado un conjunto suficientemente grande de ejemplos de entrenamiento, cada uno de los cuales ha sido etiquetado previamente con la categoría correspondiente. Una importante limitación de estos algoritmos es que precisan una cantidad grande de ejemplos (documentos) etiquetados para poder alcanzar una precisión apropiada. El etiquetado de hasta miles de documentos ha de ser realizado por una persona, especialista en el área de interés de los documentos, y por tanto es un proceso muy costoso.

Esto ha dado pie a nuevas familias de algoritmos de aprendizaje denominado semi-supervisado, capaces de aprender de un número limitado de ejemplos de entrenamiento etiquetados utilizando adicionalmente documentos no etiquetados, generalmente fácilmente disponibles. En un trabajo anterior (Pérez & Cardoso, 2009) hemos experimentado en una variedad de dominios con tres enfoques representativos de este tipo de algoritmos: la aplicación de *expectation maximization* (EM) al uso de datos etiquetados y no etiquetados; los algoritmos de *co-training*; y el algoritmo *co-EM* que combina características de los dos anteriores. En nuestros experimentos la implementación de *co-training* ha dado los mejores resultados para la categorización de textos, por lo que enfocamos en ella nuestra descripción.

La categorización de texto suele desarrollarse en tres etapas: pre-procesamiento de los datos, construcción del clasificador y categorización de los nuevos documentos, A continuación se describen los dos primeros pasos. El tercero es realizado mediante un anotador (ver Sección 3).

4. 1. Pre-procesamiento de los datos

En esta fase se transfiere el texto original, resultante de la tokenización, a un formato compacto que será utilizado en las fases siguientes. Incluye:

- Lematización (*stemming*), que reduce una palabra a su raíz (stem). Este proceso ayuda al recall, que es una medida sobre el número de documentos que se pueden encontrar con una consulta. Se ha utilizado el algoritmo propuesto e implementado en Snowball (snowball.tartarus.org).
- Eliminación de palabras de función (stopwords: artículos, preposiciones, conjunciones, etc. o) y otras dependientes del dominio. En el caso de las resoluciones rectorales, pueden ser 'Universidad Católica de Salta', 'rector', 'resuelve', etc. Esta eliminación está basada en un diccionario y ocurre después de la lematización.
- Selección de atributos, para reducir la dimensionalidad (número de atributos) del espacio de datos eliminando los que parezcan irrelevantes, como por ejemplo palabras que aparezcan en todos los documentos.
- Asignar distintos pesos a los atributos (o el mismo peso). Es común utilizar tf–idf (*term frequency–inverse document frequency*) para evaluar la importancia de una palabra en el corpus de documentos. Estamos realizando experimentos para elegir la opción más adecuada en nuestro corpus.

Con los pasos anteriores cada documento de la colección se convierte a una representación compacta adecuada para los algoritmos de aprendizaje. En nuestro caso dicha representación es el formato arff y el filtro *StringToWordVector* de Weka (www.cs.waikato.ac.nz/ml/weka/) puede ser configurado para aplicar la mayoría de las transformaciones anteriores, exceptuando la conversión de formatos de documentos.

4.2. Construcción del clasificador

Los documentos, transformados en conjuntos de atributos y valores por la etapa anterior, se utilizan para construir un clasificador que asignará categorías a nuevos documentos. Hemos evaluado una variedad de algoritmos. La limitación del aprendizaje supervisado es la disponibilidad de una cantidad importante de ejemplos de entrenamiento ya categorizados. Esta categorización o etiquetado es un proceso manual y laborioso. Para disminuir este esfuerzo surgen los algoritmos semi-supervisados, que son capaces de aprender a partir de

un conjunto de ejemplos etiquetados y no etiquetados. A continuación se describen brevemente algunos de los algoritmos utilizados en los experimentos.

4.2.1. SMO

El aprendizaje de máquinas de vectores soporte es un método supervisado que ha demostrado buenas propiedades para la categorización de documentos. El algoritmo SMO (*sequential minimal optimization*) que hemos utilizado es la implementación en Weka del algoritmo de optimización minimal secuencial para entrenar máquinas de vectores soporte usando un kernel polinomial (Witten & Frank, 2005).

4.2.2. Co-training

Blum y Mitchell (1998) introdujeron la idea de *co-training*, al reconocer que los atributos de algunos conjuntos de datos pueden descomponerse naturalmente en dos subconjuntos, y cada uno de ellos sirve efectivamente para clasificar cada documento, es decir, son redundantemente predictivos. Cada uno de esos subconjuntos actúa como una perspectiva diferente de la tarea de clasificación. La Tabla 1

Entradas: colecciones iniciales de documentos etiquetados *E* y sin etiquetar *N.*
•Dividir el conjunto de atributos en dos, *A* y *B.*
Repetir mientras existan documentos en *N*:
•Construir un clasificador *CA* usando los atributos *A* de los documentos de *E*.
•Construir un clasificador *CB* usando los atributos *B* de los documentos de *E*.
•Para cada clase *C*, elegir el documento de *N* que ha sido etiquetado por *CA* como perteneciente a *C* con mayor confianza, sacarlo de *N* y añadirlo a *E.*
•Para cada clase *C*, elegir el documento de *N* que ha sido etiquetado por *CB* como perteneciente a *C* con mayor confianza, sacarlo de *N* y añadirlo a *E.*
Salida: dos clasificadores *CA* y CB cuyas predicciones son combinadas multiplicando las probabilidades y renormalizándolas.

Tabla 1: El algoritmo de *co-training*

describe el algoritmo. Nuestra implementación genera los modelos en cada iteración con el algoritmo SMO.

Co-training depende de que las dos perspectivas (conjuntos de atributos A y B) sean redundantes e independientes - para cada instancia, los atributos de A son condicionalmente independientes de los de B dada la clase de la instancia. Aunque parezca una suposición poco realista en la práctica, hay resultados que muestran que funciona bien en ciertos conjuntos de datos que no satisfacen completamente estos requisitos, y esta consideración sigue siendo una pregunta abierta (Blum & Mitchell, 1998; Nigam & Ghani; 2000). Nuestros experimentos han estado dirigidos a entender el comportamiento en la categorización de textos en nuestro problema particular.

4.2.3. Expectation maximization (EM)

EM está basado en una técnica sencilla pero efectiva para clasificar textos, Naive Bayes, que sólo aprende de datos etiquetados (Witten & Frank, 2005). La idea es utilizar Naive Bayes para aprender clases para un pequeño conjunto etiquetado y después ampliarlo a un conjunto grande de datos no etiquetados utilizando el algoritmo EM de clustering iterativo (Nigam et al, 2000; Witten & Frank, 2005). El procedimiento tiene dos pasos: primero entrenar un clasificador Naive Bayes usando los datos etiquetados. Segundo, aplicarlo a los datos no etiquetados para etiquetarlos con las probabilidades de clase (el paso *E* o *expectation*). Tercero, entrenar un nuevo clasificador usando ahora las etiquetas de todos los datos (el paso *M* o maximización). Cuarto, repetir estos pasos hasta llegar a la convergencia (cuando el cambio en las probabilidades asignadas es menor que un cierto umbral). Nuestra implementación en Weka tomó como punto de partida la de Ray Mooney (disponible en www.cs.utexas.edu/users/ml/risc).

4.3. Configuración de los experimentos

Para evaluar nuestro enfoque hemos utilizado un corpus de 1000 resoluciones rectorales pertenecientes al año 2007 a las que se han asignado una de 21 categorías manualmente. De este corpus hemos extraído conjuntos de entrenamiento de diversos tamaños para experimentar comparando el comportamiento de los algoritmos supervisados (SMO) y no supervisados (*co-training*, EM) en función del número de ejemplos disponibles. Estos algoritmos fueron seleccionados en base a resultados de experimentos anteriores (Pérez y Cardoso, 2009).

El corpus ha sido preprocesando utilizando las utilidades de filtrado de Weka con los pasos descritos en la sección 4.1. En este preprocesamiento también hemos variado los valores de diversos parámetros, especialmente las maneras de seleccionar atributos y de asignar pesos a los atributos.

- Para reducir la dimensionalidad se ha realizado la selección de los 100 atributos más relevantes. Se ha experimentado con dos maneras de seleccionar estos atributos: los 100 más relevantes en la colección completa para todas las clases, y los 100 más relevantes en la colección completa para cada una de las clases. Esta segunda opción selecciona en la práctica entre 700 y 900 atributos.
- Los dos mecanismos más populares para asignar pesos a los atributos son binario y tfidf. En el primer caso, los pesos son 0 o 1 indicando ausencia o presencia de una palabra (el atributo) en el documento. Tfidf asigna mayor peso a los atributos que aparecen más frecuentemente, es decir, los considera representativos del contenido del documento, pero además equilibra esto reduciendo el peso de un atributo si aparece en muchos documentos, es decir, es menos discriminante.

Para evaluar la calidad de la clasificación de cada modelo en cada variación del corpus, algoritmo, y parámetros de preprocesamiento hemos examinado dos medidas: precisión y F1. La precisión (*accuracy* o porcentaje de documentos correctamente clasificados) es la más utilizada para evaluar algoritmos de aprendizaje, pero por si sola puede no ser la más representativa en la categorización de documentos ya que cada clase típicamente tiene muchos más ejemplos negativos que positivos. Por ello se utiliza F1 que combina con igual peso *precision* (número de documentos correctamente categorizados como pertenecientes a una determinada clase dividido por el número total de documentos etiquetados como pertenecientes a tal clase) y *recall* (número de documentos correctamente categorizados como pertenecientes a una determinada clase dividido por el número total de documentos que realmente pertenecen a esa clase, es decir, incluyendo los que el algoritmo no haya conseguido identificar como tales). F1 se calcula promediando las F1 de cada una de las clases (*macro-averaging*) (Yang & Joachims, 2008).

4.4. Resultados y discusión

La Figura 5 muestra los valores de la medida F1 en función del número de ejemplos de entrenamiento disponibles. F1 toma valores entre 0 y 1. Los resultados usando como métrica la precisión—

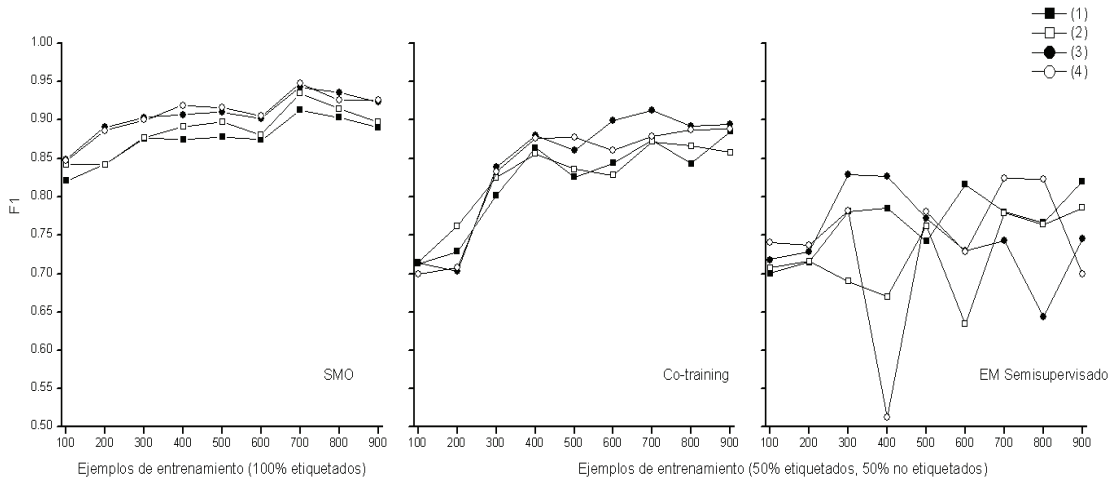


Figura 5: Variación de F1 con el número de ejemplos de entrenamiento según los mecanismos de reducción de la dimensionalidad y de asignación de pesos a los atributos.

accuracy—son muy similares y no se muestran aquí. SMO precisa ejemplos etiquetados; *co-training* y EM, al ser semi-supervisados, pueden aprovechar ejemplos no etiquetados también, por lo que en el experimento la mitad del conjunto de entrenamiento estaba etiquetada y la otra no. Para cada algoritmo modificamos los mecanismos de reducción de la dimensionalidad y de asignación de pesos a los atributos. En los casos (3) y (4) se seleccionaron los 100 atributos más relevantes para cada una de las 21 clases, resultando en un total de entre 614 y 933 atributos (dependiendo del conjunto de entrenamiento); en (1) y (2) los 100 atributos más relevantes para todas las clases. Solamente en los casos (2) y (4) se aplica la transformación tfidf.

En general los resultados son mejores (de 1% a 6% mejores en el caso de SMO y *co-training*) cuando se eligen los atributos más relevantes para cada una de las 21 categorías (líneas 3 y 4). Esta selección lleva a un conjunto de atributos mucho mayor, lo cual aumenta el tiempo de aprendizaje del modelo considerablemente en el caso de *co-training*, aunque no así en el caso de SMO. No obstante, dado que el modelo sólo se construye una vez, éste no es un factor tan importante a la hora de elegir el algoritmo. Por otro lado, la utilización de tfidf (gráficos 2 y 4) supone algo de mejora solamente cuando se utiliza SMO.

La Figura 6 (a) compara los valores de F1 para los tres algoritmos en el caso (3), es decir, seleccionando los 100 atributos para cada una de las 21 clases y sin usar tfidf. Puede verse cómo SMO produce

mejores resultados, especialmente con menos ejemplos de entrenamiento disponibles. Como se indicó, el conjunto de entrenamiento de los algoritmos semisupervisados estaba sólo parcialmente etiquetado. Por lo tanto, en cierto modo SMO tenía a su disposición más información (más ejemplos etiquetados). Por ello además comparamos en la Figura 6 (b) los tres algoritmos solamente frente al número de instancias de entrenamiento etiquetadas. (Los algoritmos semi-supervisados reciben además un número igual de instancias no etiquetadas). En este caso no hay tanta diferencia en el rendimiento de SMO y *co-training*, pero vemos que el uso de los ejemplos no etiquetados no supone una ventaja en este dominio (a diferencia de otros dominios como los explorados en (Pérez y Cardoso, 2009)).

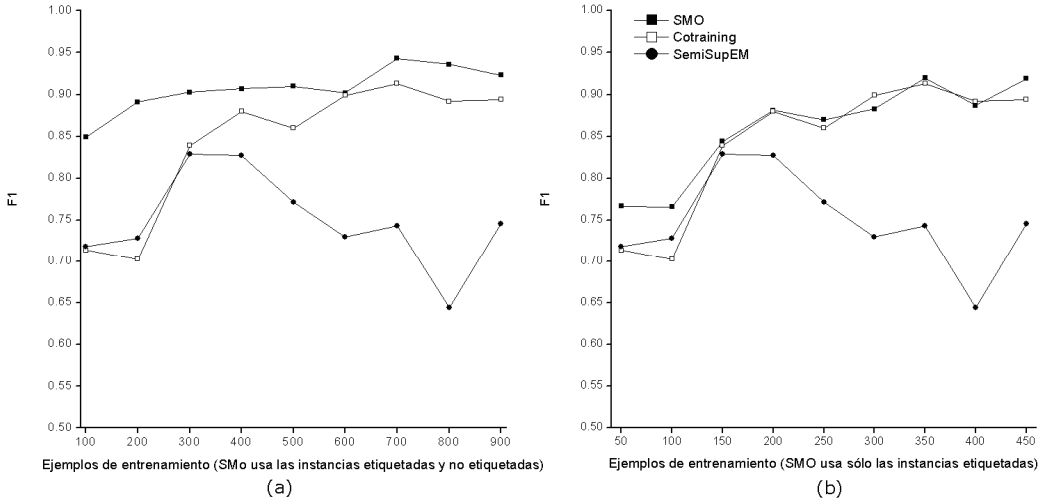


Figura 6: Comparación de los tres algoritmos.

5. Búsqueda semántica

Nuestro objetivo final es construir un buscador semántico (Guha et al, 2003) que utilice los meta-datos obtenidos automáticamente por los anotadores implementados en UIMA y las categorías asignadas automáticamente por los algoritmos de aprendizaje. Para ello hemos experimentado adaptando dos APIs de soltare libre diferentes para implementar motores de búsqueda: SemanticSearch y Lucene. Mientras que la segunda es más ampliamente utilizada para construir motores de búsqueda en general, la primera ha sido especialmente construida para su integración con UIMA, ya que como ésta última fue desarrollada por IBM. Nuestro prototipo inicial ha sido por ello construido sobre SemanticSearch y ofrece una interfaz básica para el buscador semántico. Una segunda implementación con Lucene es la base del

trabajo de un alumno de grado que ha colaborado en nuestro proyecto y está construyendo una interfaz amigable y más completa basada en web para el buscador.

Esta sección comienza exponiendo los fundamentos de la búsqueda semántica y de los motores de búsqueda. A continuación, en la sección 5.3 se describe el lenguaje *XML Fragments* que utilizamos para realizar consultas. Las secciones 5.4 y 5.5 describen SemanticSearch y el prototipo que hemos construido utilizando esta herramienta. Las secciones 5.6 y 5.7 describen Lucene y las pruebas que hemos hecho como base de un motor de búsqueda basado en esta herramienta.

5.1. Búsqueda semántica

Los motores de búsqueda tradicionales indexan las palabras que aparecen en los documentos y procesan consultas que son combinaciones booleanas de palabras. Después devuelven una lista de documentos que contienen esa combinación de palabras ordenada según diversos parámetros.

Por otro lado, muchas aplicaciones UIM pueden beneficiarse de más inteligencia en esta búsqueda, con consultas que buscan documentos no solamente basándose en las palabras que aparecen en el documento, sino en el contexto en que estas aparecen y en conceptos derivados de la interpretación del texto por los anotadores y que han sido asociados al documento en forma de anotaciones. El motor de búsqueda semántica es capaz de construir su índice no sólo con las palabras del texto sino también con las anotaciones. Obviamente la interfaz del motor de búsqueda debe permitir al usuario elaborar consultas que utilicen esas anotaciones.

La Figura 7, extraída de (Hampp & Lang, 2005), ubica la búsqueda semántica en el contexto de los paradigmas de búsqueda. Cada una de las tecnologías está construida sobre las que aparecen bajo ella. La búsqueda semántica se apoya en los elementos de la búsqueda tradicional (*keywords*) y sus operadores (+ y -, AND, OR, etc) y añade la facilidad para buscar conceptos o entidades y las relaciones entre ellos. Para ello introduce nuevos elementos, como los lenguajes *XML Fragments* (ver sección siguiente) o *XPath* para expresar las consultas. Las aplicaciones suelen ocultar esta sintaxis de los usuarios. La búsqueda en lenguaje natural trata de encontrar documentos en respuesta a una pregunta directa del usuario final. Así se evita la complejidad de la consulta de la búsqueda semántica, pero puede

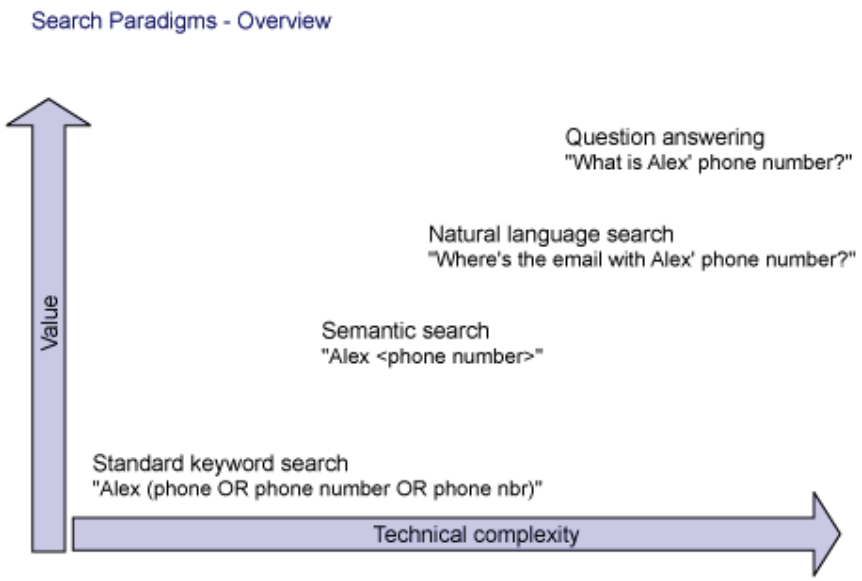


Figura 7: Comparación de paradigmas de búsqueda (Hampp & Lang, 2005).

interpretarse incorrectamente la pregunta del usuario. Suele consistir en el análisis textual de la consulta, presentada en lenguaje natural por el usuario, y su traducción automática en una consulta semántica en uno de los lenguajes mencionados. El resultado es una colección de documentos relevantes a la consulta. Finalmente, la respuesta a preguntas busca devolver la parte de un documento que incluye la respuesta. Puede también consolidar hechos procedentes de diversos documentos en una sola respuesta. Suele construirse basándose en la búsqueda en lenguaje natural, que devuelve un conjunto de documentos relevantes, con un postprocesamiento para extraer la respuesta precisa de esos documentos.

5.2. Fundamentos de un motor de búsqueda

Un motor de búsqueda almacena la información sobre una colección de documentos o páginas web que ha visitado previamente. En el caso de la web estas páginas son previamente obtenidas por un *Web crawler* (araña o gusano), navegador automatizado que va siguiendo todos los vínculos que encuentra en las páginas visitadas.

Si el motor de búsqueda permite diversos formatos, deben prepararse primero los documentos para la tokenización. Muchos documentos contienen información de formato además del contenido textual. Si esta información, por ejemplo etiquetas HTML, fuera incluida

en el índice, podría producir resultados espurios. A veces se denomina análisis de formato a la identificación y manejo del contenido de formato que está incluido en un documento de texto y que controla la forma en que el documento es presentado en una pantalla o navegador o interpretado por una aplicación (ej. un procesador de texto). Otros nombres de estas tareas son análisis de la estructura, eliminación de etiquetas, normalización del texto, limpieza de texto, o preparación del texto. Algunos formatos son propietarios y hay muy poca información disponible al respecto (ej. los utilizados por Microsoft Word o Excel), mientras que otros están ampliamente documentados.

Posteriormente los contenidos textuales de cada documento o página son analizados para determinar cómo indexarlos. Por ejemplo, podrían extraerse información especial de títulos, nombres de archivo, o campos especiales llamados meta-etiquetas, además del cuerpo del documento. Esta información se almacena en una base de datos o índice para ser utilizada por las consultas. El objetivo es recuperar la información relevante lo más eficientemente posible. Sin el índice, el motor de búsqueda tendría que escanear cada documento del corpus en cada consulta, lo que requeriría considerable tiempo y recursos. Por ejemplo, mientras que se puede consultar un índice de 10.000 documentos en milisegundos, recorrer cada palabra de 10.000 archivos llevaría minutos y hasta horas. Algunos motores de búsqueda almacenan en el índice todo o parte del documento.

Cuando un usuario realiza una consulta, el motor examina el índice y devuelve una lista de los documentos o páginas que mejor la satisfacen. Frecuentemente se devuelve también un breve resumen del documento con el título y partes del texto, resaltando las que responden a la consulta. La mayoría de los motores permiten operadores booleanos (AND, OR y NOT) en la consulta

5.3. El lenguaje XML Fragments

Para construir las consultas hemos adoptado el lenguaje de consultas *XML Fragments* (Chu-Carroll et al, 2006; Carmel at al, 2003) debido a su expresividad y a la disponibilidad de un motor de búsquedas como parte del componente *SemanticSearch* de la arquitectura UIMA (Hampp & Lang, 2005). La interfaz de búsqueda de *SemanticSearch* permite los operadores estándar de búsqueda, tales como búsqueda de texto libre y los operadores AND, OR, NOT y comodines para combinar las consultas.

Una consulta en el lenguaje *XML Fragments* consta de una estructura XML sub-especificada que combina consultas de palabras con consultas de información anotada. Esto permite buscar conceptos más específicos, e incluso relaciones entre objetos (por ejemplo, "la persona y la unidad académica deben aparecer en la misma frase" o "una persona que pertenece a cierta unidad académica" o "una unidad académica vinculada a una institución" porque aparecen cerca en el documento).

Algunos ejemplos sencillos de consultas serían:

- *ingenieros*: devuelve cualquier resolución que contenga la palabra "ingenieros"

- *<Institución/>*
  devuelve cualquier resolución que contenga una anotación de Unidad Académica

- *<Institución> Ingenieros </Institución>*
  devuelve cualquier resolución que contenga una anotación de Institución que contenga la palabra *ingenieros* (efectivamente encontrando las referencias a COPAIPA, el Consejo Profesional de Ingenieros, Agrimensores y Profesiones Afines)

- *<Institución> Ingenieros </Institución> pasantía*
  devuelve los documentos que respondan a la consulta anterior que además contengan la palabra *pasantía*

- *<UnidadAcadémica>ingen </UnidadAcadémica> <FechaResol año=2007> <FechaResol>*
  devuelve todas las resoluciones de la Facultad de Ingeniería e Informática del año 2007

- *<Institución> Ingenieros </Institución> pasantía <UnidadAcadémica>ingeniería </UnidadAcadémica>*
  resoluciones en que aparezcan la Facultad de Ingeniería e Informática, COPAIPA y la palabra *pasantía.*

- *<UnidadAcadémica> ingeniería </UnidadAcadémica> <clase categ=ConvenioPasantía </clase>*
  resoluciones en que aparece la Facultad de Ingeniería e Informática que han sido etiquetadas como convenios de pasantía por el clasificador aprendido.

Estos ejemplos se corresponden con las operaciones que se pueden realizar con *XML Fragments* definidas por Chu-Carroll et al (2006):

- conceptualización, que generaliza un literal (string) a un concepto apropiado del sistema de tipos. Por ejemplo, la consulta "*institución*" devuelve documentos en que aparezca esa palabra, mientras que <*Institución/*> busca ocurrencias de la anotación institución, aunque la palabra "*institución*" misma no aparezca en el documento.

- restricción: restringe las ocurrencias de etiquetas XML indicando qué palabras deben aparecer. Por ejemplo <*Institución*> *ingenier* <*/Institución*> devuelve ocurrencias del Consejo Profesional mientras que <*UnidadAcadémica*> *ingenier* <*/UnidadAcadémica*> devuelve documentos que se refieren a la Facultad de Ingeniería.

- relación: una anotación representa la relación entre los términos que aparecen en la consulta. Por ejemplo <*FechaResol*>*2007 Septiembre* <*FechaResol*> encuentra las resoluciones de septiembre del 2007 (y no simple resoluciones de 2007 en que aparezca la palabra septiembre).

Así *XML Fragments* permite aumentar la expresividad de las consultas de los usuarios y obtener resultados más focalizados en la búsqueda. Obviamente no se espera que el usuario final exprese sus consultas directamente en este lenguaje, sino proporcionarle una interfaz que le asista a formular la parte estructurada de las consultas, por ejemplo mediante formularios HTML donde las entradas están asociadas a elementos predefinidos correspondientes a las anotaciones, complementada con la búsqueda tradicional de *keywords*.

5.4. *SemanticSearch*

El paquete *SemanticSearch* 2.1 (http://www.alphaworks.ibm.com/tech/uima) añade a UIMA un motor de búsqueda semántica. Incluye un CAS Consumer que llena un índice con el contenido del documento así como las anotaciones generadas añadidas por los anotadores implementados en UIMA. Para consultar el índice puede usarse el lenguaje *XML Fragments* descrito en la sección anterior. Este paquete es software libre aportado por IBM y para acceder al índice existe una API para ser utilizada en una aplicación en Java.

Para construir un índice utilizando UIMA y *SemanticSearch* primero hay que correr un CPE (Collection Processing Engine) que incluye los anotadores y un CAS Consumer que toma los tokens y anotaciones, y además las oraciones del texto. Este CAS Consumer está disponible con la distribución de SemanticSearch y se llama SemanticSearchCasIndexer. Entre los anotadores debe haber uno que produce las anotaciones de tokens y oraciones.

```
<indexBuildSpecification>
  <indexBuildItem>
    <name>org.apache.uima.examples.tokenizer.Token</name>
    <indexRule>
      <style name="Term"/>
    </indexRule>
  </indexBuildItem>
  <indexBuildItem>
    <name>org.apache.uima.examples.tokenizer.Sentence</name>
    <indexRule>
      <style name="Breaking"/>
    </indexRule>
  </indexBuildItem>
  <indexBuildItem>
    <name>org.apache.uima.mio.Dni</name>
    <indexRule>
      <style name="Annotation">
            <attributeMappings>
                  <mapping>
                        <feature>dniNum</feature>
                        <indexName>dniNum</indexName>
                  </mapping>
            </attributeMappings>
        </style>
      </indexRule>
  </indexBuildItem>
……
 </indexBuildSpecification>
```

Figura 8: Ejemplo parcial del archivo de configuración de SemanticSearch.

Es necesario proporcionar un archivo de configuración (Index Build Specification) que indica cómo van a usarse las anotaciones para construir el índice. La Figura 8 muestra parte de este archivo para nuestra aplicación. Este archivo está formado por una colección de elementos (IndexBuildItems). Cada uno se refiere a un tipo de anotación y a un estilo. El primer item del ejemplo especifica que el tipo de anotación token debería ser indexado con el estilo Term. Esto significa que cada anotación token será considerada un token individual a la hora de la búsqueda. En el caso de las anotaciones relativas a nuestro dominio, todos los ítems serán del tipo *annotation*, que indica

que cada anotación será almacenada en el índice como un segmento de texto con nombre igual al nombre de la anotación y abierto a que se pueda buscar dentro de ese segmento. Además puede indicarse un campo (*feature*) específico de la anotación como el que se va a indexar.

Tras el procesamiento de la colección, el indexador construye el índice que puede ser consultado con tokens o con etiquetas XML según el lenguaje *XML Fragments*. Para más detalles sobre SemanticSearch y su integración con UIMA puede consultarse (Apache UIMA Development Community, 2008a, 2008b, 2008c; IBM, 2006).

5.5. Implementación del buscador semántico utilizando *SemanticSearch*

Nuestra primera aplicación del buscador semántico se ha desarrollado en el lenguaje de programación Java. Sus principales procesos son clasificar documentos, crear el índice para exploración de documentos y realizar búsqueda de resoluciones a partir de anotaciones y/o palabras. La Figura 9 muestra la sencilla interfaz del prototipo. El menú “Opciones” cuenta con los comandos “Crear Indice” y “Salir”. El primero recorre todos los documentos de la colección, obtiene las anotaciones y construye el índice de búsqueda. Cada vez que se incorporen nuevas resoluciones, el índice debe ser construido de nuevo.

El anotador UIMA utilizado, AnotadoresMasClase, es un anotador agregado que incluye los anotadores que reconocen personas, unidades académicas, carreras, instituciones, número y fecha de resolución y un nuevo anotador que utiliza el modelo generado en la etapa de entrenamiento y que determina a que categoría pertenece el documento que se analiza. Incluye también un anotador estándar de UIMA que produce palabras (tokens) y oraciones (*SimpleTokenAndSentenceAnnotator*).

La aplicación realiza también la búsqueda de resoluciones mediante consultas que pueden combinar palabras (*keywords*) y anotaciones. Para ello pueden utilizarse tres listas desplegables, donde aparecen los criterios de búsqueda y tres cuadros para introducir el texto a buscar. Las posibles formas de realizar consultas son:

- Con todas las palabras introducidas en el cuadro correspondiente
- Con algunas palabras

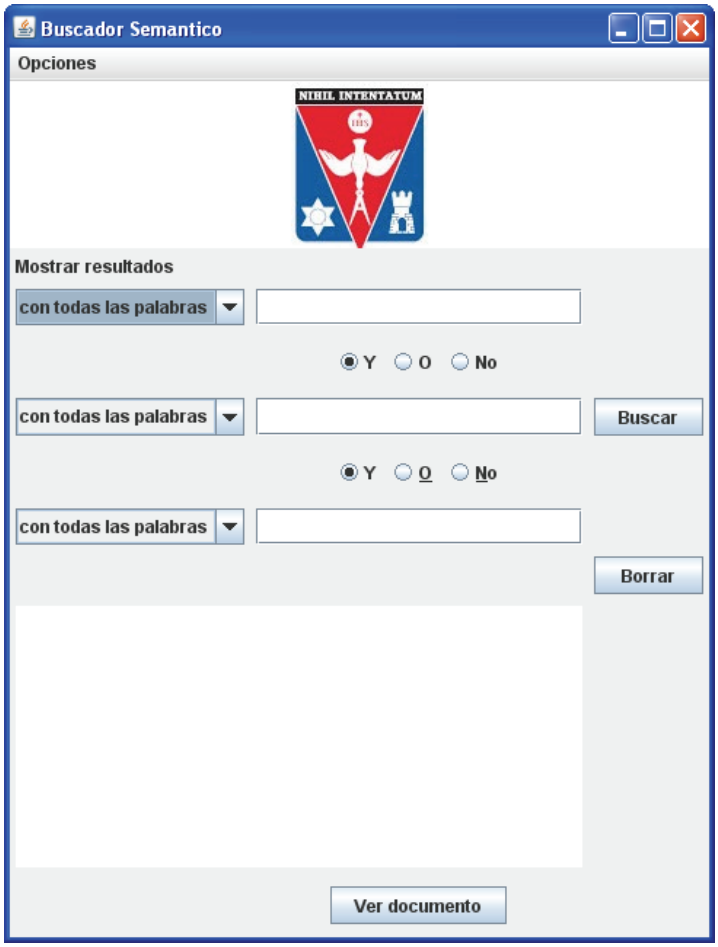


Figura 9: Interfaz del prototipo del buscador semántico.

- Con una frase exacta
- Sin determinadas palabras
- Sin determinadas palabras
- Por Nombre, Apellido, Institución, Unidad Académica, DNI, Título, Año de Resolución, Número de Resolución o Categoría de Resolución.

En el caso de que se realice búsqueda por tipo de resolución (Clase) se despliega otra lista con las categorías disponibles. Además los botones de opciones entre los cuadros de textos permiten realizar consultas más avanzadas, mediante operadores que combinan las anteriores. La Figura 10 muestra algunos ejemplos de consultas con los resultados correspondientes. Las consultas generadas por la interfaz son traducidas al lenguaje *XML Fragments*.

Para buscar documentos que simplemente contengan un cierto tipo de anotación, basta colocar "." en el cuadro correspondiente. Por ejemplo, si se quiere buscar resoluciones de cualquier Unidad Académica, se elige esta opción y en el cuadro correspondiente se escribe "."

La búsqueda se activa presionando el botón "Buscar" y los nombres de los documentos encontrados que cumplan con las combinaciones se mostrarán en el panel inferior. El botón "Ver documento" abrirá el documento seleccionado de esta lista. Las Figuras 11, 12 y 13 muestran ejemplos de consultas con el resultado obtenido y uno de los documentos propuestos abierto.

5.6. El proyecto Lucene

Apache Lucene es una librería escrita en Java para implementar motores de búsqueda en texto de alto rendimiento y con una gran gama de opciones. Se trata por tanto de una alternativa a SemanticSearch. Está siendo utilizada en múltiples proyectos y aplicaciones que precisan búsqueda en texto completo, especialmente portable a múltiples plataformas (Paul, 2004; Gospodnetić, 2004). Apache Lucene es un proyecto de software libre y abierto bajo la Licencia Apache que permite su uso en programas tanto abiertos como comerciales.

El centro de la arquitectura lógica de Lucene se encuentra el concepto de Documento (*Document*) que contiene Campos (*Fields*) de texto. Cada campo representa información sobre el texto que se desea indexar, y puede también incluir metadatos sobre el documento, tales como su autor, o la fecha de creación, aunque no sean campos en los que se va a realizar la búsqueda.

Esta flexibilidad permite a Lucene ser independiente del formato del archivo. Por ello cualquier archivo de texto (PDF, HTML, documentos de Microsoft Word, etc) puede ser indexado siempre que se pueda extraer la información textual del mismo.

Antes de la indexación, un campo de texto se convierte en la representación final del índice que es una colección de términos. Para ello primero se divide el texto en instancias de tokens de Lucene, que pueden después ser modificados mediante una serie de filtros (por ejemplo, para eliminar *stopwords).*

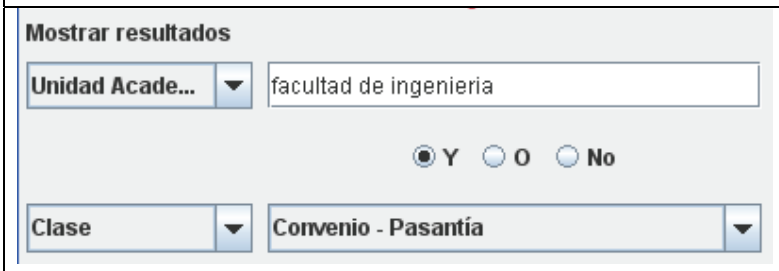
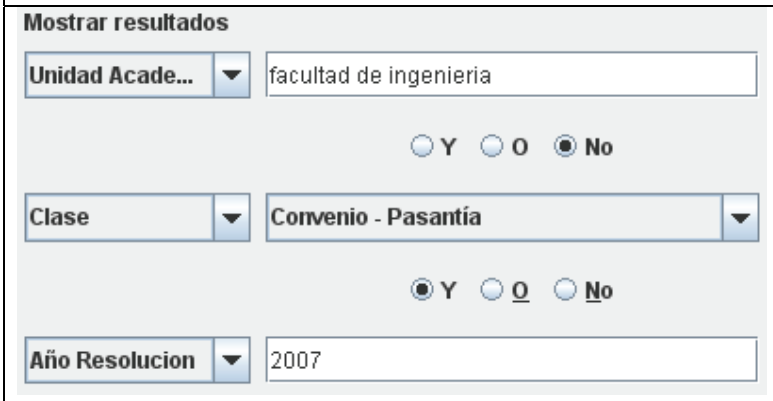
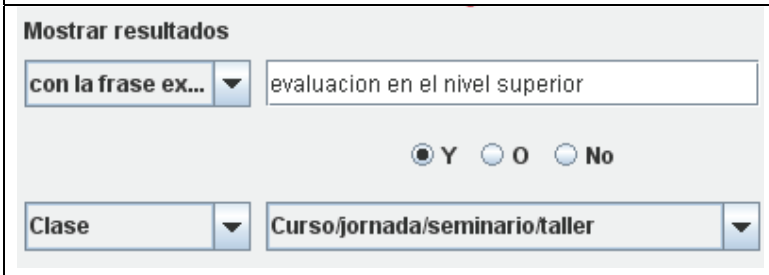
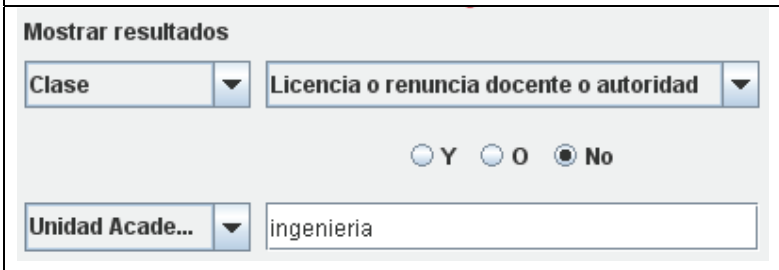
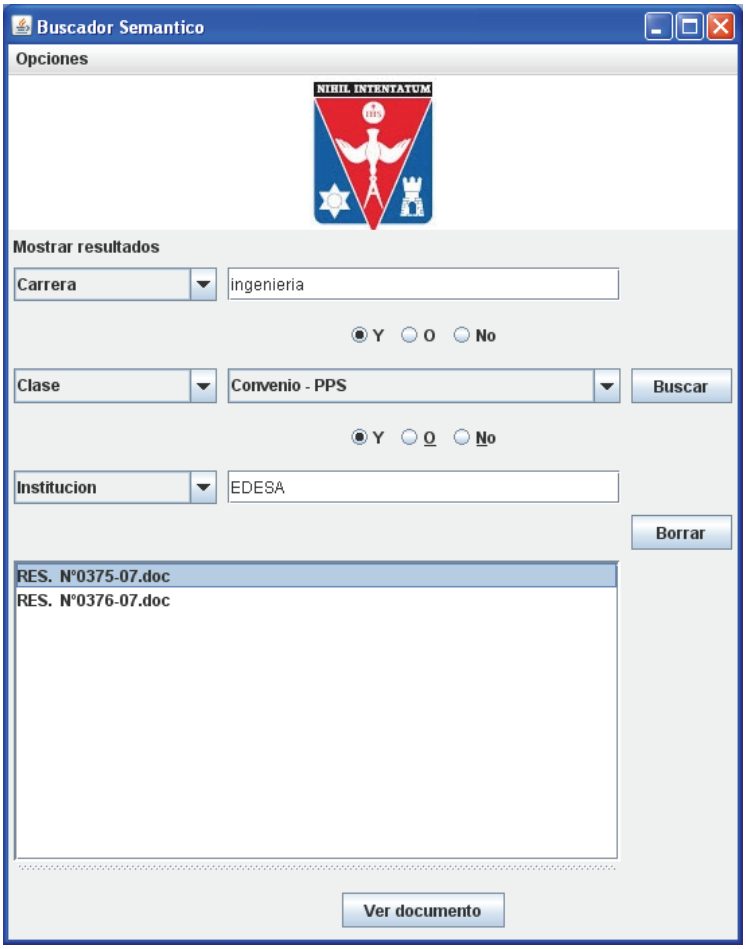
 *<Institucion>Arzobispado +de +salta</Institucion>*	Devuelve resoluciones que tengan como institución al Arzobispado de Salta
 *+<UA>facultad de ingenieria</UA>* *+<Clase valor="ConvPasant"></Clase>*	Devuelve resoluciones de la Unidad Académica Facultad de Ingeniería e Informática y que traten sobre convenios o pasantías
 *+<UA>facultad de ingenieria</UA>* *+<Clase valor="ConvPasant"></Clase>* *+<FechaResol anio="2007"></FechaResol>*	Devuelve resoluciones de la Facultad de Ingeniería que no traten sobre convenios ni pasantías pero que sean del año 2007
 *+"evaluacion en el nivel superior"* *+<Clase valor="Curso"></Clase>*	Devuelve resoluciones que contengan la frase exacta "evaluación en el nivel superior" y se refieran a un curso, jornada, seminario o taller
 *+<Clase valor="LicenRenunc"></Clase> +<UA>ingenieria</UA>*	Devuelve resoluciones sobre licencias o renuncias de docentes o autoridades que no sean de la Facultad de Ingeniería e Informática

Figura 10: Ejemplos de consultas en el prototipo.



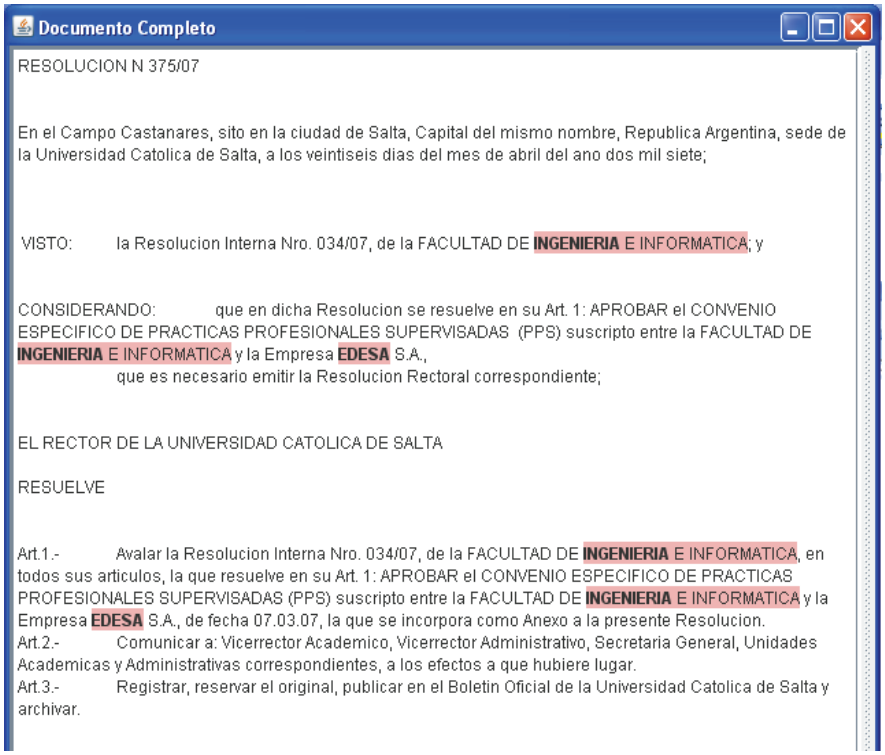
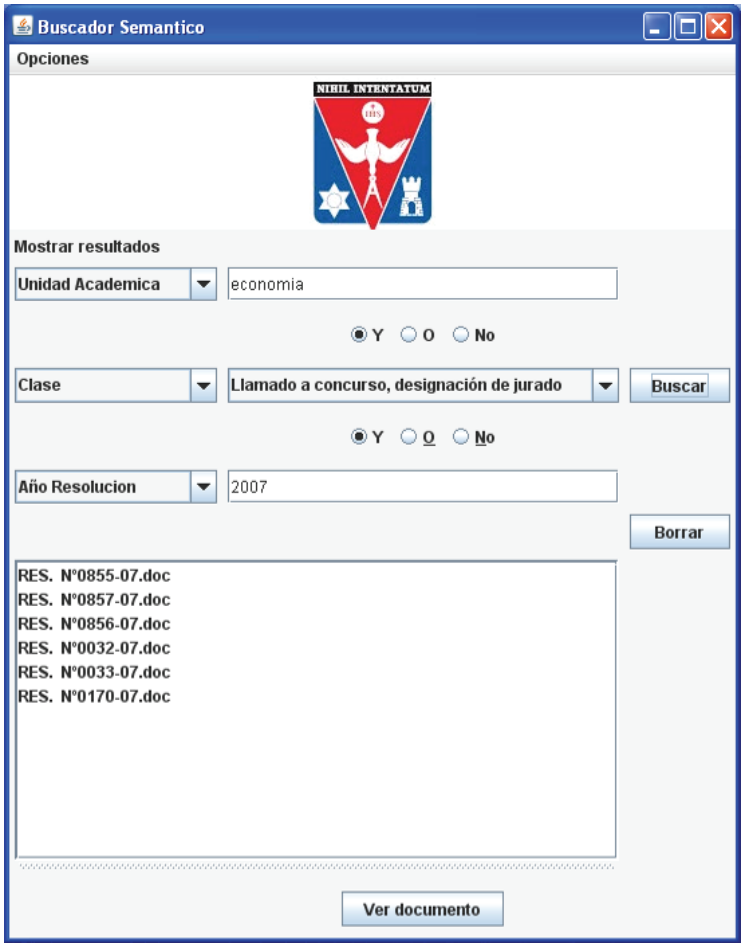


Figura 11. Ejemplo del resultado de la búsqueda de documentos.



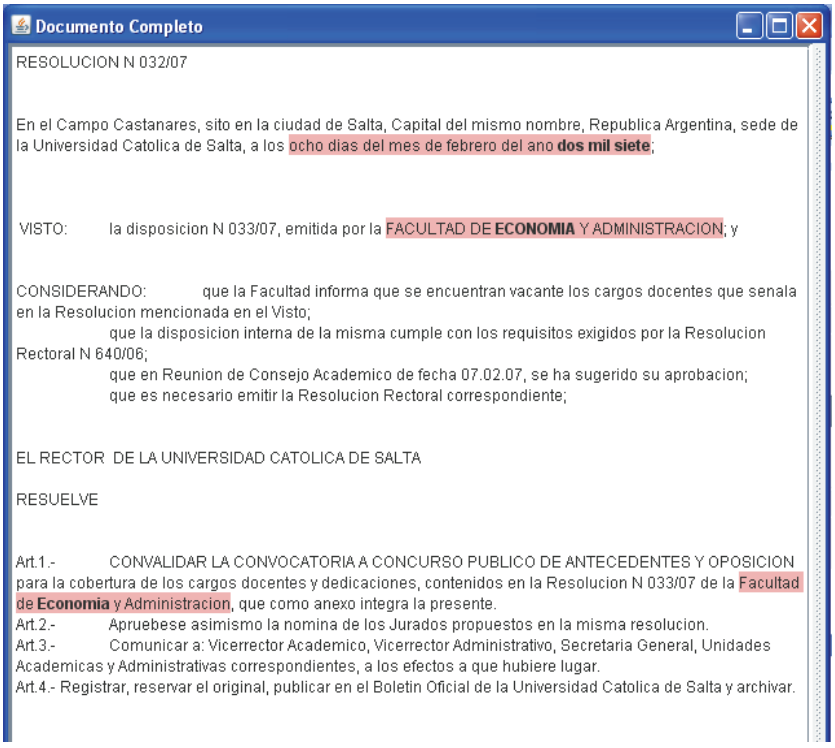


Figura 12. Ejemplo del resultado de la búsqueda de documentos.

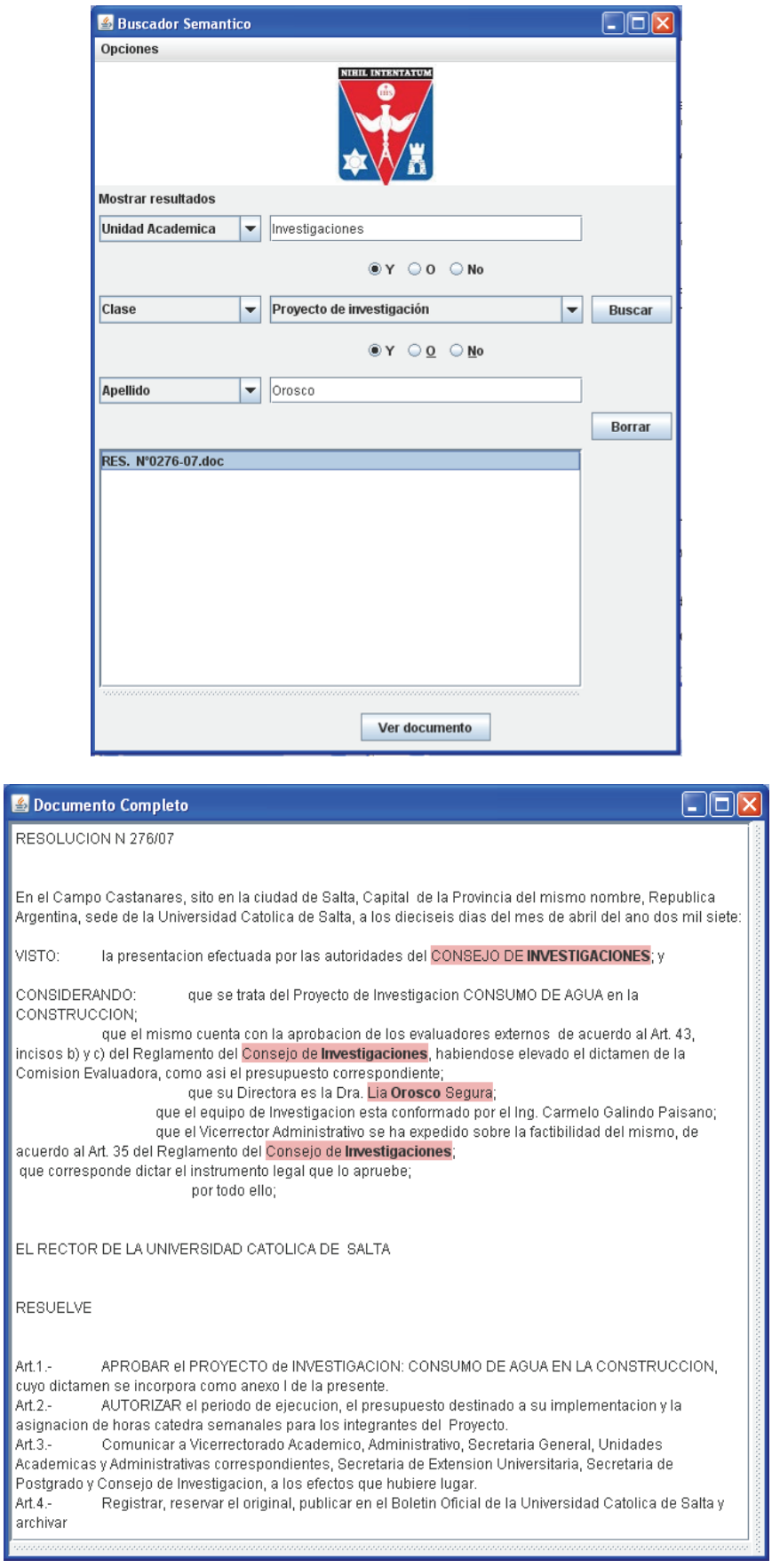


Figura 13. Ejemplo del resultado de la búsqueda de documentos.

5.7. Implementación utilizando Lucene

LuCas (Fäßler et al, 2009) es un componente para UIMA del tipo CAS Consumer que sirve de puente entre UIMA y la librería Lucene para construir motores de búsqueda. LuCas almacena los datos procedentes del CAS (anotaciones) en un índice de Lucene, para que después puedan ser recuperados de forma muy eficiente. LuCas ha sido recientemente incorporado a la sandbox de UIMA (http://incubator.apache.org/uima/ sandbox.html) y es software libre bajo la licencia Apache Software License.

LuCas va llenando automáticamente un índice de Lucene convirtiendo las anotaciones de UIMA en instancias de "documentos" de Lucene según un archivo de configuración que define el mapeo. Esto permite que Lucene recupere documentos no sólo en base a la información contenida en el texto sino también en base a los metadatos añadidos durante el procesamiento de UIMA en forma de anotaciones. La Figura 14 muestra una parte del archivo de configuración de LuCas. En este ejemplo, se crea una campo en el índice de Lucene llamado "institución" para cada anotación UIMA del tipo "org.apache.uima.mio.institucion". En el segundo ejemplo, la entrada en el índice de Lucene es a partir de uno de los campos (o *features*) "apellido" de la anotación de UIMA del tipo "org.apache.uima.mio.persona". la definición del tercer campo permite la indexación de todas las palabras del documento. Puede consultarse (Fäßler et al, 2009) para más información sobre LuCas.

La Figura 15 muestra una consulta al índice generado con Lucene. Utiliza una interfaz disponible como componente denominada Luke, pensada para la etapa de desarrollo con el fin de verificar el correcto funcionamiento del índice y las consultas. Estamos desarrollando una interfaz más amigable y robusta para el buscador semántico que esté basada en web. Esta interfaz está siendo construida sobre el motor construido con Lucene.

6. Conclusiones

La minería de textos, y la categorización de documentos en particular, son un campo de investigación y aplicación prometedor dado que más y más las organizaciones están interesadas en aprovechar el gran cuerpo de conocimiento no estructurado de que disponen. Las técnicas de recuperación de la información y de aprendizaje automático, combinadas con la potencia de la arquitectura UIMA y su reuso de

```
<field name="institucion" index="yes">
      <filter name="lowercase" />
      <annotations>
            <annotation sofa="_InitialView"
                        type="org.apache.uima.mio.Institucion"
                        tokenizer="standard" />
      </annotations>
</field>

<field name="apellido" index="yes">
      <filter name="lowercase" />
      <annotations>
            <annotation sofa="_InitialView"
                        type="org.apache.uima.mio.Persona"
                        tokenizer="standard" />
                  <features>
                        <feature name="apellido" />
                  </features>
      </annotations>
</field>

<field name="content" index="yes" stored="no">
      <filter name="lowercase" />
      <filter name="stopwords" filePath="stopwords.txt"
              ignoreCase="true" />
      <annotations>
        <annotation sofa="_InitialView"
                    type="org.apache.uima.TokenAnnotation"
                    tokenizer="standard" />
      </annotations>
</field>
```

Figura 14: Ejemplo parcial del archivo de configuración de LuCas.

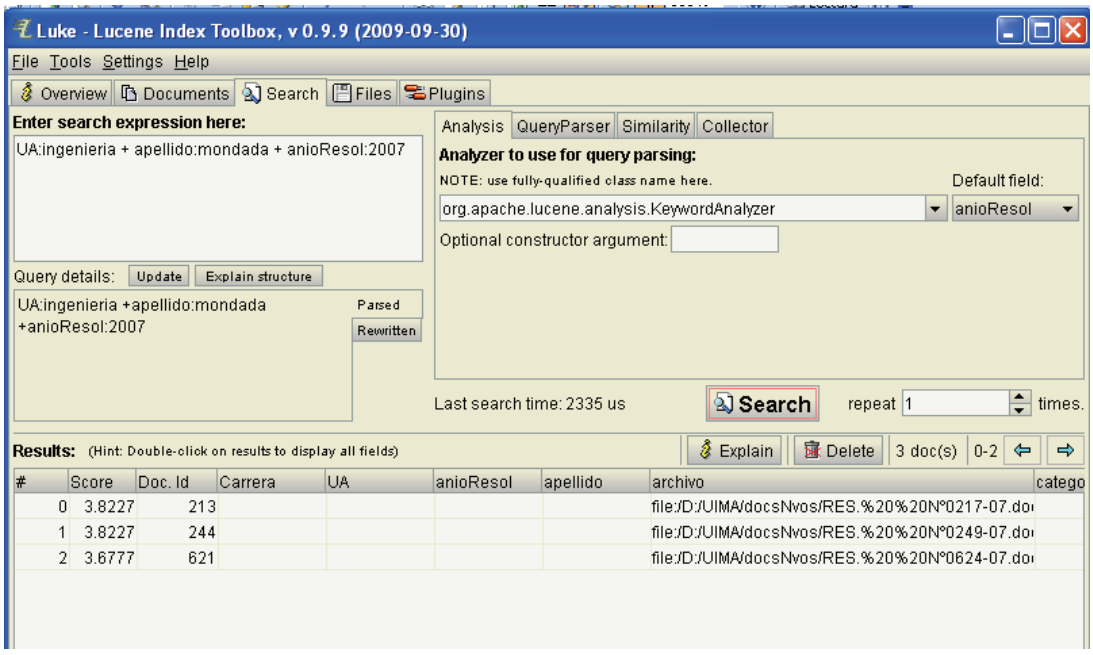


Figura 15. Resultado de una búsqueda con Lucene utilizando la interfaz Luke.

componentes, facilitan la tarea de extracción de conocimiento. En particular, hemos investigado e implementado anotadores que extraen del texto entidades, relaciones entre ellas, y la información necesaria para asignar automáticamente categorías a documentos de texto en base a un corpus relativamente pequeño de ejemplos.

El enfoque utilizado para la extracción de entidades, en particular instituciones y carreras, funciona relativamente bien, pero su calidad depende de la calidad del diccionario y de la precisión en la equiparación de las ocurrencias del texto con las del diccionario. A menudo una misma entidad, por ejemplo el nombre de una carrera, aparece de muy diversas formas en el texto. El diccionario debe incluirlas todas o variaciones suficientemente cercanas. Una clara posibilidad de trabajo futuro es flexibilizar este enfoque, posiblemente aplicando algoritmos de aprendizaje automático para detectar estas entidades.

La clasificación automática de documentos utilizando el modelo aprendido es de bastante calidad comparada con las etiquetas asignadas manualmente. En base a los experimentos realizados, el algoritmo semi-supervisado *cotraining* tiene un rendimiento bastante bueno en cuanto a los resultados de la clasificación y requiere menos ejemplos etiquetados para aprender al poder aprovechar los ejemplos no etiquetados. No obstante los modelos aprendidos por el algoritmo SMO (máquinas de vectores soporte) son muy buenos para este problema sin necesidad de tener muchos ejemplos etiquetados de entrenamiento. Por tanto lo hemos elegido para la implementación en este problema.

Hemos desarrollando un prototipo que integra todas las anotaciones poniéndolas a disposición de un buscador semántico y así en última instancia de un usuario que pueda efectuar consultas y visualizar los documentos resultantes de las mismas mediante una interfaz apropiada. El prototipo utiliza la API SemanticSearch y presenta una interfaz básica suficiente como prueba de concepto. Se está construyendo un sistema más robusto y una interfaz más amigable utilizando la API Lucene en un proyecto de grado asociado a este trabajo de investigación.

El conocimiento y la experiencia obtenida de este proyecto son base para futuras implementaciones que pueden hacer extensivas estas técnicas a otros problemas de clasificación e interpretación de textos, y en una perspectiva más amplia a otros problemas de gestión del conocimiento.

Agradecimientos

Este trabajo ha sido financiado en parte por la Convocatoria 2007 del Consejo de Investigaciones de la Universidad Católica de Salta (Resolución Rectoral 723/08). Las autoras agradecen la colaboración de la Secretaria General de la UCASAL, Prof Constanza Diedrich por su asesoramiento sobre la taxonomía de resoluciones rectorales. También agradecen la colaboración de los alumnos David Zamar e Iván Ramos en diversas etapas del trabajo

Bibliografía

Alias-I, 2008, LingPipe 3.8.2. http://alias-i.com/lingpipe (acceso 1 Septiembre 2009).

Apache UIMA Development Community, 2008, *UIMA Tutorial and Developers' Guides*, version 2.2, http://incubator.apache.org/uima/downloads/releaseDocs/2.2.2-incubating/docs/html/

Apache UIMA Development Community, 2008, *UIMA Tools Guide and Reference*, version 2.2, http://incubator.apache.org/uima/downloads/releaseDocs/2.2.2-incubating/docs/html/

Apache UIMA Development Community, 2008, *UIMA References*, version 2.2, http://incubator.apache.org/uima/downloads/releaseDocs/2.2.2-incubating/docs/html/

Blum, A. & Mitchell, T., 1998, Combining labeled and unlabeled data with co-training. En *Proceedings of the Eleventh Annual Conference on Computational Learning theory, COLT' 98*. ACM, New York, NY, 92-100.

Carmel, D., Maarek, Y. S., Mandelbrod, M., Mass, Y., & Soffer, A., 2003, Searching XML documents via XML fragments. En *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval* (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM, New York, NY, 151-158. DOI= http://doi.acm.org/10.1145/860435.860464.

Chu-Carroll, J., Prager, J., Czuba, K., Ferrucci, D., y Duboue, P., 2006, Semantic search via XML fragments: a high-precision approach to IR. En *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, SIGIR '06. ACM, New York, NY, 445-452.

Fäßler, E., Landefeld, R., Tomanek, K. & Hahn, U., 2009, LuCas - A Lucene CAS Indexer. En Proceedings of the 2nd UIMA@GSCL Workshop (German Society for Computational Linguistics and Language Technology)

Feldman, R. & Sanger, J., 2007, *The Text Mining Handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press, New York.

Ferrucci, D. & Lally, A. 2004. Building an example application with the unstructured information management architecture. *IBM Systems Journal* 43, 3, 455-475.

Gospodnetić, O. & Hatcher, E., 2004, *Lucene in Action*, Manning, Greenwich, CT

Hampp, T. & Lang, A., 2005, Semantic search in WebSphere Information Integrator OmniFind edition: The case for semantic search. *IBM Developer Works*

IBM, 2006, *UIMA SDK Semantic Search Engine: Search and Index API (SIAPI)*

Moore, C., 2002, Diving into data, *Infoworld*, 25 de octubre, http://www.infoworld.com/article/02/10/25/021028feundata_1.html

Nigam, K. & Ghani, R., 2000, Analyzing the effectiveness and applicability of co-training. En *Proceedings of the Ninth international Conference on information and Knowledge Management. CIKM '00.* ACM, New York, 86-93.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T., 2000, Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39 (2-3) 103-134.

OMG, 2007, XML Metadata Interchange (XMI), v 2.1.1

Paul, P., 2004, The Lucene Search Engine - Adding search to your applications, *JavaRanch Journal* April 2004. http://www.javaranch.com/journal/index.jsp?p=2004

Pérez, M. A. & Cardoso, A. C., 2009, Comparación de algoritmos de aprendizaje semi-supervisado. En *V Jornadas de Ciencia y Tecnología de Facultades de Ingeniería del NOA*, Salta, Septiembre 2009.

Sebastiani, F., 2005, Text categorization. En A. Zanasi (ed.) *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, WIT Press, Southampton, UK, 109-129.

Tomanek, K & Wermter, J., 2008, JULIE Lab Lingpipe Gazetteer Annotator Version 2.1. Jena University Language & Information Engineering (JULIE) Lab. Disponible en www.julieblab.de

Yang, Y. & Joachims, T., 2008, Text categorization, *Scholarpedia*, 3(5):4242

Witten, I. & Frank, E., 2005, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann.

Diseño y construcción de una planta piloto multipropósito de CO_2 supercrítico

Gerardo Tita*, M. Cornejo* y A. Ambrogi †

gerardotita@hotmail.com

Resumen

El uso de CO_2 (dióxido de carbono) supercrítico varía desde la producción de extractos naturales, sin el uso de los solventes orgánicos tradicionales, hasta el desarrollo de nuevos materiales. Esta amplia variedad de utilidades lo convierte en un proceso ideal principalmente para las industrias farmacéutica, cosmética y alimenticia. En este trabajo se presenta una planta de CO_2 supercrítico a escala piloto y su puesta en marcha. Esta planta opera a una presión máxima de trabajo de 50 MPa, una temperatura de 80º C y un caudal regulable de hasta 20 kg/hr de CO_2 funcionando con un circuito cerrado lo que permite reciclar el solvente. Tiene un extractor de 4 litros de capacidad lo que la convierte, en una planta piloto con una versatilidad interesante. Se presentan, además, las primeras curvas de extracción obtenidas en las que el rendimiento máximo posible es comparable a otras plantas de menor escala.

1. Introducción

Desde la década de los 80 que se trabaja a escala industrial con el dióxido de carbono supercrítico como solvente [1] en procesos como la descafeinización del café, la extracción de lúpulo y la extracción de especias [2]. Sin embargo en los últimos diez años hubo una fuerte aparición de nuevas aplicaciones para esta tecnología, que se realizan a escala piloto en todo el mundo. Prueba de eso son las numerosas publicaciones que nombran aplicaciones farmacéuticas para la formulación de nuevas drogas, aplicaciones ambientales para la

* Gerardo Tita: Estudiante de Ingeniería Industrial, Facultad de Ingeniería e Informática UCASAL y miembro del I.ES.I.ING.
María Cornejo: Analista químico y bromatóloga. Docente en la Facultad de Ingeniería e Informática UCASAL y miembro del I.ES.I.ING.
† Alejandor Ambroggi: C.I.T.TEC., Universidad Nacional de Río Cuarto y UCASAL.

recuperación de metales y la determinación de algunos contaminantes en el suelo [3].

Una de las aplicaciones industriales mas difundidas es la de extracción de aromas y colorantes naturales de matrices vegetales, principalmente por el alto valor del producto terminado y las exigencias de los mercados internacionales en cuanto a la presencia de restos de solventes orgánicos. En este trabajo se presentan el diseño y la construcción de una planta piloto orientada inicialmente para la extracción de oleorresina de pimentón. Asimismo se presentas las primeras curvas de extracción obtenidas con dos presiones de trabajo diferentes.

2. Fluidos Supercríticos

Una sustancia pura está en estado supercrítico cuando la misma se encuentra a valores de presión y temperatura mayores que ciertos valores característicos llamados críticos (Pc y Tc respectivamente). La Figura 1 representa un diagrama de presión-temperatura típico para una sustancia pura. Allí Tc y Pc representan la máxima temperatura y presión a la cual una sustancia pura puede existir en equilibrio líquido-vapor.

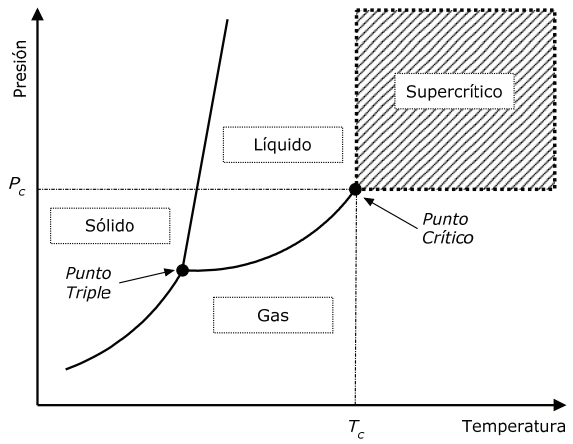


Figura 1: Diagrama Presión – Temperatura para una sustancia pura. Tc: temperatura crítica. Pc: Presión crítica

En este diagrama de equilibrio de fases se puede observar que no hay línea alguna que delimite la zona del estado supercrítico. Si se calienta una mezcla líquido-vapor a volumen constante, la densidad del líquido disminuye y la del gas aumenta hasta que en el punto crítico ellas se vuelven iguales y el menisco -o interfase que las separa- desaparece. Cuando la mezcla se aproxima al punto crítico comienzan a producirse fluctuaciones en la densidad de ambas fases en regiones de dimensiones microscópicas dando lugar a un fenómeno de dispersión lumínica típico conocido con el nombre de "opalescencia crítica".

3. Materiales y Métodos Experimentales

3.1. Proceso de extracción supercrítica

El proceso de extracción supercrítica es relativamente sencillo de realizar. Las variables a controlar son caudal de CO_2, temperatura y presión. Sin embargo, los equipos necesarios para manejar elevadas presiones del orden de los 50 MPa, son los que hacen que el proceso sea más complejo.

El circuito básico presentado en la Figura 2 consiste en cuatro etapas principales que son: extracción, expansión, separación y recuperación del solvente. A su vez, los cuatro componentes críticos necesarios en una planta son: el extractor de alta presión, una válvula reductora de presión, un separador y una bomba para elevar la presión del solvente reciclado [4].

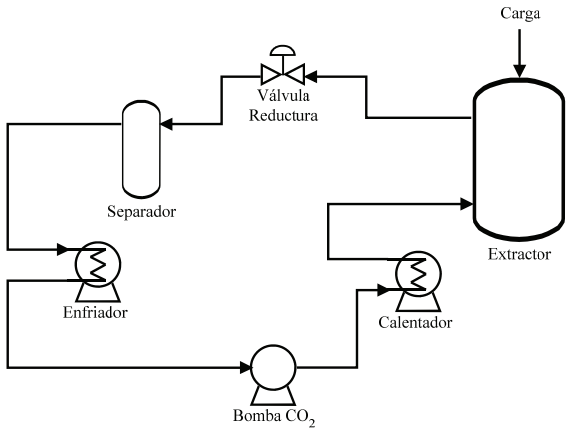


Figura 2: Esquema del Proceso de Extracción supercrítica

La Figura 3 muestra el ciclo de circulación del solvente en un diagrama T-S esquemático. El CO_2 líquido es cargado desde un cilindro exterior a un cilindro buffer en el circuito donde se lo enfría hasta el punto A, desde ahí la bomba lo comprime isoentrópicamente hasta los 50 MPa (punto B), y se lo calienta hasta la temperatura de extracción (punto C). Luego el cambio del punto de extracción (punto C) al punto final de separación (punto F) requiere que se disminuya la presión por medio de una válvula reguladora y se caliente para asegurarnos que el solvente esté en el estado gaseoso y se produzca la precipitación del soluto. El ciclo vuelve a comenzar al enfriar el CO_2 gaseoso (punto F) hasta la fase líquida (puntos F-E-D). Luego el CO_2 líquido y libre de extracto es retomado por la bomba.

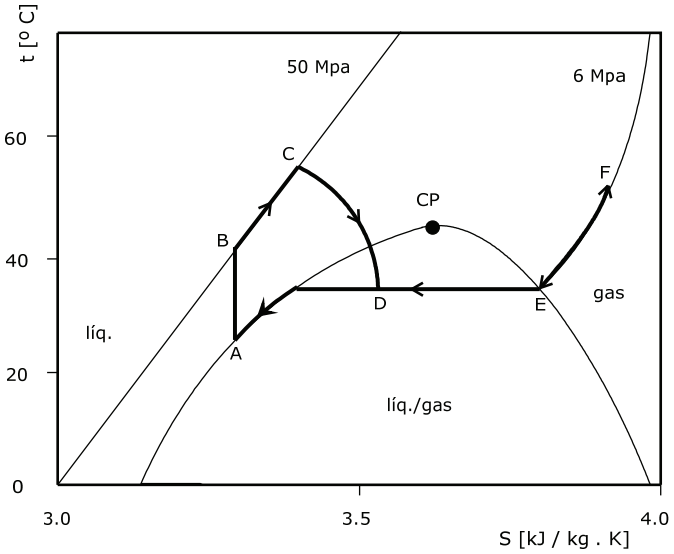


Figura 3: Diagrama T-S – Ciclo de extracción supercrítica

El proceso a escala piloto se realiza por batch, mientras que a escala industrial o semi-industrial se puede trabajar en procesos semicontinuos con extractores en paralelo. Esto es debido, principalmente, a que existen algunas limitaciones tecnológicas que hacen que la despresurización del extractor, al terminar la extracción, deba hacerse lentamente. Para solucionar esto, se utiliza más de un extractor de manera tal que mientras uno trabaja el otro se va descargando [4, 5].

3.2. Planta piloto de CO_2 *supercrítico*

La planta piloto, diseñada y construida en la Universidad Católica de Salta, se compone de tres recipientes a presión: un extractor, un separador y un buffer, todos ellos realizados en tubos sin costura de acero SAE 4340. La capacidad de cada recipiente es de 4 L; 0,7 L y 10 L respectivamente. Los tubos de acero fueron confeccionados especialmente para este proyecto por la firma Aceros Zapla S.A. La bomba que presuriza el sistema es del tipo neumática a pistón pudiendo llegar a los 70 Mpa y mantener un caudal de 60 L/hr.

Hay dos circuitos diferentes según la presión de trabajo de cada uno: el primero es para una presión de 50 MPa que va desde la salida de la bomba hasta la entrada de la válvula reguladora y, el segundo, va desde la salida de la válvula reguladora hasta la entrada de la bomba con una presión de trabajo de 6 MPa.

3.3. Puesta a punto

Las primeras pruebas de nuestra planta piloto fueron realizadas con pimentón cultivado y molido en los valles calchaquíes de la provincia de Salta, del cual se extrajo una mezcla de resina y aceite esencial llamado oleorresina. Básicamente, esta oleorresina se extrae del tejido exterior del fruto o pericarpio por el contacto con el dióxido de carbono en estado supercrítico, usado a modo de solvente, seguido de una descompresión rápida a su estado gaseoso liberando el solvente del extracto por precipitación. El pimentón inicialmente fue secado al aire libre, en canchas de secado, luego molido con un molino de bolas en su lugar de origen. Al ser recibido en la planta, se realizó un análisis de composición del tamaño de partícula por tamizado y un análisis del contenido de humedad.

En la puesta a punto se realizaron dos ensayos, con presiones de extracción de 30 y 50 MPa cada uno, mientras que las temperaturas del CO_2 se mantuvieron constantes tanto en la etapa de la extracción (80 °C) como en la de separación (40 °C). Durante toda la extracción se utilizó un caudal constante de CO_2.

4. Resultados y discusión

Durante la extracción se pueden distinguir dos o tres subetapas, según los distintos autores [4,6]. En la primera, llamada de lavado, el fenómeno de solvatación juega un papel muy importante, haciendo que la extracción sea veloz traduciéndose en una curva de extracción que crece casi linealmente en el tiempo. Esto es debido a que los aceites de

la matriz vegetal están muy disponibles a la extracción manteniendo constante la concentración del extracto en el solvente y, por lo tanto, saturando la capacidad de disolución. En esta subetapa el CO_2 disuelve la oleorresina residente en los cromoplastos más accesibles situados en la superficie de las partículas del pimentón molido.

En la Figura 4 se representan los mecanismos de transporte que actúan en la extracción: (1) Transporte convectivo de CO_2 desde el seno de la fase fluida a la superficie de la partícula, (2) difusión del CO_2 dentro de los poros y la fase sólida que conforma la partícula, (3) desorción del soluto desde la superficie sólida del poro y disolución del soluto en CO_2, (4) difusión del soluto (oleorresina) dentro de los poros, y (5) transporte convectivo del soluto al seno de la fase fluida.

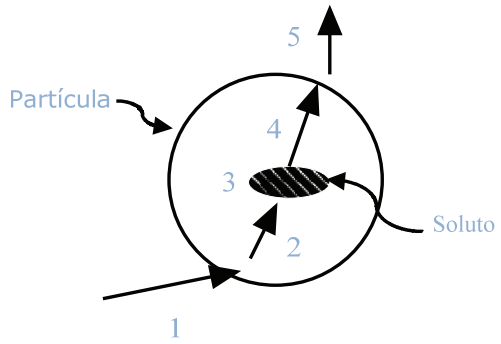


Figura 4: Esquematización de los mecanismos de transporte dentro de una partícula. 1-Transporte convectivo, 2-Difusión, 3-Desorción del soluto en el seno de la partícula, 4-Difusión y 5-Transporte convectivo.

A medida que el tiempo transcurre, la concentración del soluto disminuye debido a un aumento de la resistencia a la difusión de la matriz vegetal dando lugar a una última etapa gobernada por procesos difusivos. Aquí, el CO_2 se difunde dentro de la partícula de pimentón molido disolviendo la oleorresina contenida en los cromoplastos presentes en el seno de la partícula y difundiéndola hasta su superficie (ver Figura 4).

La tercera subetapa puede definirse como una intermedia de transición entre la primera de lavado y la última etapa difusiva.

Las Figuras 5 y 6 muestran las respectivas curvas obtenidas en los ensayos a 30 y 50 MPa. En el eje vertical se representa el

rendimiento de la extracción calculado como la relación entre la masa extraída y la masa total de extracto de la muestra, mientras que en el eje horizontal se indica el tiempo de extracción.

En las mismas Figuras 5 y 6 se representan, a modo de comparación, otras dos curvas obtenidas en un trabajo previo de Ambrogi et al. [7] (ver en referencia la descripción del equipo utilizado). Para el caso de la extracción a 50 MPa Ambrogi et al. observan que en la primera subetapa se extrae el 75 % del total de extracto disponible con solo el 15 % de solvente utilizado para la extracción total del soluto. Comparando estos datos con los obtenidos en nuestra planta, los valores de estos porcentajes se repiten con la diferencia de que en [7] el 100 % de la extracción se alcanza aproximadamente a los 160 min, mientras que en nuestro caso esta situación se logra luego de 360 min. Esta discrepancia en ambas curvas puede deberse quizás a una deficiencia en el sistema de condensación y recuperación del CO_2 (enfriamiento de la etapa 3 del ciclo termodinámico) de nuestra planta. Esta deficiencia se evidencia a partir del segundo proceso de extracción en adelante y no así en la primera extracción en donde la totalidad del CO_2 se encuentra en estado líquido (consideramos como el primer proceso de extracción al que se realiza cuando se llena el equipo con el cilindro de CO_2 por primera vez desde la entrega del proveedor). Actualmente se está trabajando en el mejoramiento del rendimiento de esta etapa.

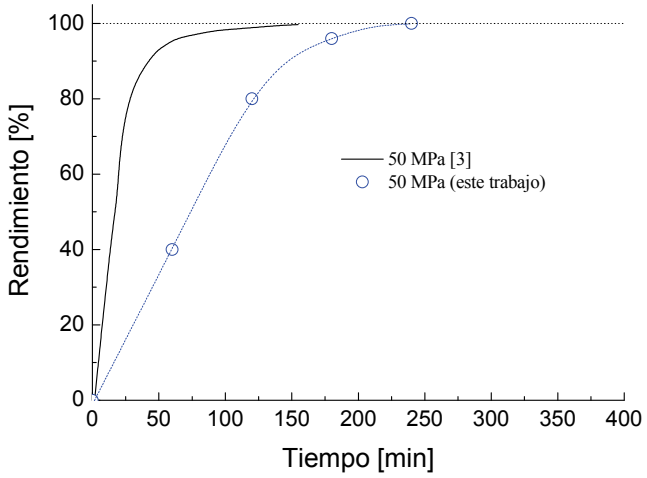


Figura 5: Extracción supercrítica de oleorresina de pimentón, 50 Mpa, 80 ºC

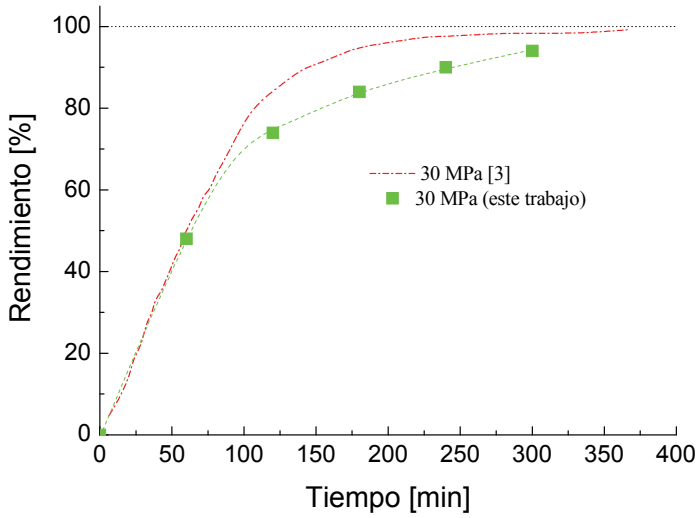


Figura 6: Extracción supercrítica de oleorresina de pimentón, 30 Mpa, 80 ºC

Para el caso de la extracción a 30 MPa ambas curvas presentan una notable similitud, sobre todo en la primera subetapa de lavado. La discrepancia observada en la segunda etapa puede deberse a la diferencia entre la materia prima usada en [7] y la usada en el presente trabajo, como ser: distinto origen de la materia prima, contenido de humedad, tamaño de partícula y cantidad de aceite original [4].

5. Conclusiones

Se ha diseñado y construido una planta de CO_2 supercrítico a nivel experimental con componentes de bajo costo. Las primeras pruebas muestran un comportamiento similar al de otras plantas comerciales construidas en el extranjero por empresas de I+D. Los tiempos de extracción son aceptables y satisfactorios, prueba de ello son la cantidad de extracto obtenido en unos minutos más o en el mismo tiempo que la planta comercial. Estos datos nos permiten decir que la planta piloto es apta para realizar estudios de balance de masa y energía, ensayos con una variedad muy importante de materiales diversos para obtener productos de alto valor comercial; y además abre el espectro de posibilidades para realizar tareas de I +D con plantas de CO_2 supercrítico, en la industria cosmética, farmacéutica y alimenticia.

La planta construida con bajo presupuesto puede ser utilizada en un sinnúmero de aplicaciones ya conocidas como así también para explorar otras nuevas aplicaciones como por ejemplo la recuperación de metales [8], extracción de elementos raros [9], y el secado de aerogel [10].-

Agradecimientos

Este trabajo fue realizado en el marco del proyecto “Producción de Oleorresina de pimentón de los valles calchaquíes de la provincia de Salta”, P.F.I.P. (Proyectos Federales de Innovación Productiva) financiado por el MinCyT, Ministerio de Ciencia y Técnica de la Nación. Se agradece al Dr. Osvaldo Andrade por su dirección en los inicios del proyecto, y también a la firma Aceros Zapla S.A. que por medio del Ing. Rafael Possetti (†) donó los aceros para la fabricación de los recipientes.

(†) Dedicamos este trabajo al Ing. Alejandro Ambrogi quien falleció durante su desarrollo, y dejó un vacío imposible de llenar.

Referencias

[1] Brogle, H, “CO_2 in solvent extraction”, *Chemistry and Industry*, 1982, 385-390.

[2] Palmer, M. V., Ting S. S. T., “Applications for Supercritical fluid technology in food processing”, *Food chemistry*, 52, 1995, 345-352

[3] Herrero, M., Mendiola J. A., Cifuentes A., Ibañez E., “Supercritical fluid extraction: Recent advances and applications”, *Journal of Chromatography*, 1217, 2010, 2495-2511.

[4] Mukhopadhyay, M, "Natural Extracts using supercritical carbon dioxide", *CRC Press*, 2000, pp 5-9

[5] Casas L., Mantell C., Rodriguez M., López E., Martínez de la Ossa E., “Industrial design of multifunctional supercritical extraction plant for agro-food raw materials”, *Chemical Engineering Transactions*, Vol. 17., 2009, 1585-1590.

[6] Taylor, L., *“Supercritical fluid extraction”*, John Wiley & Sons Inc., 1996, pp 101-126.

[7] Ambrogi, A, Cardarelli, D.A, Eggers, R, Fractional extraction of páprika using Supercritical Carbon Dioxide and on-line determination

of Carotenoids, *Journal of Food Science*, Vol. 76, nro 9, 2002, 3236-3241.

[8] Riviera de la Rosa, J., Barbarín-Castillo, J.M., Rodríguez-Díaz, J., Isidro-Almaguer, J.(2005): Solubilidad de cobre acomplejado en CO_2 supercrítico, *Tecnol. Ciencia Ed. 20*, 18-22.

[9] Shimizu R., Sawada K, Enokida Y. Yamamoto Y,. (2005: Supercritical fluid extraction of rare earth elements from luminescent material in waste fluorescent lamps, *J. of Supercritical Fluid 33*, 235-241.

[10] Tang Q., Wang T., (2005): Preparation of silica aerogel from rice hull ash by supercritical carbon dioxide drying, *J. of Supercritical Fluid 35*, 91-94.

La metodología de Kimball para el diseño de almacenes de datos (*Data warehouses*)

Gustavo R. Rivadera*

grivadera@ucasal.net

Resumen

Los almacenes de datos (data warehouses en inglés) toman cada día mayor importancia, a medida que las organizaciones pasan de esquemas de sólo recolección de datos a esquemas de análisis de los mismos. Sin embargo a pesar de la gran difusión de los conceptos relacionados con los almacenes de datos, no existe demasiada información disponible en castellano en cuanto a las metodologías para implementarlos. En este breve artículo intentaremos brindar una explicación general de una de las metodologías más usadas, la metodología de Kimball.

Palabras Claves: Metodologías de implementación de almacenes de datos- Almacenes de datos - Metodología de Kimball

1. Introducción

Un almacén de datos (data warehouse, DW) según Inmon (Inmon 02, Imhoff & Galemmo 03), es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se trata, sobre todo, de un historial completo de la organización, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos (especialmente con herramientas OLAP, de procesamiento analítico en línea). Por otra parte Kimball (Kimball 98) la define como "una copia de los datos transaccionales estructurados específicamente para consultas y análisis". Actualmente

* Ingeniero en Computación, desarrollador independiente de software, analista del Ministerio Público de la Provincia de Salta, docente de las Cátedras de Modelos y Simulación, Análisis Estratégico de Datos y Bases de Datos III, en la Facultad de Ingeniería e Informática, UCASAL. Actualmente cursa la Maestría en Ingeniería del Software en el Instituto Tecnológico de Buenos Aires (ITBA).

uno de los mayores impedimentos para construir este tipo de almacenes de datos es la falta de conocimiento de metodologías adecuadas para su implementación, y la disciplina para cumplirlas. En este breve artículo describiremos la metodología más utilizada actualmente: la metodología de Kimball†.

2. Metodologías actuales

Existen muchas metodologías de diseño y construcción de DW. Cada fabricante de software de inteligencia de negocios busca imponer una metodología con sus productos. Sin embargo, se imponen entre la mayoría dos metodologías, la de Kimball y la de Inmon. Para comprender la mayor diferencia entre estas dos metodologías, debemos explicar además de la noción de DW mencionando en la introducción, la idea de Data mart. Un Data mart (Kimball et al 98) es un repositorio de información, similar a un DW, pero orientado a un área o departamento específico de la organización (por ejemplo Compras, Ventas, RRHH, etc.), a diferencia del DW que cubre toda la organización, es decir la diferencia fundamental es su alcance.

Desde el punto de vista arquitectónico, la mayor diferencia entre los dos autores es el sentido de la construcción del DW, esto es comenzando por los Data marts o ascendente (Bottom-up, Kimball) o comenzando con todo el DW desde el principio, o descendente (Top-Down, Inmon).

Por otra parte, la metodología de Inmon se basa en conceptos bien conocidos del diseño de bases de datos relacionales (Inmon 02, Imhoff & Galemmo 03); la metodología para la construcción de un sistema de este tipo es la habitual para construir un sistema de información, utilizando las herramientas habituales, al contrario de la de Kimball, que se basa en un modelado dimensional (no normalizado) (Kimball et al 98, 08).

3. ¿Cuál metodología adoptar?

Pensamos que la metodología más acorde a los negocios de nuestra región es la de Kimball, por cuanto proporciona un enfoque de menor a mayor, muy versátil, y una serie de herramientas prácticas que

† En este artículo se han consultado las siguientes referencias técnicas para la metodología de Kimball: Mundy & Thornthwaite 2006, Kimball et al 1998, Kimball & Caserta 2004, Kimball & Ross 2002, Kimball & Merz 2000, Kimball & Ross 2010.

ayudan a la implementación de un DW. Es acorde a nuestras empresas porque se pueden implementar pequeños datamarts en áreas especificas de las mismas (compras, ventas, etc.), con pocos recursos y de poco irlos integrándolos en un gran almacén de datos. Por tanto, detallaremos esta metodología en lo que resta de este artículo.

4. La metodología de Kimball en detalle

La metodología se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle) (Kimball et al 98, 08, Mundy & Thornthwaite 06). Este ciclo de vida del proyecto de DW, está basado en cuatro principios básicos:

- Centrarse en el negocio: Hay que concentrarse en la identificación de los requerimientos del negocio y su valor asociado, y usar estos esfuerzos para desarrollar relaciones sólidas con el negocio, agudizando el análisis del mismo y la competencia consultiva de los implementadores.

- Construir una infraestructura de información adecuada: Diseñar una base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de requerimientos de negocio identificados en la empresa.

- Realizar entregas en incrementos significativos: crear el almacén de datos (DW) en incrementos entregables en plazos de 6 a 12 meses. Hay que usa el valor de negocio de cada elemento identificado para determinar el orden de aplicación de los incrementos. En esto la metodología se parece a las metodologías ágiles de construcción de software.

- Ofrecer la solución completa: proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios. Para comenzar, esto significa tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible. También se deberá entregar herramientas de consulta ad hoc, aplicaciones para informes y análisis avanzado, capacitación, soporte, sitio web y documentación.

La construcción de una solución de DW/BI (Datawarehouse/Business Intelligence) es sumamente compleja, y Kimball nos propone una metodología que nos ayuda a simplificar esa complejidad. Las tareas de esta metodología (ciclo de vida) se muestran en la figura 1.

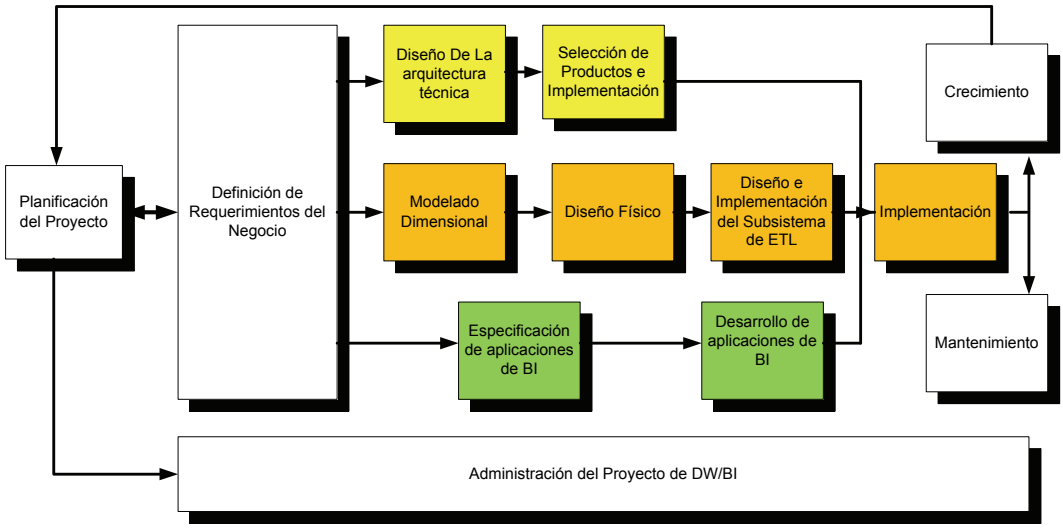


Fig. 1: Tareas de la metodología de Kimball, denominada *Business Dimensional Lifecycle* (Kimball et al 98, 08, Mundy & Thornthwaite 06)

De la figura 1, podemos observar dos cuestiones. Primero, hay que resaltar el rol central de la tarea de definición de requerimientos. Los requerimientos del negocio son el soporte inicial de las tareas subsiguientes. También tiene influencia en el plan de proyecto (nótese la doble fecha entre la caja de definición de requerimientos y la de planificación). En segundo lugar podemos ver tres rutas o caminos que se enfocan en tres diferentes áreas:

- *Tecnología (Camino Superior).* Implica tareas relacionadas con software específico, por ejemplo, Microsoft SQL Analysis Services.
- *Datos (Camino del medio).* En la misma diseñaremos e implementaremos el modelo dimensional, y desarrollaremos el subsistema de Extracción, Transformación y Carga (*Extract, Transformation, and Load - ETL*) para cargar el DW.
- *Aplicaciones de Inteligencia de Negocios (Camino Inferior).* En esta ruta se encuentran tareas en las que diseñamos y desarrollamos las aplicaciones de negocios para los usuarios finales.

Estas rutas se combinan cuando se instala finalmente el sistema. En la parte de debajo de la figura se muestra la actividad general de administración del proyecto. A continuación describiremos cada una de las tareas.

4.1. Planificación

En este proceso se determina el propósito del proyecto de DW/BI, sus objetivos específicos y el alcance del mismo, los principales riesgos y una aproximación inicial a las necesidades de información.

En la visión de programas y proyectos de Kimball, Proyecto, se refiere a una iteración simple del KLC (Kimball Life Cycle), desde el lanzamiento hasta el despliegue.

Esta tarea incluye las siguientes acciones típicas de un plan de proyecto:

- Definir el alcance (entender los requerimientos del negocio).
- Identificar las tareas
- Programar las tareas
- Planificar el uso de los recursos.
- Asignar la carga de trabajo a los recursos
- Elaboración de un documento final que representa un plan del proyecto.

Además en esta parte definimos cómo realizar la administración o gestión de esta subfase que es todo un proyecto en si mismo, con las siguientes actividades:

- Monitoreo del estado de los procesos y actividades.
- Rastreo de problemas
- Desarrollo de un plan de comunicación comprensiva que direccione la empresa y las áreas de TI

4.2. Análisis de requerimientos:

La definición de los requerimientos es en gran medida un proceso de entrevistar al personal de negocio y técnico, pero siempre conviene

tener un poco de preparación previa. Se debe aprender tanto como se pueda sobre el negocio, los competidores, la industria y los clientes del mismo. Hay que leer todos los informes posibles de la organización; rastrear los documentos de estrategia interna; entrevistar a los empleados, analizar lo que se dice en la prensa acerca de la organización, la competencia y la industria. Se deben conocer los términos y la terminología del negocio.

Parte del proceso de preparación es averiguar a quién se debe realmente entrevistar. Esto normalmente implica examinar cuidadosamente el organigrama de la organización. Hay básicamente cuatro grupos de personas con las que hablar desde el principio: el directivo responsable de tomar las decisiones estratégicas; los administradores intermedios y de negocio responsables de explorar alternativas estratégicas y aplicar decisiones; personal de sistemas, si existen, la gente que realmente sabe qué tipos de problemas informáticos y de datos existen; y por último, la gente que se necesita entrevistar por razones políticas.

A partir de las entrevistas, podemos identificar temas analíticos y procesos de negocio. Los temas analíticos agrupan requerimientos comunes en un tema común (ver tabla 1).

Tema Analítico	Análisis o requerimiento inferido o pedido	Proceso de negocio de soporte	Comentarios
Planificación de ventas	Análisis histórico de ordenes de revendedores	Ordenes de compras	Por cliente, por país, por región de ventas
	Proyección de ventas	Ordenes de compras	La proyección es un proceso de negocio que usa las órdenes como entradas

Tabla 1: Temas analíticos

Por otra parte, a partir del análisis se puede construir una herramienta de la metodología denominada matriz de procesos/dimensiones (Bus Matrix en inglés).

Una dimensión es una forma o vista o criterio por medio de cual se pueden sumariar, cruzar o cortar datos numéricos a analizar, datos que se denominan medidas (measures en inglés).

Esta matriz tiene en sus filas los procesos de negocio identificados, y en las columnas, las dimensiones identificadas.

Un ejemplo de esta matriz se puede observar en la tabla 2. Cada X en la intersección de las filas y columnas significa que en el proceso de negocio de la fila seleccionada se identifican las dimensiones propuestas.

	Dimensiones					
Proceso de Negocio	Tiempo	Producto	Empleados	Clientes (Revendedores)	Geografía de ventas	Importes
Proyección de ventas	X	X	X	X	X	X
Compras	X	X	X	X	X	X
Control de llamadas	X	X	X	X	X	
…						

Tabla 2: Matriz de procesos/dimensiones (*Bus Matrix*).

Finalmente se busca priorizar los requerimientos o procesos de negocios más críticos.

4.3. Modelado Dimensional

La creación de un modelo dimensional es un proceso dinámico y altamente iterativo. Un esquema general se puede ver en la figura 2.

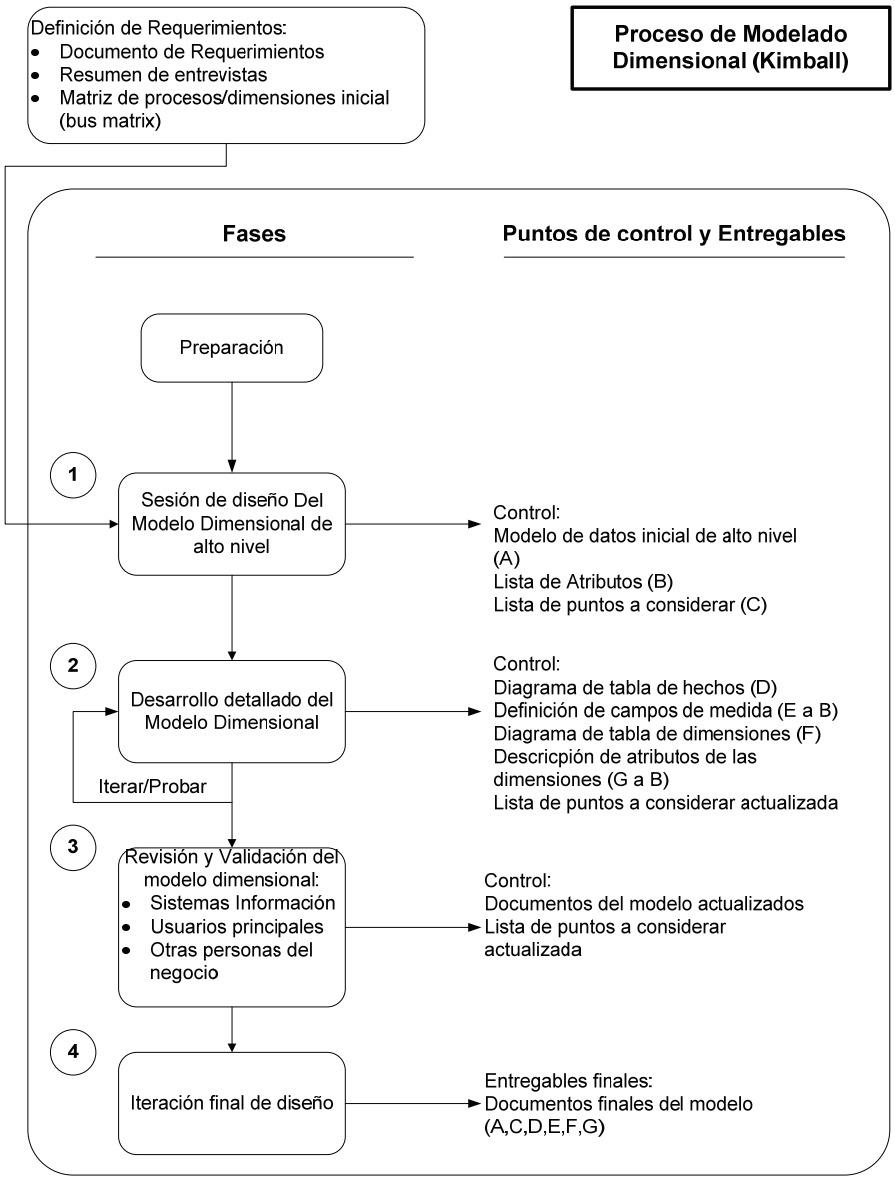


Fig. 2: Diagrama de flujo del proceso dimensional de Kimball (Mundy & Thornthwaite 06)

El proceso de diseño comienza con un modelo dimensional de alto nivel obtenido a partir de los procesos priorizados de la matriz descrita en el punto anterior.
El proceso iterativo consiste en cuatro pasos:

1. Elegir el proceso de negocio.
2. Establecer el nivel de granularidad.
3. Elegir las dimensiones.
4. Identificar medidas y las tablas de hechos.

4.3.1 Elegir el proceso de negocio

El primer paso es elegir el área a modelizar. Esta es una decisión de la dirección, y depende fundamentalmente del análisis de requerimientos y de los temas analíticos anotados en la etapa anterior.

4.3.2. Establecer el nivel de granularidad

La granularidad significa especificar el nivel de detalle. La elección de la granularidad depende de los requerimientos del negocio y lo que es posible a partir de los datos actuales. La sugerencia general es comenzar a diseñar el DW al mayor nivel de detalle posible, ya que se podría luego realizar agrupamientos al nivel deseado. En caso contrario no sería posible abrir (drill-down) las sumarizaciones en caso de que el nivel de detalle no lo permita.

4.3.3. Elegir las dimensiones

Las dimensiones surgen naturalmente de las discusiones del equipo, y facilitadas por la elección del nivel de granularidad y de la matriz de procesos/dimensiones. Las tablas de dimensiones tienen un conjunto de atributos (generalmente textuales) que brindan una perspectiva o forma de análisis sobre una medida en una tabla hechos. Una forma de identificar las tablas de dimensiones es que sus atributos son posibles candidatos para ser encabezado en los informes, tablas pivot, cubos, o cualquier forma de visualización, unidimensional o multidimensional.

4.3.4. Identificar las tablas de hechos y medidas

El último paso consiste en identificar las medidas que surgen de los procesos de negocios. Una medida es un atributo (campo) de una tabla que se desea analizar, sumarizando o agrupando sus datos, usando los criterios de corte conocidos como dimensiones. Las medidas habitualmente se vinculan con el nivel de granularidad del punto 4.3.2., y se encuentran en tablas que denominamos tablas de hechos (fact en inglés). Cada tabla de hechos tiene como atributos una o más medidas de un proceso organizacional, de acuerdo a los requerimientos. Un registro contiene una medida expresada en números, como ser cantidad, tiempo, dinero, etc., sobre la cual se desea realizar una operación de agregación (promedio, conteo, suma, etc.) en función de

una o más dimensiones. La granularidad es el nivel de detalle que posee cada registro de una tabla de hechos.

4.3.5. Modelo gráfico de alto nivel

Para concluir con el proceso dimensional inicial se realiza un gráfico denominado modelo dimensional de alto nivel (o gráfico de burbujas, *Bubble chart*, en el léxico de Kimball), como ilustra la figura 3.

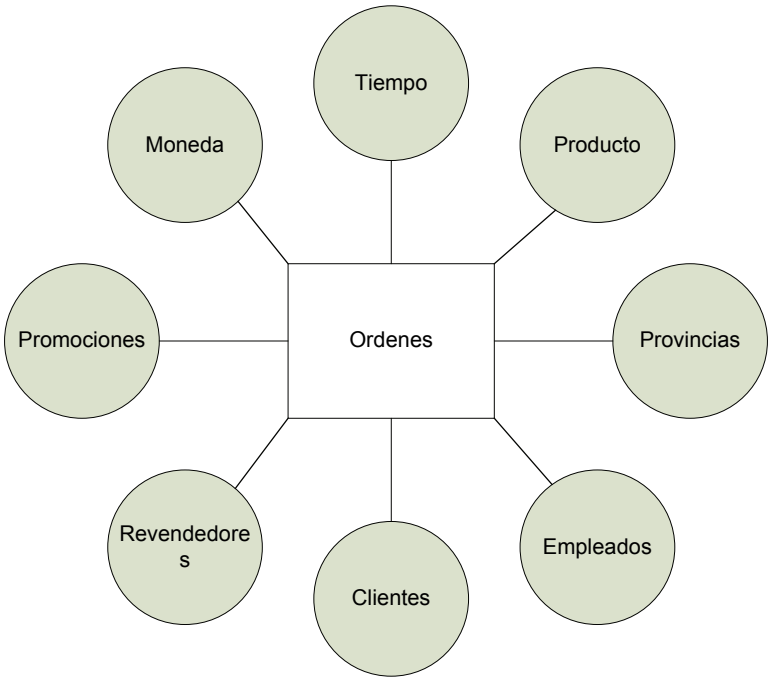


Fig. 3: Ejemplo de Modelo final de alto nivel de la sesión inicial de diseño (Mundy & Thornthwaite 06)

4.3.6. Identificación de atributos de dimensiones y tablas de hechos

La segunda parte de la sesión inicial de diseño consiste en completar cada tabla con una lista de atributos bien formada. Una lista de este tipo se muestra en la figura 4. Esta lista o grilla se forma colocando en las filas los atributos de la tabla, y en las columnas la siguiente información:

- Características relacionadas con la futura tabla dimensional del almacén de datos (*target*), por ejemplo tipo de datos, si es clave

primaria, valores de ejemplo, etc. Por razones de espacio no describiremos todas las columnas, para mayor información puede consultarse la referencia (Mundy & Thornthwaite 06).

- El origen de los datos (*source*, por lo general atributos de las tablas transaccionales).
- Reglas de conversión, transformación y carga (*ETL rules*), que nos dicen como transformar los datos de las tablas de origen a las del almacén de datos.

Table Name:	DimOrderInfo
Table Type	Dimension
View Name	OrderInfo
Description	OrderInfo is the "junk" dimension that includes miscellaneous information about the Order transaction
Used in schemas	Orders
Generate script?	Y

	Target										
Column Name	Description	Datatype	Size	Key?	FK To	NULL?	Default Value	Unknown Member	Example Values	SCD Type	Source System
Extended Property?	Y				Y				Y	Y	Y
OrderInfoKey	Surrogate primary key	smallint		PK ID		N		-1	1, 2, 3, 4...		ETL P rocess
BKSalesReasonID	Sales reason ID from source system	smallint				N		-1			AW
Channel	Sales channel	char	8					Unknown	Reseller, Internet	1	AW
SalesReason	Reason for the sale, as reported by the customer	varchar	30					Unknown		1	AW
SalesReasonType	Type of sales reason	char	10					Unknown	Marketing, Promotion, Other	1	AW
AuditKey	What process loaded this row?	int		FK	Audit Dim	N		-1		1	Derived

	Source						
SCD Type	Source System	Source Schema	Source Table	Source Field Name	Source Datatype	ETL Rules	Comments
Y	Y	Y	Y	Y	Y	Y	Y
	ETL Process					Standard surrogate key	
	AW	Sales	SalesReason	SalesReasonID	int	Convert to char; left-pad with zero. R for reseller row.	We need to insert a single row for Reseller
1	AW	Sales	SalesReason	Derived		Internet' for real sales reasons. 'Reseller' for reseller row.	
1	AW	Sales	SalesReason	Name	nvarchar(50)	Convert to varchar; "Reseller' for reseller row.	
1	AW	Sales	SalesReason	ReasonType	nvarchar(50)	Convert to varchar; "Reseller' for reseller row	
1	Derived					Populated by ETL system using standard technique	

Fig. 4: Lista de atributos (Mundy & Thornthwaite 06)

4.3.7. Implementar el modelo dimensional detallado

Este proceso consiste simplemente en completar la información incompleta de los pasos anteriores. El objetivo en general es identificar todos los atributos útiles y sus ubicaciones, definiciones y reglas de negocios asociadas que especifican cómo se cargan estos datos. Para este cometido se usa la misma planilla del punto anterior.

4.3.8. Prueba del modelo

Si el modelo ya esta estable, lo que se hace habitualmente es probarlo contra los requerimientos del negocio. Haciendo la pregunta práctica de ¿Cómo podemos obtener esta información en particular del modelo? Para las pruebas podemos usar diseños de reportes estructurados, de usuarios actuales, diseños de cubos prospectivos, etc.

4.3.9. Revisión y validación del modelo

Un vez que tenemos confianza plena en el modelo, ingresamos en esta etapa final (ver figura 2), lo cual implica revisar el modelo con diferentes audiencias, cada una con diferentes conocimientos técnicos y del negocio. En el área de sistemas deberían revisarlo los programadores y analistas de los sistemas, y el DBA si existe. También debería revisarse con usuarios y personas del negocio que tengan mucho conocimiento de los procesos y que quizás no hayan participado del diseño del modelo. Finalmente podemos hacer un documento que enuncie una serie de preguntas del negocio (tomadas a partir de los requerimientos), y las conteste por medio del modelo.

4.3.10 Documentos finales

El producto final, como se puede ver en la Figura 2, son una serie de documentos (solo mencionamos los más importantes), a saber:

- Modelo de datos inicial de alto nivel
- Lista de atributos
- Diagrama de tablas de hechos
- Definición de campos de medida
- Diagrama de tablas de dimensiones
- Descripción de los atributos de las dimensiones
- Matriz DW (o DW Bus Matrix) completa

4.4. Diseño Físico

En esta parte, intentamos contestar las siguientes preguntas:

- ¿Cómo puede determinar cuán grande será el sistema de DW/BI?
- ¿Cuáles son los factores de uso que llevarán a una configuración más grande y más compleja?
- ¿Cómo se debe configurar el sistema?

- ¿Cuánta memoria y servidores se necesitan? ¿Qué tipo de almacenamiento y procesadores?
- ¿Cómo instalar el software en los servidores de desarrollo, prueba y producción?
- ¿Qué necesitan instalar los diferentes miembros del equipo de DW/BI en sus estaciones de trabajo?
- ¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?
- ¿Cómo conseguir un plan de indexación inicial?
- ¿Debe usarse la partición en las tablas relacionales?

4.5. Diseño del sistema de Extracción, Transformación y Carga (ETL).

El sistema de Extracción, Transformación y Carga (ETL) es la base sobre la cual se alimenta el Datawarehouse. Si el sistema ETL se diseña adecuadamente, puede extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el DW en un formato acorde para la utilización por parte de las herramientas de análisis.

4.6 Especificación y desarrollo de aplicaciones de BI

Una parte fundamental de todo proyecto de DW/BI está en proporcionarles a una gran comunidad de usuarios una forma más estructurada y por lo tanto, más fácil, de acceder al almacén de datos. Proporcionamos este acceso estructurado a través de lo que llamamos aplicaciones de inteligencia de negocios (Business Intelligence Aplications).

Las aplicaciones de BI son la cara visible de la inteligencia de negocios: los informes y aplicaciones de análisis proporcionan información útil a los usuarios. Las aplicaciones de BI incluyen un amplio espectro de tipos de informes y herramientas de análisis, que van desde informes simples de formato fijo a sofisticadas aplicaciones analíticas que usan complejos algoritmos e información del dominio. Kimball divide a estas aplicaciones en dos categorías basadas en el nivel de sofisticación, y les llama informes estándar y aplicaciones analíticas.

4.6.1. Informes estándar

Los informes estándar son la base del espectro de aplicaciones de BI. Por lo general son informes relativamente simples, de formato predefinido, y parámetros de consulta fijos. En el caso más simple, son informes estáticos prealmacenados. Los informes estándar proporcionan a los usuarios un conjunto básico de información acerca de lo que está sucediendo en un área determinada de la empresa. Este tipo de aplicaciones son el caballo de batalla de la BI de la empresa. Son informes que los usuarios usan día a día. La mayor parte de lo que piden las personas durante el proceso de definición de requisitos se clasificaría como informes estándar. Por eso es conveniente desarrollar un conjunto de informes estándar en el ciclo de vida del proyecto. Algunos informes estándares típicos podrían ser:

- Ventas del año actual frente a previsión de ventas por vendedor
- Tasa de renovación mensual por plan de servicio
- Tasa quinquenal de deserción por unidad académica
- Tasas de respuestas de correo electrónico por promoción por producto (marketing)
- Recuento de audiencia y porcentaje de la audiencia total por la red de televisión por día de la semana y hora del día (Sistema de marketing televisivo)
- Reclamos del año actual hasta la fecha frente a previsión, por tipo de vehículo
- Volumen de llamadas por producto como un porcentaje del total de ventas

4.6.2. Aplicaciones analíticas

Las aplicaciones analíticas son más complejas que los informes estándar. Normalmente se centran en un proceso de negocio específico y resumen cierta experiencia acerca de cómo analizar e interpretar ese

proceso de negocio. Estas aplicaciones pueden ser muy avanzadas e incluir algoritmos y modelos de minería de datos, que ayudan a identificar oportunidades o cuestiones subyacentes en los datos. Otra característica avanzada en algunas aplicaciones analíticas es que el usuario puede pedir cambios en los sistemas transaccionales basándose en los conocimientos obtenidos del uso de la aplicación de BI. En el otro extremo del espectro, algunas aplicaciones analíticas se venden como soluciones cerradas o enlatados, y son independientes de las aplicaciones particulares de la empresa. Algunas aplicaciones analíticas comunes incluyen:

- Análisis de la eficacia de la promociones
- Análisis de rutas de acceso en un sitio Web
- Análisis de afinidad de programas
- Planificación del espacio en espacios comerciales
- Detección de fraudes
- Administración y manejo de categorías de productos

5. Conclusiones

La metodología de Kimball proporciona una base empírica y metodológica adecuada para las implementaciones de almacenes de datos pequeños y medianos, dada su gran versatilidad y su enfoque ascendente, que permite construir los almacenes en forma escalonada. Además presenta una serie de herramientas, tales como planillas, gráficos y documentos, que proporcionan una gran ayuda para iniciarse en el ámbito de la construcción de un Datawarehouse.

Referencias

Imhoff & Galemmo, *Mastering Data Warehouse Design: Relational and Dimensional Techniques*, Wiley Publishing, 2003

Inmon, *Building the Data Warehouse*, (Third Edition). John Wiley & Sons, 2002

Kimball & Caserta, *The Data Warehouse ETL Toolkit*, Indianapolis, Wiley, 2004.

Kimball & Merz, *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*, Wiley, 2000.

Kimball & Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (Second Edition), New York, Wiley, 2002.

Kimball & Ross, *The Kimball Group Reader; Relentlessly Practical Tools for Data Warehousing and Business Intelligence*, Indianapolis, Wiley, 2010.

Kimball et al., *The Data Warehouse Lifecycle Toolkit*. 2nd Edition. New York, Wiley, 2008

Kimball et al., *The Data Warehouse Lifecycle Toolkit*. New York, Wiley, 1998.

Mundy & Thornthwaite, *The Microsoft Data Warehouse Toolkit—With SQL Server 2005 and the Microsoft Business Intelligence Toolset, Indianapolis*, Wiley, 2006.

Páginas web útiles

http://www.kimballgroup.com : Este sitio contiene mucha información y artículos sobre la metodología, y además una serie de planillas de Excel usadas en cada paso de la metodología.
http://www.bi-bestpractices.com/view-articles/4768
http://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimensional/
http://kle.sisorg.com.mx/articulo04.html

Estimación de la peligrosidad sísmica que afecta a la ciudad de Salta

Lía Orosco y Mika Haarala-Orosco*

liaorosco@gmail.com

Resumen

En este artículo se estima la amenaza sísmica sobre la Ciudad de Salta, haciendo uso de una metodología probabilista. Se considera el catálogo instrumental del período 1964-2008 del Centro Sismológico Internacional (ISC) cuyas magnitudes fueron homogeneizadas a magnitud momento (Mw).

Se definen cinco zonas sismogénicas de las que cuatro corresponden a sismos superficiales y una a sismos de mediana profundidad. Se determinaron las curvas de excedencia de magnitudes para cuatro sismos de diseño y con varias relaciones de atenuación calculados para otros ambientes sísmicos similares al que presenta la Provincia de Salta, se calcularon las aceleraciones en roca en la ciudad, para cada zona sismogénica considerada.

Las fuentes sismogénicas ubicadas en el frente de deformación de Sierras Subandinas y en el Valle de Lerma y sus alrededores, son las determinantes para definir la peligrosidad sísmica de la Ciudad de Salta.

Palabras Claves: riesgo sísmico, acelerogramas, potencial de daño

---

* Lía Orosco es Doctora Ingenieria de Caminos, Canales y Puertos por la Universidad Politécnica de Cataluña en el área de la Ingeniería Sísmica. Es Profesora Titular de la Cátedra Construcciones de Hormigón Armado, en la Facultad de Ingeniería e Informática de la UCASAL. Mika Haarala Orosco, es Msc en Matemáticas, por la Universidad de Joensuu, Finlandia. Es investigador del Instituto de Estudios Intedisciplinarios de Ingeniería, de la UCASAL, en el área de Matemática Aplicada y Estadística.

1. Introducción

En su acepción más simple y básica, el riesgo sísmico de una ciudad es el resultado de dos elementos o conceptos básicos: la peligrosidad o amenaza y la vulnerabilidad de la misma. La peligrosidad engloba el fenómeno natural en sí mismo, y aquellos aspectos que hacen que este peligro se vea amplificado o reducido; entre los más importantes se tienen los "efectos de sitio" y entre estos se cita como ejemplo, la amplificación de la respuesta debido a suelos blandos.

La vulnerabilidad también comprende varios aspectos (Carreño et al., 2007). La vulnerabilidad física, relacionada al estado de las obras que el hombre construye (edificios, puentes, torres de electricidad, acueductos, etc.) es la más conocida y evidente. El ser humano puede en mayor o menor medida, según las circunstancias, manejar la vulnerabilidad, y por ende, reducir o aumentar el riesgo sísmico del sitio donde vive y se desarrolla.

Toda acción encaminada a reducir el riesgo sísmico implica básicamente reducir la vulnerabilidad. Con respecto a la peligrosidad, si bien no es posible actuar sobre los mecanismos que producen los sismos, si es necesario conocer el fenómeno lo máximo posible. De esta manera, se contribuye al objetivo de reducir el riesgo, ya que este conocimiento es básico para hacer efectivas medidas apropiadas de reducción de vulnerabilidad que entre otras comprenden la definición de la acción sísmica con fines de diseño sismorresistente, tanto para obras nuevas como para reforzar obras existentes, pensar escenarios de catástrofes para diseñar políticas de gestión de la emergencia, trazar planes de prevención, etc.

Este trabajo es un aporte al conocimiento de la peligrosidad sísmica que amenaza la ciudad de Salta. Esta amenaza se define en términos probabilistas o deterministas, o bien en base a una combinación de ambas filosofías según el caso.

En regiones como el noroeste argentino, se está lejos aún de poder definir niveles de amenazas en base a criterios puramente deterministas, ya que se debería tener un conocimiento profundo y

exacto de las estructuras que provocan los sismos que afectan la ciudad, y no se posee esta información. Por ello, el criterio probabilista se impone, ya que se tiene alguna información dada sobre todo por la red sísmica mundial y en los últimos tiempos por la red nacional, a cargo del Instituto Nacional de Prevención Sísmica (INPRES).

La peligrosidad sísmica será definida en términos de **máxima aceleración en "roca"** en el sitio, ya que la norma sismorresistente en vigencia en el país, define la acción sísmica en términos de espectros elásticos de pseudoaceleración.

En base a los principios sentados por Cornell (1.968), la peligrosidad o amenaza sísmica puede expresarse en forma probabilista como:

$$\upsilon(a) = \sum_{i=1}^{N_s} \upsilon_i \times \iint \left[\mathrm{P}(a > a_0 / m, r \right] f(m)\, f(r)\, \mathrm{d}m\, \mathrm{d}r \qquad (1)$$

donde $\upsilon(a)$ es la probabilidad que un valor de aceleración en roca exceda un cierto valor umbral a_0; υ_i representa la tasa de ocurrencia de sismos con magnitudes más elevadas que un cierto valor límite m_0 en la zona sismogénica i (N_s es el número de zonas que afectan el sitio); $f(m)$ y $f(r)$ son funciones de densidad de probabilidad independientes de la magnitud y la distancia, mientras que $p(a > a_0 / m, r)$ es la probabilidad condicional y está relacionada a leyes de atenuación regional, de tal forma que es uno menos la función de distribución acumulada para la atenuación de a.

Cada componente de la fórmula anterior es considerada en este trabajo, ya sea en forma probabilista o determinista, por lo que el modelo utilizado podría definirse como "híbrido".

Por lo tanto, las principales consideraciones a realizarse para el

logro del objetivo de este trabajo son:

a) Definición de escenarios sísmicos donde las zonas potenciales que pueden afectar al sitio bajo estudio son identificadas y analizadas.

b) Definición de un marco probabilista para determinar la tasa de excedencia de la magnitud en cada una de la zonas sismogénicas, dado un cierto período de tiempo.

c) Cálculo de la aceleración máxima en roca en la Ciudad de Salta, obtenidas aplicando leyes de atenuación, definidas en función de la magnitud y distancias para valores picos de la aceleración en roca. Se elige este parámetro pues la acción sísmica está definida en el Reglamento Sismorresistente en vigencia por medio de espectros (en realidad pseudoespectros elásticos) de aceleración Para la región bajo estudio no hay definidas leyes de atenuación propias, debida a la falta de registros de sismos fuertes; por ello, se han considerado varias expresiones propuestas en la bibliografía a fin de reducir la incertidumbre por la falta de tales leyes locales.

2. Definición de las fuentes sismogénicas que afectan el sitio

Se tomó como área de estudio un radio de 200 km alrededor del sitio. Para el análisis probabilista se considera un catálogo derivado del ISC Comprehensive Bulletin (2.001) para el periodo 1.964 - Julio 2.008 del Catálogo ISC (International Seismic Center); 2008 es el último año que figura en este catálogo, ya que existe un período en que se hacen los cálculos para definir con menos incertidumbre los parámetros de cada evento. La Figura 1 muestra el sitio de estudio (con una cruz), la zona de influencia considerada (radio de 200 km) y los eventos en profundidad. Los colores señalan el rango de magnitud de los mismos.

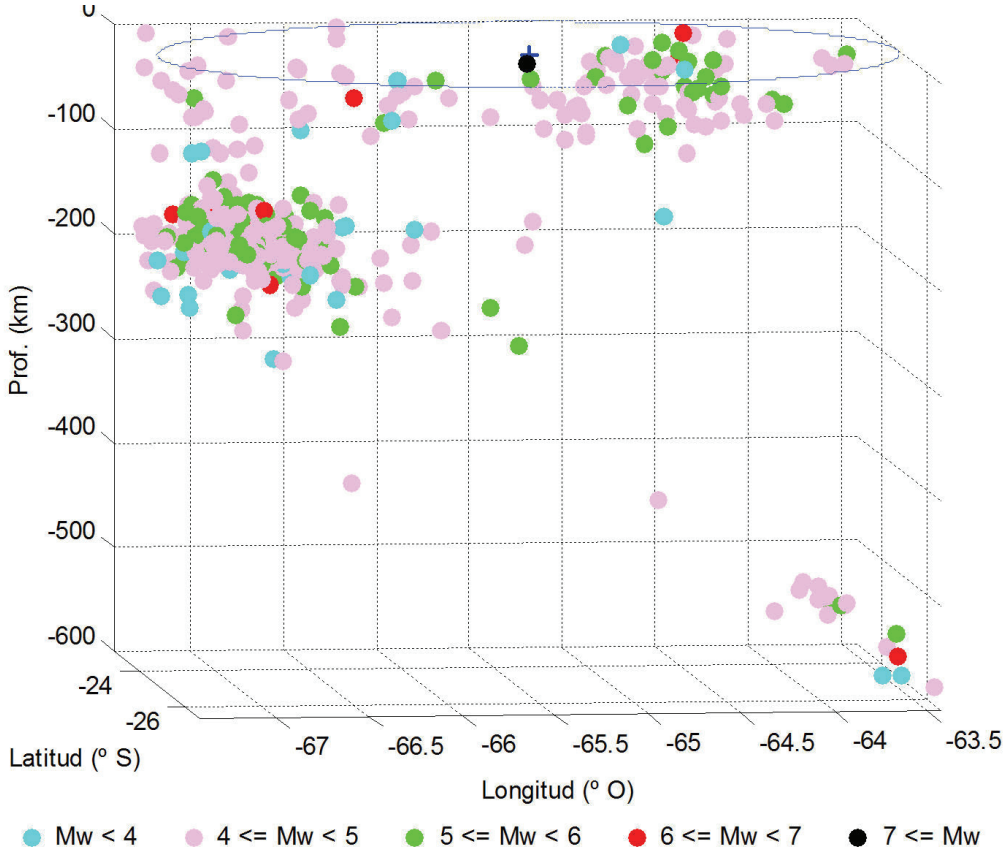


Figura 1: Ubicación de focos en profundidad de eventos ubicados en un radio de 200 km de la ciudad de Salta

Se hicieron intentos de homogeneizar las distintas clases de magnitudes que incluyen los distintos catálogos, con expresiones disponibles en la literatura para la zona (Sudamérica) que son lineales; pero ellas no son adecuadas como lo demuestra la Figura 2. Esta representa los eventos tomados en una región de latitud [+5º -55º] y longitud [-60º -90º] en dos tipos de magnitud: en el eje de las abscisas se tienen los valores en magnitud momento y en la ordenada, se muestran los valores en magnitud de onda de cuerpo (período corto). Un análisis de regresión no lineal, da como resultado una curva que tiene la siguiente expresión

$$m_b = \mathrm{e}^{\left(\frac{-15{,}70}{\mathrm{M_w}} + 5{,}58\right)^{0{,}5}} \qquad (2)$$

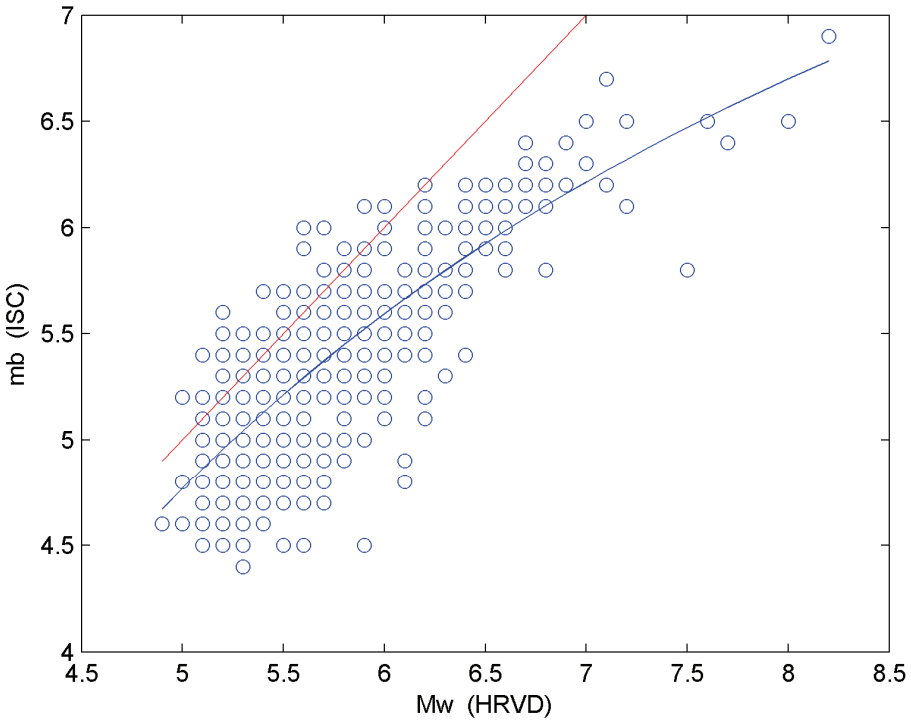


Figura 2: Relación entre M_w y m_b. La gráfica resume datos del catálogo ISC y el catálogo de la Universidad de Harvard (HRVD), tal como fueron publicados en el boletín ISC. La curva en azul, muestra la regresión no lineal.

Otra operación que se realiza sobre el catálogo es el filtrado de réplicas ya que debe considerarse sólo el movimiento principal que define el evento sísmico. En este caso, se tomó el criterio de Davis y Frolich (1991)

$$d_{st} = (d^2 + C^2T^2)^{1/2} \qquad (3)$$

donde d_{st} está en km, d es la distancia entre dos eventos (km), T el tiempo entre ellos (días) y C una constante que relaciona distancias y tiempo y que se sugiere sea igual a 1km/día para sismos en Sudamérica. Se tomó como corte para definir réplicas, un valor d_{st}=75 km.

Se hace una clasificación de la sismicidad según la profundidad de los focos, de tal manera de identificarlos como superficiales (menor a 70 km de profundidad focal); sismicidad media, cuando estos focos se ubican en el intervalo 70 km a 350 km y finalmente los sismos profundos cuando se originan a más de 350 km de profundidad.

En la Figura 3 se aprecia la sismicidad superficial; en la Figura 4 se muestra la sismicidad de profundidad media; en principio, son las que serán consideradas en el análisis.

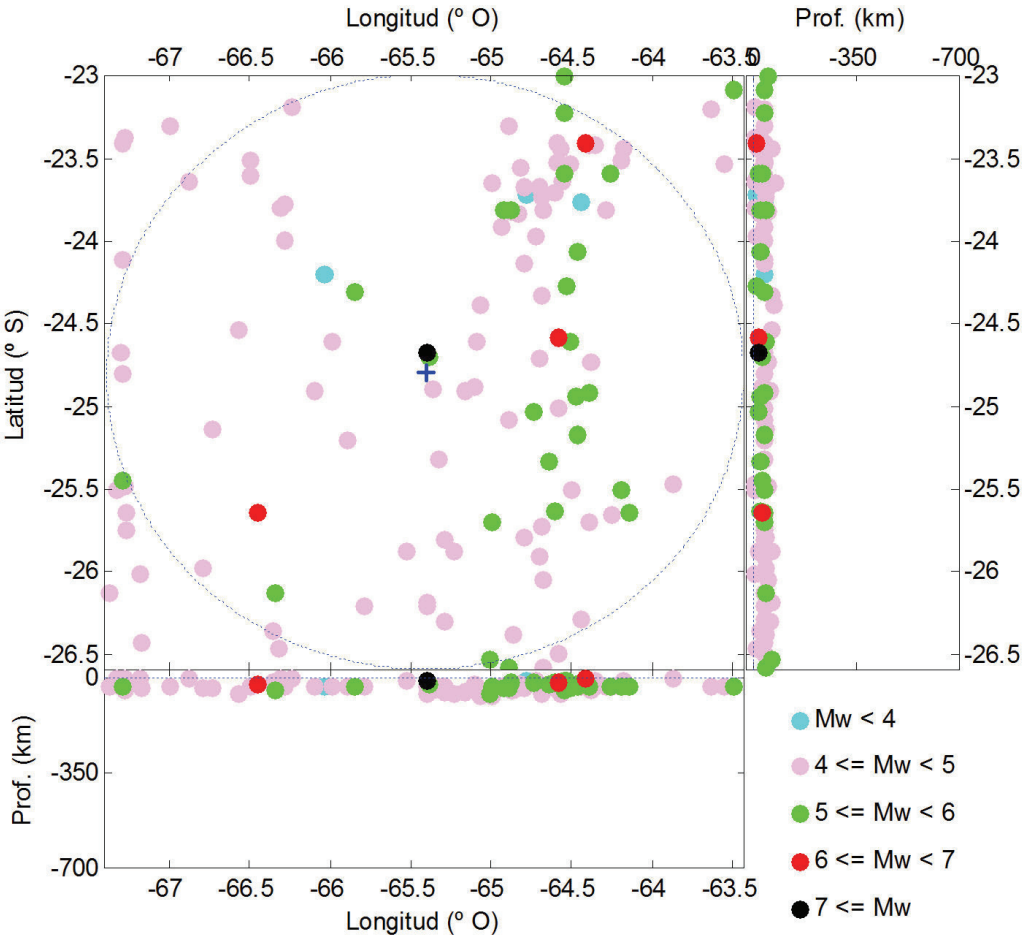


Figura 3: Sismicidad superficial que afecta la Ciudad de Salta

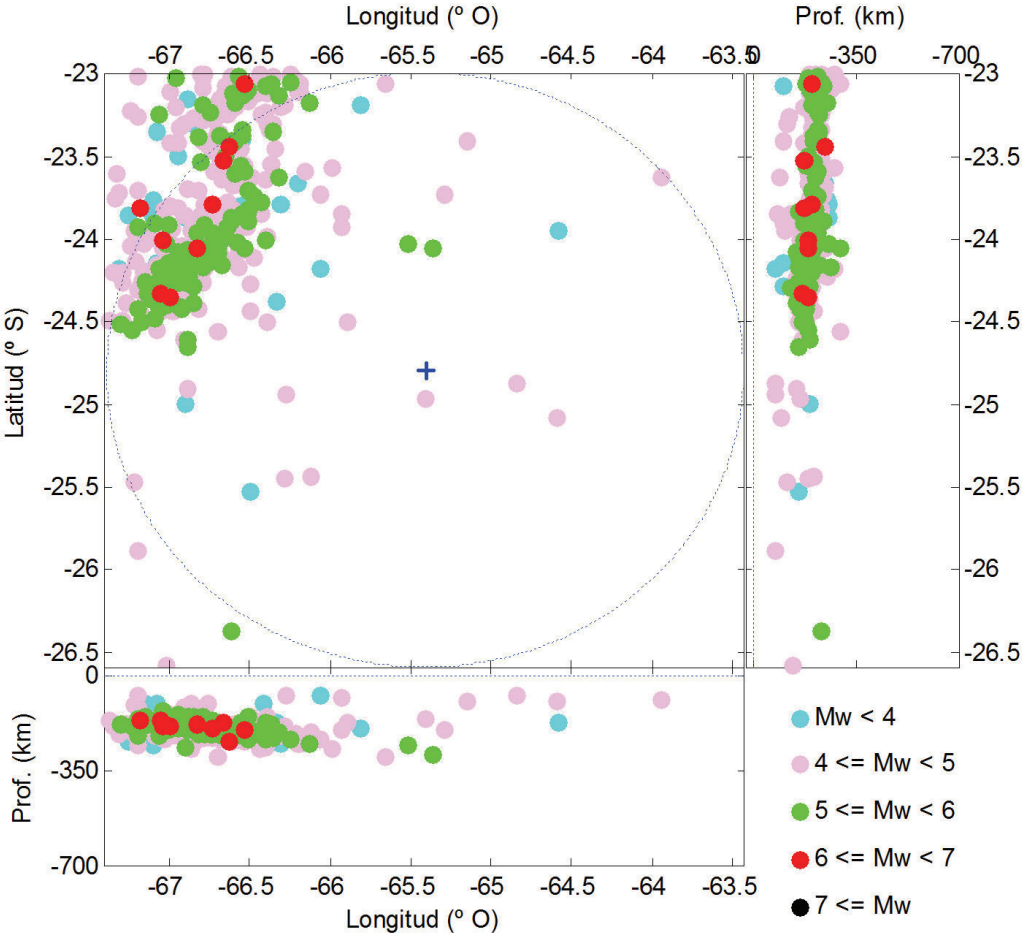


Figura 4: Sismicidad media que afecta la Ciudad de Salta

Se han definido cinco zonas sismogénicas que afectan a la Ciudad de Salta, considerando un sitio céntrico de la misma, de coordenadas 65,41ºO , 24,79ºS, señalado con una cruz azul en las figuras. Las zonas 1 a 4 comprenden sismos superficiales. Se definen las zonas en base a la distribución de eventos que resultan de la proyección en superficie de los focos. En cada una de ellas, se considera que la sismicidad es homogénea. La zona 1 (Figura 5) engloba los sismos que se alinean al este de la ciudad y comprenden la franja donde se han originado los sismos históricos más intensos. Corresponden al frente de

deformación este que provoca el fenómeno de subducción de la Placa de Nazca, debajo de la Sudamericana. En general se origina un ambiente con predominio de tensiones de compresión, lo que origina movimientos de fallas inversas y superficiales.

La zona 2 representa la sismicidad superficial difusa situada en el mismo Valle de Lerma. La zona 3 representa los sismos originados al oeste de la Ciudad de Salta, mientras se consideró una cuarta zona más alejada situada al sur. La zona 5 (Figura 6) engloba los numerosos sismos de subducción situados a más de 70 km de profundidad.

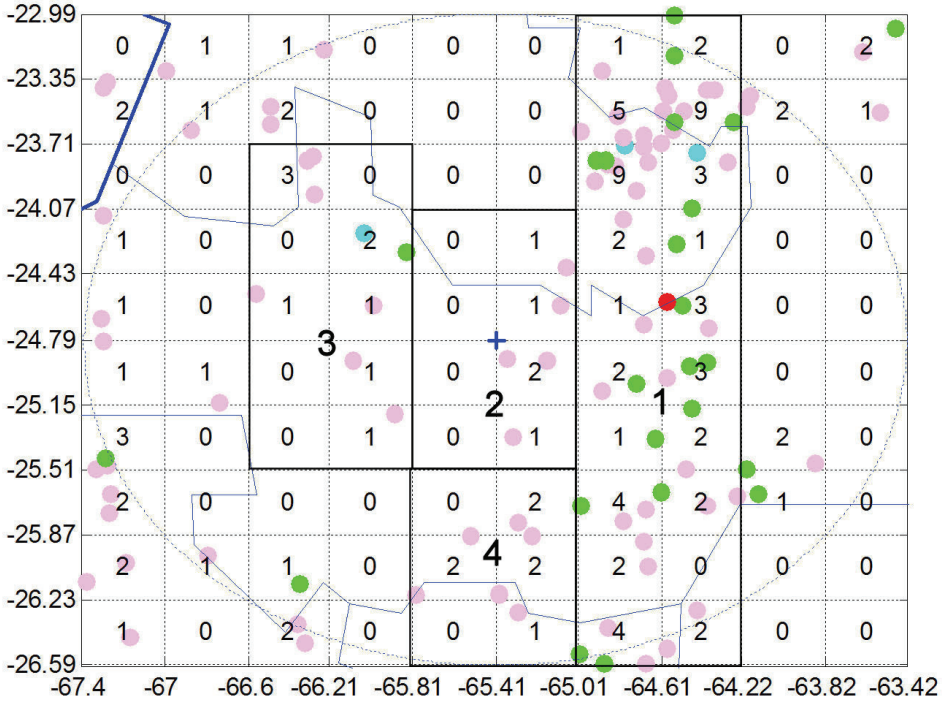


Figura 5: Zonas sismogénicas asociadas a la sismicidad superficial

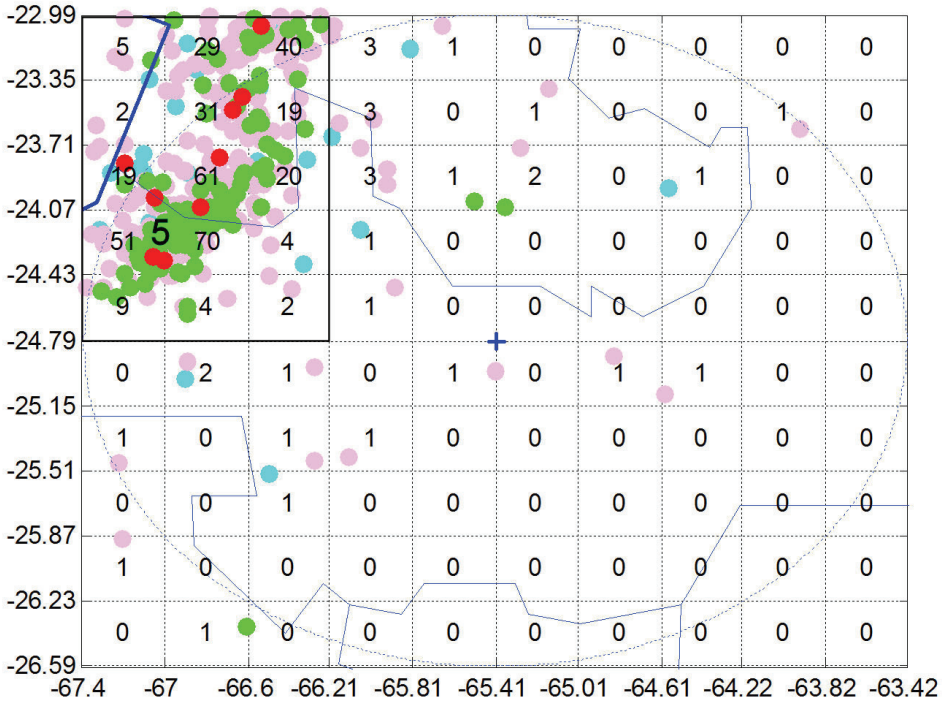


Figura 6: Zona sismogénica asociada a la región de subducción de la placa de Nazca – sismicidad de profundidad media.

3. Ley Gutemberg-Richter

La bien conocida ley de Gutemberg-Richter (1.944) es utilizada para representar la tasa de ocurrencia de eventos con una magnitud igual o mayor que cierto valor

$$\log N = a - b\,M \qquad (4)$$

donde N es el número de eventos cuyas magnitudes son mayores o iguales que M y las constantes a y b están relacionadas a la sismicidad de la región. El parámetro b describe la relación entre sismos intensos y pequeños que ocurren en la región. Así, cuanto más grande sea este valor, indicará que hay una mayor proporción de sismos pequeños que

grandes en la zona; por el contrario valores bajos denotan una elevada actividad sísmica.

Como función de densidad de probabilidad ($f(m)$), frecuentemente se adopta un modelo exponencial truncado

$$f(m) = \frac{\beta \, \mathrm{e}^{-\beta(m-m_0)}}{1-\mathrm{e}^{-\beta(m_u-m_0)}} \qquad (5)$$

donde $\beta = \log b$. Debe adoptarse para las zonas sismogénicas consideradas un valor límite inferior y superior para la magnitud.

4. Terremotos mínimo y máximo a considerar

Con respecto a la mínima magnitud a considerar en los cálculos probabilistas, fue definida en 4,2.

Los terremotos máximo posibles se define en base a las características de la zona sismogénica considerada. Para definir la amenaza sísmica de la ciudad de Salta se deberían conocer estos parámetros de las fallas. Una falla queda perfectamente determinada cuando se estima el terremoto máximo probable (TMP) que pueda ocurrir en la falla en base a las características geométricas y sismotectónicas de la misma; el período de recurrencia para tal evento y la ley de recurrencia de la misma.

No se cuenta con estudios que puedan permitir el cálculo con certeza del período de retorno del TMP para fallas en esta zona, sobre todo lo relacionado a la tasa de movimiento de las mismas. Es posible inferir valores en base a fórmulas empíricas obtenidas haciendo uso de datos de varios sitios del mundo de elevada actividad sísmica. Estas fórmulas relacionan parámetros de falla tal como superficie o longitud de ruptura con la magnitud alcanzada. Por ello, se considera que estas expresiones pueden señalar a un límite superior en el cálculo del TMP,

en la hipótesis de una ruptura total

Con este recaudo, utilizando expresiones para fallas inversas superficiales (predominantes en la determinación de la peligrosidad de esta zona según los informes geológicos) como la propuesta por Idda en 1.959 (Aguiar, 2.003),

$$M_{\max} = 6,04 + 0,79 \log L_r \qquad (6)$$

o la definida por Bonilla et al. (1.984)

$$M_{\max} = 5,71 + 0,916 \log L_r \qquad (7)$$

dan valores para una longitud de 100 km de **7,62 y 7,54** respectivamente, para fallas ubicadas en la zona sismogénica 1. Esta longitud es la de la falla donde se alinean los sismos históricos de 1.692 y 1.908.

Slemmons (1.982) propuso una fórmula considerando la longitud total de falla en base a datos del oeste norteamericano:

$$M_s = 6,618 + 0,0012 L_t \qquad (8)$$

con una desviación estándar de 0,221. Para el sistema de fallas antes mencionado, se tiene una magnitud máxima de **6,78**, menor a la anotada en el catálogo histórico (estimada en 7 por el INPRES).

De todos modos, estos valores deben ser considerados quizás a título de interés, pues no se tienen datos exactos acerca de la cinética de estas fallas, datación de deformación “on going”, o estimación de

desplazamientos sufridos en eventos pasados. Además, las longitudes consideradas en los cálculos, corresponden a su traza tal como aparecen en los mapas geológicos.

El terremoto máximo en términos probabilistas se estima considerarando un período de tiempo extenso, como 10.000 años, que es un límite inferior para considerar que una falla sea activa o no (Serva,1992). Estudios más recientes elevan este período a 15.000 años o más aún, dependiendo de la obra a emplazarse en el sitio considerado, o los fines del estudio.

Considerando el corto período de tiempo que comprende el catálogo sísmico, la previsiones a tan largo plazo llevan implícito una gran incertidumbre, por lo que los valores obtenidos no son fiables.

5. Tasa de ocurrencia de terremotos

Dos filosofías se usan generalmente para determinar este parámetro; una sostiene que suceden según un proceso de Poisson, por lo que cada uno de ellos es independiente con respecto a cualquier otro. Otra corriente considera que cuando sucede un evento, libera las tensiones en la falla por lo que reduce temporalmente la probabilidad de ocurrencia de sismos (las réplicas no se encuadran en este concepto). En este trabajo se adoptó un modelo tipo Poisson (Stewart et al., 2.001).

Un proceso tipo Poisson se define como:

$$\mathrm{P_t}(N=n)=\frac{\mathrm{e}^{-\lambda t}(\lambda t)^n}{n!} \qquad n=1,2,3,\ldots \quad \lambda>0 \qquad (9)$$

donde N (la variable aleatoria) es el número de veces que sucede un evento en el intervalo de tiempo $[t_0,t_1]$ (cuya duración en t unidades de tiempo) y λ es el parámetro de Poisson y es el valor medio de

ocurrencias del evento por unidad de tiempo. Este parámetro describe el número de eventos sísmicos que exceden un valor umbral.

6. Distancia fuente-sitio

Se considera este parámetro en forma determinista y se lo define como el punto más cercano según el catálogo. Por ello, al definirse las zonas simogenéticas, la distancia al sitio es la que corresponde al evento más cercano al punto bajo estudio.

7. Peligrosidad sísmica

La amenaza sísmica en el sitio fue estimada considerando la probabilidad de excedencia para la magnitud en cada zona sismogénica considerada.

La probabilidad de excedencia del número de eventos cuya magnitud es mayor que un cierto valor de referencia, para diferentes intervalos de tiempo, se determina como:

$$\mathrm{P_t}(M > m) = 1 - \mathrm{e}^{\lambda t} \qquad (10)$$

El parámetro λ se considera como el número de eventos con magnitudes mayores que un cierto valor y es estimado utilizando la relación de Gutemberg-Richter.

8. Resultados

En las Figuras 7, 8 y 9 pueden apreciarse los eventos considerados en cada una de las zonas. Las figuras ubicadas a la derecha muestran las curvas representativas de la ley Gutemberg-Richter, con su correspondiente intervalo de confidencia (1%). En ellos,

los cuadros a la izquierda muestran loe eventos considerados en el análisis y los cuadros a la derecha, los resultados de la regresión.

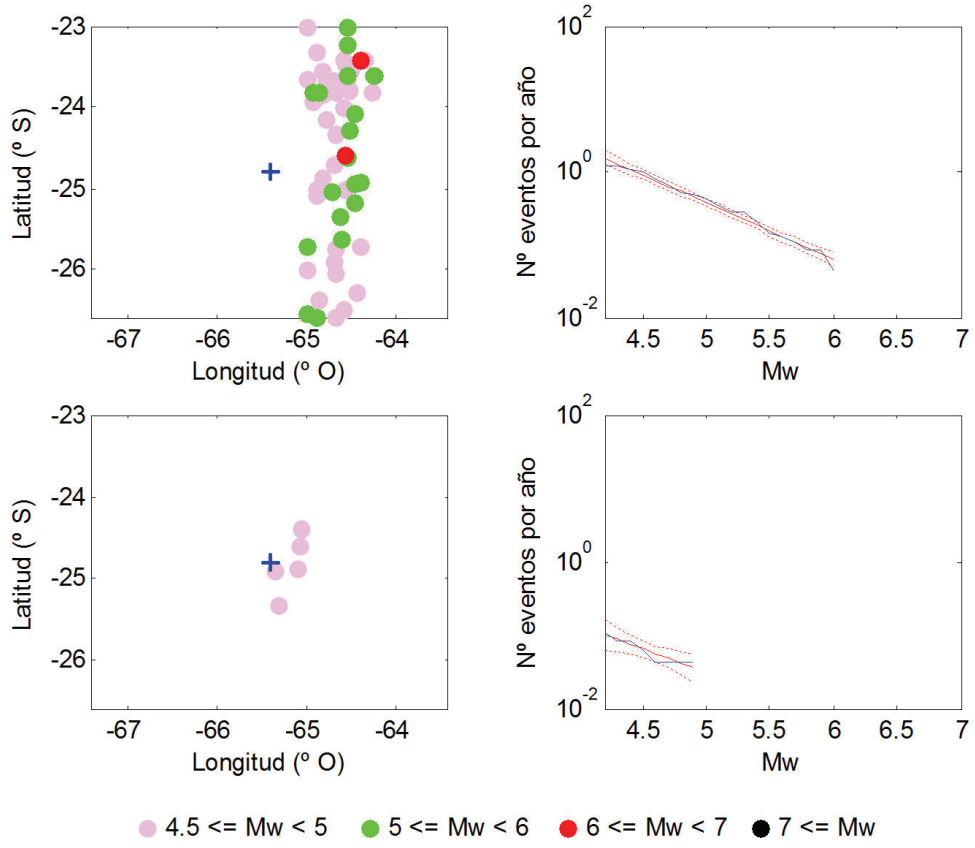


Figura 7: Resultados de la regresión para las zonas sismogénicas 1 (parte superior) y 2 (parte inferior).

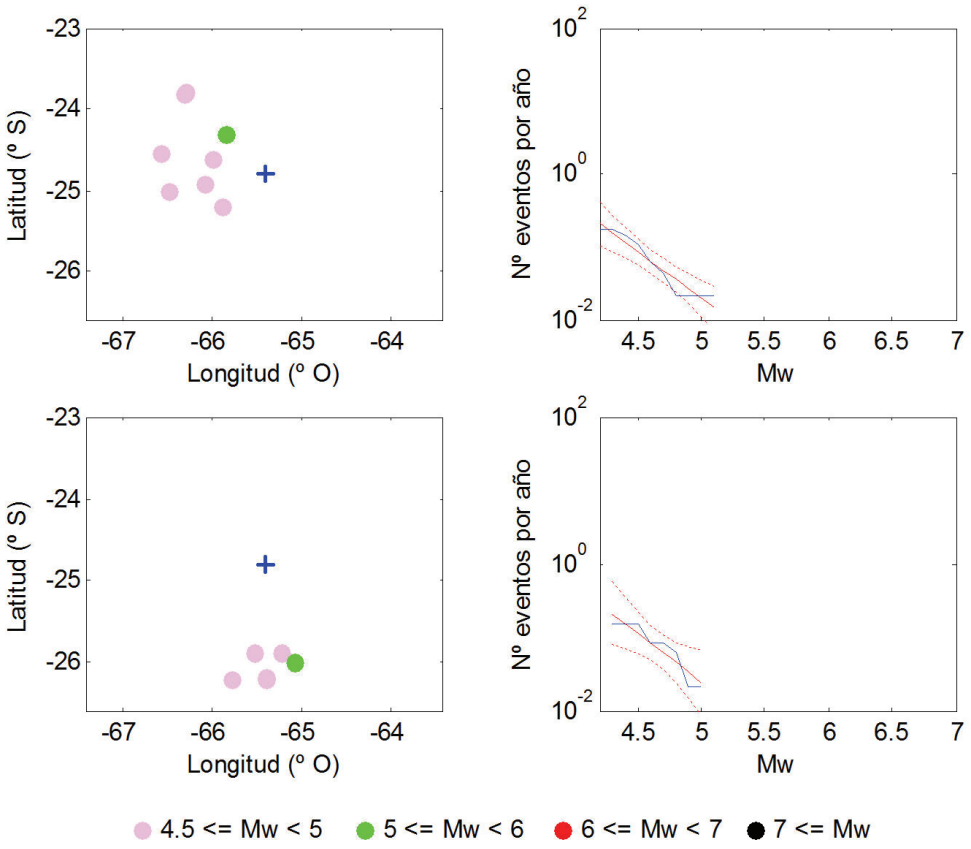


Figura 8: Resultados de la regresión para las zonas sismogénicas 3 y 4.

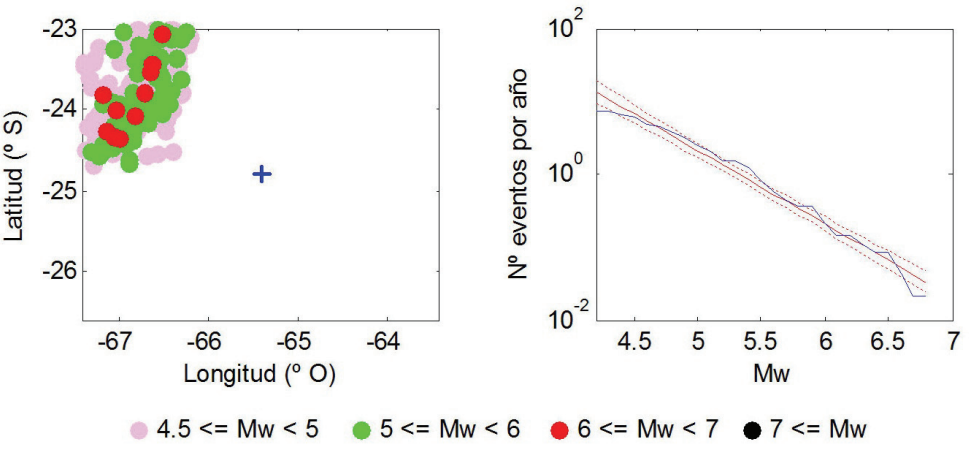


Figura 9: Resultados de la regresión para la zona sismogénica 5

La Tabla 1 muestra los parámetros de la ley de Gutemberg-Richter (a y b), obtenidos por regresión lineal.

Tabla 1: Resultados de la regresión (Desviación estándar entre paréntesis)

Zona sismogénica	**A**	**B**
1	3,44 (0,13)	0,77 (0,03)
2	1,68 (0,46)	0,64 (0,10)
3	4,64 (0,62)	1,27 (0,14)
4	5,08 (1.02)	1,34 (0,22)
5	5,34 (0.17)	1,00 (0,03)

Las Figuras 10 a 14 muestran las curvas de excedencia para las magnitudes esperadas considerando distintos períodos de tiempo (por ejemplo, la vida útil de las estructuras de obras civiles, tiempo de exposición para definir niveles de amenaza con fines de elaboración de planes de contingencia, prevención y mitigación de desastres, etc.).

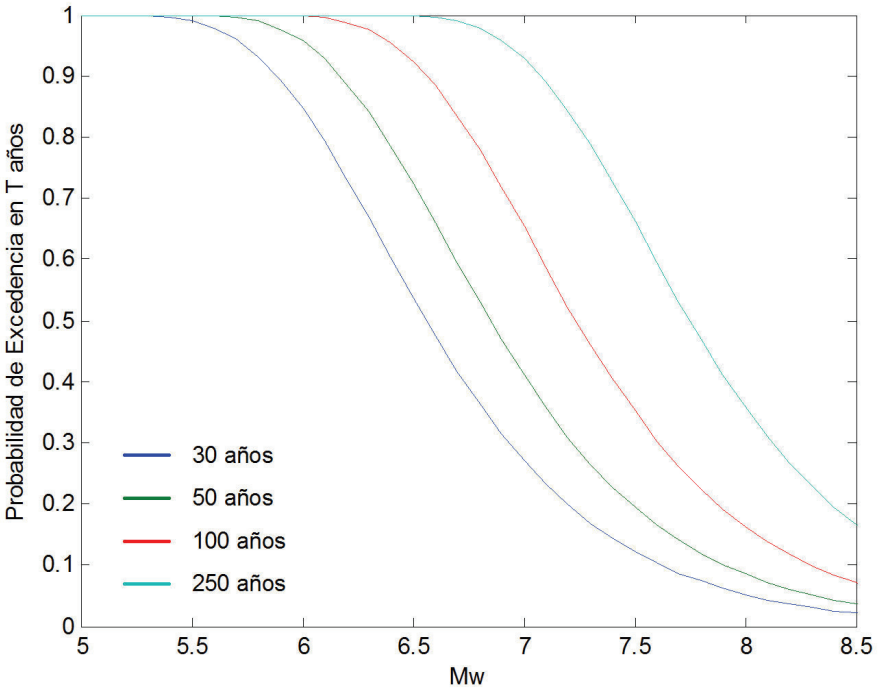


Figura 10: Curvas de excedencia de magnitudes para eventos originados en la zona sismogénica 1

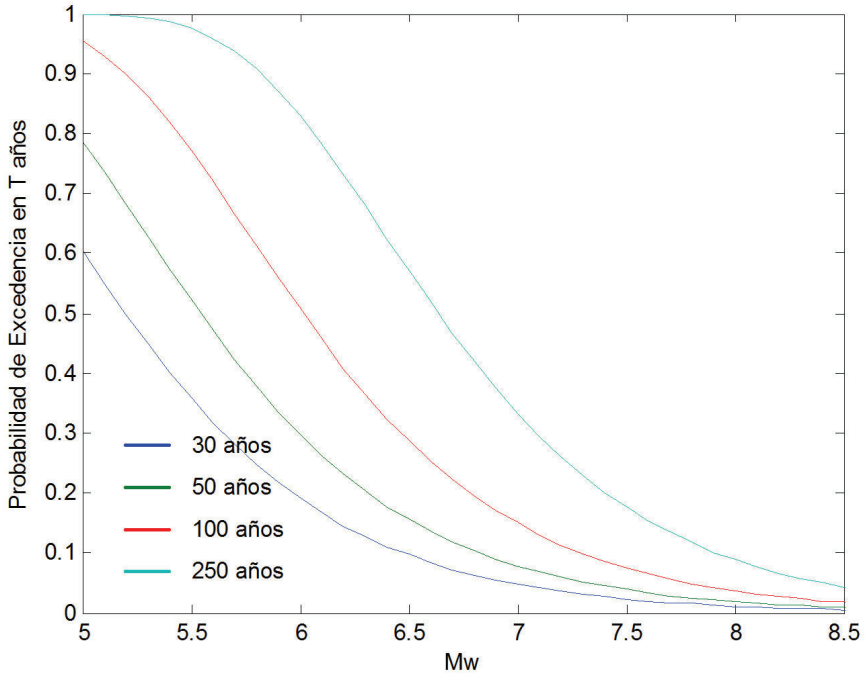


Figura 11: Curvas de excedencia de magnitudes de eventos originados en la zona sismogénica 2

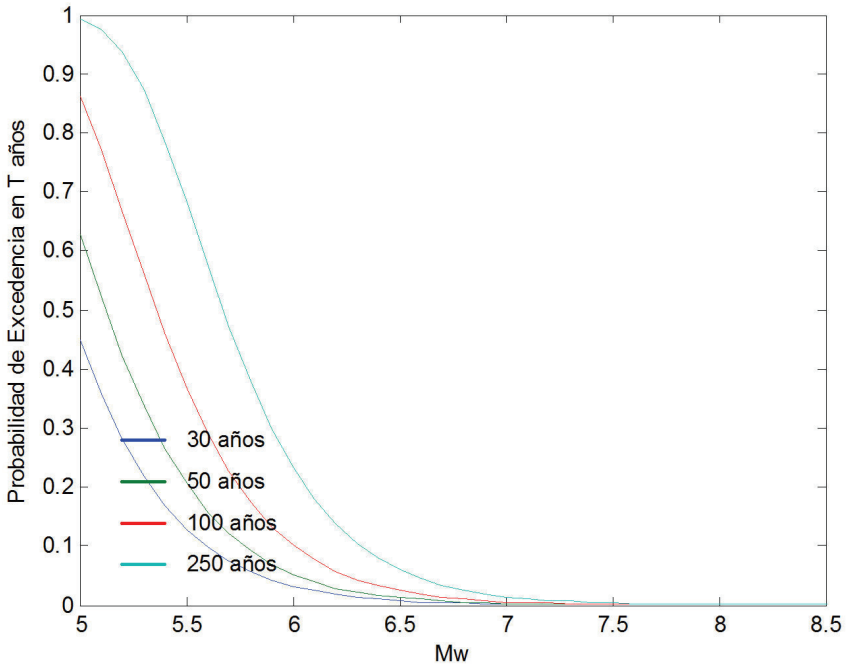


Figura 12: Curvas de excedencia de magnitudes en eventos originados en la zona sismogénica 3

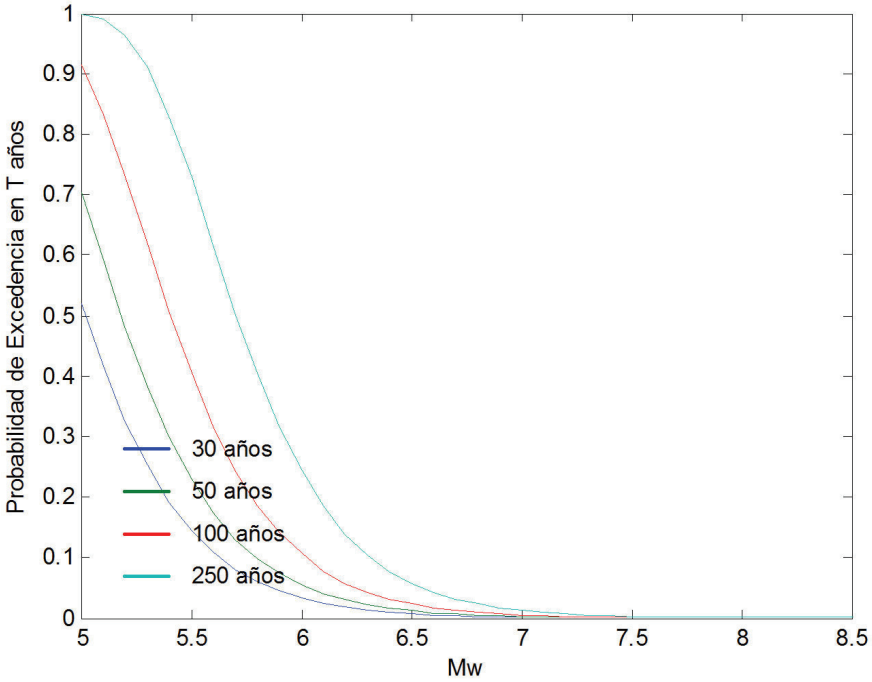


Figura 13: Curvas de excedencia de magnitudes de eventos originados en la zona sismogénica 4

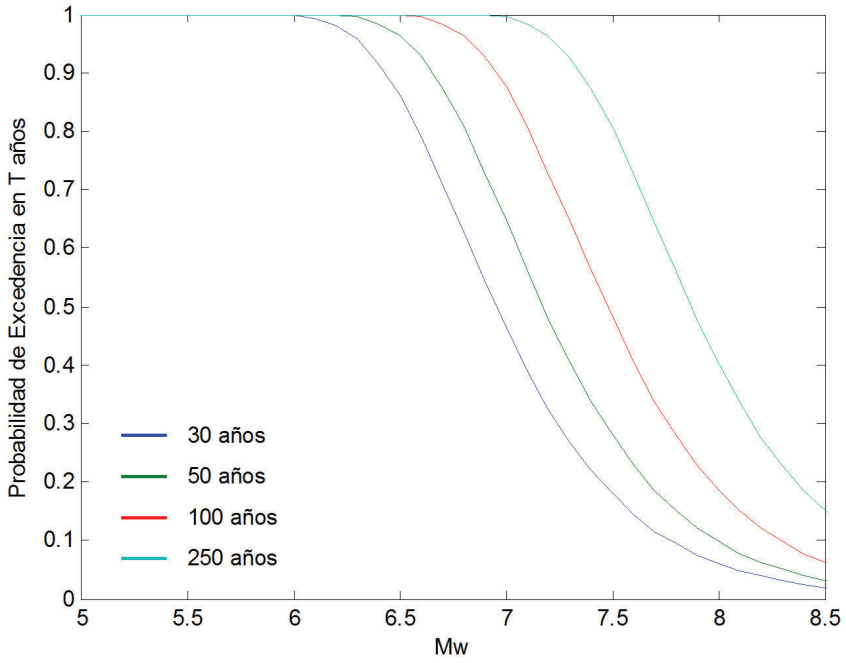


Figura 14: Curvas de excedencia de magnitudes de eventos originados en la zona sismogénica 5

9. Análisis de los resultados

Los resultados plasmados en las curvas de excedencia muestran que para grandes períodos de retorno, en forma probabilista se esperan magnitudes en las fallas de la zona sismogénica 1 mayores a 8. En realidad cada falla tiene un terremoto máximo posible, dada por las características físicas de la misma que hace que sea capaz de liberar hasta una determinada cantidad de energía. Ese conocimiento no se tiene aún para las fallas que amenazan la ciudad, por lo que en estos momentos sólo es posible inferir probabilidades, para períodos de tiempo acotados.

En esta zona se espera un sismo de magnitud de aproximadamente Mw=7,8, para 475 años de período de retorno.

La Figura 15 muestra la sismicidad histórica de la provincia, según el catálogo de sismos históricos SISRA. Se observa que los sismos que más han influido en la ciudad de salta, se ubican en lo que en este

trabajo se ha designado como zona 1. Considerando las magnitudes locales dadas por SISRA, un evento de magnitud 7 o mayor se produciría cada (2010-1692)/2= 159 años promedio en la zona.

Analizando la Figura 10, se tiene que una magnitud Mw de 7, tiene un 40% de probabilidad de ocurrir en 50 años, lo que da un período de retorno de 98 años. Si se considera al sismo de 1844 como de magnitud 7 (fue asignada una magnitud de 6,5) se tendría entonces un período promedio de 106 años. Considerando estos tres sismos emblemáticos en la historia sísmica que afectaron la ciudad de Salta, es en términos probabilistas esperable un sismo de estas características, proveniente de la zona sismogénica 1, en los próximos 30 años.

Considérese que el período 1692-2010 comprende 318 años, mientras que los cálculos se realizaron con una historia sísmica de sólo 44 años, lo que considerando los períodos de retorno de sismos de gran intensidad es insuficiente para estimar con certeza la sismicidad esperada en períodos mayores, por ejemplo aun del doble de duración del catálogo. Pero a la vez, el conocimiento que se tiene de los parámetros de las fuentes simogénicas potencialmente peligrosas es tan escaso, que esta metodología, con todas sus limitaciones, es por ahora lo que se cuenta para definir la acción sísmica. De hecho los espectros de diseño de norma sismorresistente que rige el cálculo de estructuras, es el resultado de un análisis probabilista.

De todos modos, se hace notar que en la zona sismogénica 2, que incluye sismos registrados en el valle de Lerma o en las montañas adyacentes, no incluye el sismo del 27 de Febrero pasado. De todos modos, para un período de retorno de 475 años, el análisis probabilista dio como plausible experimentar un sismo Mw=6,8, originado en el interior del Valle de Lerma, con los 5 eventos que se aprecian en la Figura 7.

El estudio de las fallas del frente de deformación este (zona 1) como las que se encuentran en el Valle de Lerma y sus adyacencias (zona sismogénica 2) surge como una actividad prioritaria a llevar a cabo para

poder estimar la amenaza sísmica en la ciudad con mayor confiabilidad

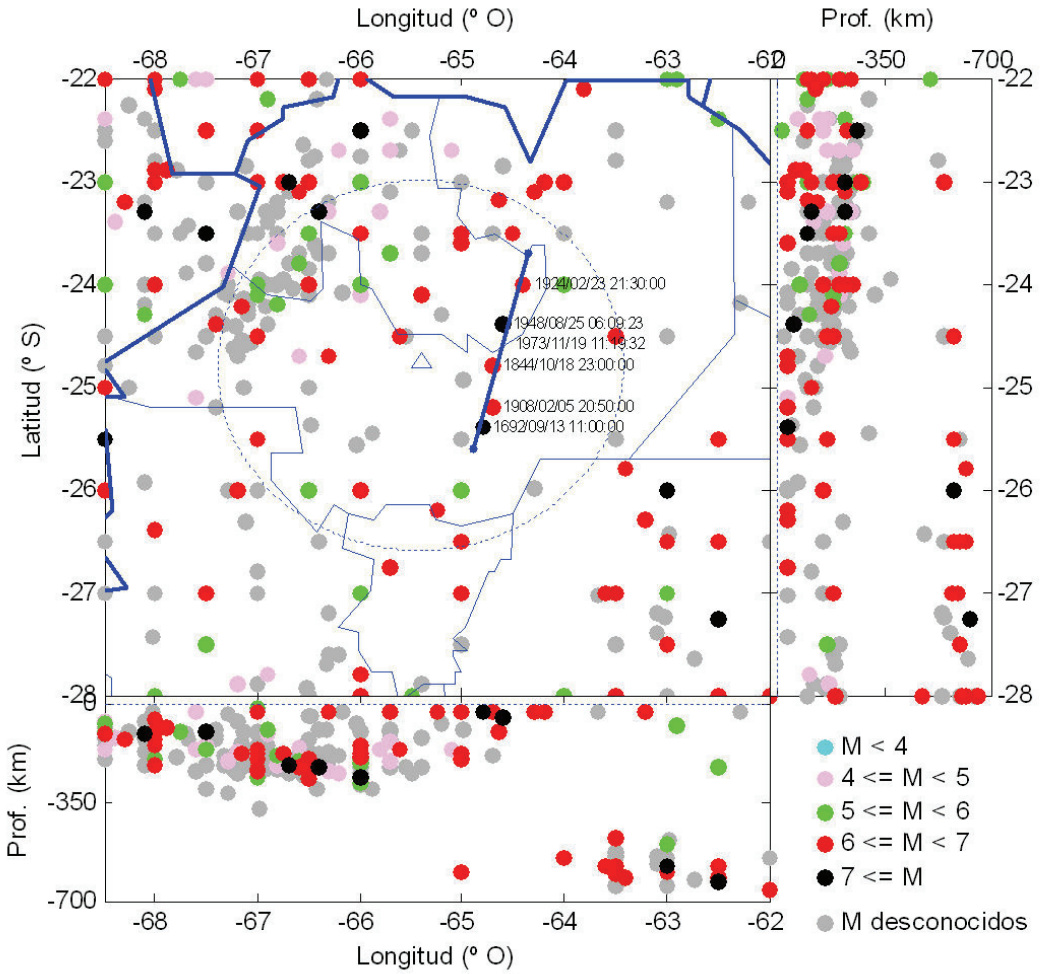


Figura 15: sismicidad histórica en la zona de influencia a la ciudad de Salta.

Además, ciudades como Rosario de la Frontera, Metán y Orán han crecido grandemente y son ciudades ubicadas en cercanías de fallas con actividad cuaternaria reconocida, por lo que la vulnerabilidad ha aumentado también grandemente.

Los sismos originados al oeste, que afectaron históricamente a La Poma, San Carlos, Cafayate, han sido experimentados en la Ciudad de Salta, pero con baja intensidades macrosísmicas por lo que no son prioritarias para la estimación de la peligrosidad de la ciudad de Salta, pero sí para las localidades de los Valles Calchaquíes.

10. Determinación de aceleraciones en "roca" en la Ciudad de Salta

Para el cálculo sismorresistente de las estructuras la norma en vigencia define la acción sísmica en términos de espectros de respuesta elástico de la pseudoaceleración del suelo. Por ello la aceleración del suelo es el parámetro más familiar al ingeniero estructuralista, por lo que la amenaza definida en magnitudes será traducida a aceleraciones en roca en la ciudad de Salta.

Se explican en primer lugar las leyes de atenuación utilizadas en el cálculo de la aceleración en roca para el centro de la ciudad de Salta.

10.1 Leyes de atenuación

A fin de determinar las aceleraciones en roca en el sitio de la Ciudad de Salta, se hizo uso de fórmulas de atenuación propuestas para otras zonas sísmicas, ya que no se cuentan con leyes de atenuación definidas para el norte del país.

Mario Bufaliza (INPRES, 1.995)

Está basada en datos de la región centro-oeste de Argentina y se expresa como:

$$\log a = -0.7837 + 0.353M - 1.5981 \log R - 0.00099R \qquad (11)$$

donde M es la magnitud de ondas superficiales para el caso en que M > 6 y de ondas de cuerpo cuando M < 6; R es la distancia hipocentral (en km) y a es la mayor de las aceleraciones pico considerando las dos componentes horizontales, en porcentaje de g. Para sitios cercanos a las fallas se modifica la ecuación para considerar la saturación del movimiento, lo que conduce a la siguiente expresión: (en g)

$$\log a = -0,7837 + 0,353M - 1,5981 \log D - 0,00099D \quad (12)$$

donde D se define como:

$$D = \sqrt{R^2 + (4,32\ 10^{0.073M})^2} \quad (13)$$

Donovan (1.973)

Fue deducida en base a datos de todo el mundo, en gals

$$a = 1320\ \mathrm{e}^{0,58M}\ (R + 25)^{-1,52} \quad (14)$$

Castano (INPRES, 1.977)

Su expresión dada en gals, es apropiada para campo lejano con distancias epicentrales mayores a los 100 km:

$$a = 1150\ \mathrm{e}^{0,7M}\ (R + 4M)^{-1,6} \quad (15)$$

Patwardhan at al. (citado en Idriss, 1.978)

Para sismos someros de corteza propuso,

$$a = 186\ \mathrm{e}^{1,04M_s}\left(R + 0,864\ \mathrm{e}^{0,463M_s}\right)^{-1,90} \quad (16)$$

Woodward-Clyde Consultants (INPRES, 1.982)

Proponen para sismos de corteza superficiales con datos del oeste de Estados Unidos, la siguiente expresión en g.

$$a = 0,141\, e^{1,04 M_s} (R + 0,775\, e^{0,5 M_s})^{-1,75} \quad (17)$$

Gil – Nafa - Zamarbide (INPRES, 1.982)

Desarrollaron para la zona central-oeste de Argentina la siguiente relación en g.

$$a = 0,063\, e^{M_s} (R + 0,193\, e^{0,714 M_s})^{-1,4} \quad (18)$$

Riera (1.994)

Es adecuada para eventos con distancias epicentrales mayores a 10 km y se expresa en gals.

$$a = 5993\, S^{0.34}\, \Delta^{-1} \left(1 + 0,408\, S^{0,29} \ln \Delta\right)^{-1} \quad (19)$$

donde S es el "área de fractura de la falla ", que se define para zonas intra-placa como:

$$S = 10^{\frac{(M - 7,078)}{0,709}} \quad (20)$$

Se aprecia que estas fórmulas dependen de la magnitud y la

distancia al punto considerado.

Crouse (1991)

Relación de atenuación para zonas de subducción, con focos menores a 400 km, (en gals):

$$a = 578.25\ \mathrm{e}^{1.76M+0.00916h}\left(\mathrm{R}+1.58\,\mathrm{e}^{0.608\,\mathrm{M}}\right)^{-2.73} \qquad (21)$$

10.2 Niveles de diseño considerados

En los últimos años, se impone el diseño por desempeño según lineamientos de VISION 2.000 (SEOAC, 1.995) que especifica cuatro niveles de diseño. Los mismos se resumen en la Tabla 2. El nivel de diseño de la norma argentina corresponde al sismo “raro”.

Tabla 2: Sismos de diseño, para el diseño por desempeño (VISION 2.000)

Sismo	Vida útil	Probabilidad de excedencia	Período de retorno
Frecuente	30 años	50 %	43 años
Ocasional	50 años	50 %	72 años
Raro	50 años	10 %	475 años
Muy raro	100 años	10 %	970 años

Dependiendo del tipo de estructura y las consecuencias de su colapso para el medio ambiente y la vida humana se definen distintos niveles de diseño (Serva, 1.992; United States Committee on Large Dams, 1.985).

Se determinarán por lo tanto las aceleraciones en suelo firme para estos 4 niveles de diseño, desagregado por cada una de las zonas sismogénicas propuestas en este trabajo.

La Tabla 3 resume datos necesarios para el cálculo de las aceleraciones en la ciudad de Salta, originadas por cada fuente sismogénica considerada en el análisis. La segunda columna muestra la máxima magnitud registrada en cada zona (que figura en el catálogo utilizado); la tercera columna, muestra el valor medio (entre paréntesis desviación estándar) de los valores de magnitud del catálogo; R es la distancia al "punto más cercano" y en las dos últimas columnas se observa el valor esperado de las magnitudes para los cuatro niveles de diseño explicados antes, considerando una distribución tipo Poisson y Gumbel tipo I, respectivamente. La distribución tipo Gumbel I se incluye a título informativo.

Tabla 3: Características de cada zona sismogénica

Zonas	**Magnitud Máxima**	**Magnitud media**	**R (km)**	**Magnitudes esperadas**	
				Poisson	**Gumbel**
1	6.01	4.77 (0.52)	66	6.56	6,34
				6,85	6,55
				7,90	**7,30**
				8,31	7,59

2	4,93	4,60 (0,32)	41	5,19	5,40
				5,54	5,53
				6,82	**6,01**
				7,31	6,19
3	5,14	4,42 (0,38)	72	4,94	5,41
				5,12	5,55
				5,77	**6,08**
				6,01	6,28
4	5,03	4,65 (0,27)	118	5,01	5,55
				5,18	5,69
				5,79	**6,20**
				6,03	6,39
5	6,80	4,75 (0,53)	131	6,95	6,77
				7,17	6,97
				7,99	7,68
				8,30	7,94

10.3 Resultados

En las Tablas 4, 5, 6 y 7 se anotan los valores de aceleración obtenidas para cada sismo de diseño y cada zona sismogénica considerada, con los correspondientes valores medios y su desviación estándar (última fila).

Tabla 4:Aceleraciones en roca para un “sismo frecuente” (en g)

Ley de atenuación	**Sismo de diseño 1**				
	Zona 1	**Zona 2**	**Zona 3**	**Zona 4**	**Zona 5**
Bufaliza	0,04	0,03	0,01	0,004	0,01
Donovan	0,06	0,05	0,02	0,013	0,04
Castano				0,015	0,05
Patwardhan	0,04	0,02	0,01	0,004	0,02
Gil-Nafa-Zamarbide	0,09	0,05	0,02	0,011	0,05
W-Clyde Consult.	0,05	0,03	0,01	0,005	0,03
Crouse	0,09	0,05	0,02	0,006	0,05
Riera	0,03	0,02	0,01	0,004	0,02
Valor medio (Desv. estándar)	**0,06 (0,02)**	**0,03 (0,01)**	**0,01 (0,01)**	**0,008 (0,005)**	**0,03 (0,02)**

Tabla 5: Aceleraciones en roca para un “sismo ocasional” (en g)

Ley de atenuación	**Sismo de diseño 2**				
	Zona 1	**Zona 2**	**Zona 3**	**Zona 4**	**Zona 5**
Bufaliza	0,04	0,03	0,01	0,004	0,02
Donovan	0,08	0,06	0,03	0,014	0,04
Castano				0,016	0,05
Patwardhan	0,05	0,03	0,01	0,004	0,02
Gil-Nafa-Zamarbide	0,11	0,06	0,02	0,013	0,07
W-Clyde Consult.	0,07	0,04	0,01	0,006	0,03
Crouse	0,12	0,06	0,02	0,008	0,07
Riera	0,04	0,03	0,01	0,005	0,02
Valor medio (Desv. estándar)	**0,07 (0,03)**	**0,05 (0,02)**	**0,02 (0,01)**	**0,009 (0,005)**	**0,04 (0,02)**

Tabla 6: Aceleraciones en roca para un sismo “raro” (en g)

Ley de atenuación	**Sismo de diseño 3**				
	Zona 1	**Zona 2**	**Zona 3**	**Zona 4**	**Zona 5**
Bufaliza	0,11	0,09	0,02	0,007	0,03
Donovan	0,14	0,12	0,04	0,021	0,06
Castano				0,024	0,09
Patwardhan	0,11	0,09	0,02	0,008	0,05
Gil-Nafa-Zamarbide	0,21	0,16	0,04	0,023	0,12
W-Clyde Consult.	0,15	0,12	0,02	0,011	0,07
Crouse	0,23	0,16	0,04	0,017	0,13
Riera	0,06	0,07	0,02	0,008	0,02
Valor medio	**0,14**	**0,12**	**0,03**	**0,015**	**0,07**
(Desv. estándar)	**(0,06)**	**(0,04)**	**(0,01)**	**(0,007)**	**(0,04)**

Tabla 7: Aceleraciones en roca para un sismo “muy raro” (en g)

Ley de atenuación	**Sismo de diseño 4**				
	Zona 1	**Zona 2**	**Zona 3**	**Zona 4**	**Zona 5**
Bufaliza	0,15	0,15	0,02	0,008	0,04
Donovan	0,18	0,16	0,04	0,024	0,08
Castano				0,028	0,11
Patwardhan	0,15	0,13	0,02	0,009	0,06
Gil-Nafa-Zamarbide	0,26	0,22	0,05	0,028	0,15
W-Clyde Consult.	0,20	0,16	0,03	0,014	0,09
Crouse	0,28	0,21	0,05	0,023	0,16
Riera	0,08	0,09	0,02	0,010	0,03
Valor medio (Desv. estándar)	**0,18 (0,07)**	**0,16 (0,04)**	**0,03 (0,01)**	**0,018 (0,009)**	**0,09 (0,05)**

11. Conclusiones y recomendaciones

Se han determinado en este trabajo las aceleraciones probables en roca en la ciudad de Salta, debido a sismos generados en cinco zonas sismogénicas con influencia en la ciudad, según se desprende de la sismicidad histórica e instrumental registrada en un radio de 200 km.

En general se desprende que hay dos zonas sismogénicas con especial incidencia en la peligrosidad sísmica de la ciudad: el frente de deformación este situado a una distancia promedio de 90-100 km y la sismicidad que se registra en el Valle de Lerma o sus límites. Por ello, es prioritario el estudio de las fallas, cuya actividad generan los movimientos sísmicos plausibles de ocasionar daños en la Ciudad de Salta.

A fin de obtener los valores de aceleración en la superficie en cada sitio de la ciudad, se debe considerar la influencia de los mantos cuaternarios de suelo y en especial, las últimas capas de suelo más blandos.

REFERENCIAS

Aguiar, R., 2.003, Espectros sísmicos de riesgo uniforme para verificar desempeño estructural en países latinoamericanos, *Conferencia ofrecida en el XI Seminario Iberoamericano de Ingeniería Sísmica*, Mendoza, Argentina.

Bonilla, M.G., Mark, R.K. and Lienkaemper, 1.984, Statistical relations among earthquake magnitude, surface length and surface fault displacement, *Bulletin of the Seismological Society of America*, V 74, pp 2.379-2.411.

Carreño, M.L., Cardona, O. and Barbat, A.H., 2007, Urban seismic risk analysis: an holistic approach, *Natural Hazards*, Vol 40, pp 137-172

Cornell, C. A., 1968, Engineering seismic risk analysis, *Bull Seismological Sociey of America*, Vol 58, pp1583-1606

Crouse. C.B., 1991, Ground motion attenuation equations for earthquakes on the Cascadian subduction zones, *Earthquake Spectra*, vol 7 (2), pp 201-236.

Davis y Frolich, 1991, Single-link cluster analysis and earthquake aftershocks; decay laws and regional variations. *J. Geophys. Res*. 96, 6335– 6350.

Donovan N. C., 1.973, A statistical evaluation of strong motion data including the February 9, 1.971, San Fernando earthquake. *Proceedings of 5th World Conference on Earthquake Engineering*, Vol 1.

Gutemberg, B. and Richter, C. F., 1.944, Frequency of earthquakes in California, *Bulletin of Seismic Society of America*, 34, pp 1.985-1.988.

Haarala Orosco, M., 2.006, *EARTHSTAT, programa para el cálculo de la peligrosidad sísmica*, La Caldera, Salta.

Idriss, I.M.,1978, Characteristics of earthquake ground motion, *Proceeding of the ASCE Geotehnical Engineering Division Speciality Conference: Earthquake Engineering and Soil Dynamics*, Vol III, pp 1151-1265.

INPRES, Instituto Nacional de Prevención Sísmica, 1.977, *Zonificación sísmica de la República Argentina, Publicación Técnica Nº 5*, San Juan, Argentina.

INPRES, 1.982, Microzonificación sísmica del Valle del Tulum – Provincial de San Juan – *Informe Técnico*, Vol 2.

INPRES, 1.995, Microzonificación sísmica de Mendoza y Gran Mendoza *Informe Técnico*.

INPRES-CISOC-103, 1.983, Normas Argentinas para Construcciones Sismorresistentes, Parte I, (Construcciones en general), *INPRES*, San Juan, Argentina

International Seismological Centre, *On-line Bulletin*, Internatl. Seis. Cent., Thatcham, United Kingdom, 2.001. http://www.isc.ac.uk/Bull

Riera, J and Doz, G., 1.994, Sobre la definición de la excitación sísmica considerando las características básicas de falla, *Lectures of Postgraduate Course in Civil Engineering*, UFRGS, Brasil.

SEAOC, 1.995, Vision 2.000: Report on performance based seismic engineering of buildings, *Structural Engineering Association of California*, Sacramento, EEUU.

Serva, L, 1.992, An analysis of the world major regulatory guides for NPP seismic design, *ENEA, Technical Report* RT/DISP/92/03, ISSN/0393-3016.

Slemmons, B. D., 1.984, Evaluation of seismic hazards in earthquake-resistant design: identification and characterization of active faults, *EERI Seminar*, July 1.984, Stanford University.

Stewart, J. F., Chiou, S-J., Bray, J. Graves, R., Somerville, P., Abrahamson, N., 2.001, Ground Motion Evaluation Procedures for Performance –Based Design, *Technical Report*, PEER 2.001/09, Pacific Earthquake Engineering Research Center

US Geological Service, On-Line Catalogue http://neic.usgs.gov/neis/epic

United Status Committee on Large Dams, 1.985, Guidelines for selecting parameters for dam projects.

La difícil tarea de la seguridad informática. Análisis de un caso en una organización típica salteña.

Fredi Aprile, Sergio Appendino, H. Beatriz P. de Gallo[1]

(faprile@copaipa.org.ar), (sappendino@copaipa.org.ar), (bgallo@copaipa.org.ar)

Resumen

Las tareas relacionadas con la seguridad informática en cualquier empresa parecen no ser tan simples. La historia que se presenta a continuación, refleja casos reales de acontecimientos surgidos en este ámbito y región, que genera una situación problemática, y de que manera la podemos afrontar.

Palabras Claves: Seguridad - Seguridad informática - Información - Comité de seguridad.

1. La Historia.

La falta de conciencia en la seguridad de los datos es un problema presente en cualquier empresa o institución.

La información procesada, almacenada y consultada por los medios tecnológicos actuales es un recurso más de cualquier organización. En este contexto, la jefatura de una empresa ha confiado en una persona, el *Lic. Juan Seguro*, la responsabilidad de la seguridad informática. Antes de asumir esta tarea, él había ingresado al

[1] Fredi Rene Aprile es Licenciado en Análisis de Sistemas. Analista Senior en el Poder Judicial de Salta en las áreas de seguridad informática, auditorias e informática forense. Es docente de la UCASAL. En el ámbito privado, se desempeña como consultor informático con especialización en la seguridad informática.
Sergio Appendino es Ingeniero en Sistemas, CISA (Auditor Interno de Sistemas Certificado), Docente de UCASAL y Perito Informático en la Corte de Justicia Provincial y Federal de Salta.
H. Beatriz P. de Gallo, es Ingeniera en Computación, Master en Administración de Negocios, docente e investigadora de la UCASAL. Es además perito informático de la Corte de Justicia de Salta. En el ámbito privado, se desempeña en la organización y gestión de proyectos de capacitación in company.

organismo como analista y programador de sistemas, pero siempre se había manifestado interesado en la seguridad informática por su tesis de graduación y algunos cursos e investigaciones que estaba realizando.

Después de dos años de asumir la tarea, Juan se enfrentó con un serio problema: la seguridad de la información es un concepto nuevo en la empresa y por lo tanto observa que existe una “Falta de conciencia generalizada sobre la seguridad de los datos por parte de todos los usuarios”.

2. La inversión.

La magnitud de los riesgos asociados en materia de seguridad informática involucra la inversión de nuevas tecnologías y de recursos humanos. Cada pedido de presupuesto a los niveles superiores debe argumentarse el doble que en las otras actividades.

Sin embargo la inversión se aprobó inmediatamente luego que la página institucional fue hackeada, o aquella vez en que un virus paralizó a toda la organización bloqueando el acceso a los sistemas y al uso de los servicios.

¿Juan deberá esperar que suceda otro incidente de seguridad para que se aprueben nuevas inversiones?

3. ¿Qué es lo que debemos proteger?

Existe abundante información en la empresa. Ante la consulta de Juan de qué información debe proteger, la respuesta de todos los gerentes es “Todo se debe proteger”. Juan pensó en su momento que sería lindo también que su casa estuviese protegida con policías de seguridad en todas las puertas, alarmas, sistemas de vigilancia, investigadores, etc. pero este costo es más elevado que su propia casa.

Lo mismo sucede en cualquier organización, es imposible proteger absolutamente todo o proteger información que por su característica es pública. Por ello se debe definir cuáles son los activos a proteger, cómo clasificar la información, qué dato es público, qué información es privada o sensible, quienes deben/pueden acceder. ¿Quién tiene la responsabilidad de definir qué información es crítica y sensible y definir sus accesos?. ¿Quién es realmente el dueño de los datos?

4. Uso de Internet.

¿Debe permitirse el uso del Chat y del correo electrónico personal?

Se contrató para toda la organización el servicio de Internet, pero por cuestiones de seguridad se limitó el acceso solo a páginas autorizadas y libres de riesgos. A los dos días de implementar esta política, sufrió serios reclamos de los gerentes en el sentido de que se estaba limitando el uso a la información. ¿Debería permitir el uso libre de Internet?. ¿Bajo qué condiciones?

5. Software y licencia de uso.

Después de realizar un análisis de toda la red, Juan detectó numerosos programas instalados vía Internet, mails o medios removibles en la mayoría de los equipos. Por ello se encargó de desinstalar todo y bloquear las futuras instalaciones por una cuestión de seguridad, compatibilidad con los sistemas de la organización y debido a que la mayoría de esos programas no contaban con las licencias legales de uso. Al día siguiente fue acusado con todos los términos posibles que se pueda imaginar por parte de los empleados y gerentes, ya que no podían escuchar música, los protectores de pantalla desaparecieron, no se podía instalar nada, no podían chatear con otros empleados, etc. , etc. . ¿Debería volver atrás con la desinstalación de los programas?. ¿Cómo debe hacer frente a todas las críticas?

6. Los dispositivos removibles

La creciente proliferación de nuevos medios tecnológicos removibles (pendrive, mp3, mp4, memorias de cámaras digitales, etc.) conlleva el riesgo de que cualquier empleado introduzca o extraiga información o programas de/hacia las pcs. Esto tiene un riego adicional si en dichos dispositivos se

encuentran virus informáticos o cualquier programa no autorizado con contenido malicioso que puede causar daños e incompatibilidad en los equipos o los sistemas. Considerando esto, cada vez que bloqueó el acceso a los dispositivos de almacenamiento, tuvo serias quejas de los usuarios porque no podían intercambiar información con otras personas. ¿Debería permitir el uso de medios removibles?

7. Las claves de acceso.

Se habilitó usuario y contraseñas para todas las personas para el acceso a los sistemas, pero algunas de ellas estaban muy molestas ya que tenían una tarea más de recordar esos datos y en algunos casos todos se prestaban o intercambiaban las contraseñas. ¿Cómo debería capacitar a los empleados en este aspecto?. ¿De quién es la responsabilidad de la confidencialidad de las claves?

8. Conflicto de intereses.

Se han producido alteraciones del clima laboral con compañeros de trabajo en cuanto a la operatividad de los sistemas. Cada vez que se adiciona seguridad a las aplicaciones informáticas, se agrega una tarea más a los usuarios y los programadores indican que disminuye la perfomance de los sistemas. La seguridad y la operatividad de los sistemas a veces no se complementan, entonces ¿Quién debería definir ese equilibrio?

Todas las políticas que trata de implementar Juan son consensuadas por la jefatura de sistemas. Aún así, no son bien recibidas por los altos niveles jerárquicos debido a que se imparten desde un área con menor jerarquía en la organización.

Por ejemplo: un gerente general comentó que ningún empleado o subjefe le puede decir a él de que manera debe navegar en Internet o que programa debe instalar. Por otro lado, no es posible aplicar sanciones por incumplimiento a empleados de otros sectores que no sean del área de sistemas.

Sirva esto como ejemplo, para mostrar que los recaudos técnicos que se puedan tomar, no son suficientes para proteger la información, y se requiere de un marco normativo que contenga y establezca las reglas de uso de la información mediante decisiones tomadas, debatidas, consensuadas y aprobadas por los altos niveles jerárquicos

de la organización. ¿Qué organismo, área, unidad organizativa se debería crear? . ¿Qué nivel de jerarquía debería tener?

9. Hasta aquí el problema…

¿Cómo lo solucionamos? pues básicamente con políticas de seguridad que deben ser evaluadas y aprobadas en un organismo interno (Comité de Seguridad), conformado por adecuados niveles de decisión en la organización y que le otorgan al área de seguridad el presupuesto necesario para llevar a cabo la implantación de las políticas, controlar su seguimiento y evaluar acciones correctivas.
El Comité de Seguridad debería fijar en estas Políticas los objetivos que tiendan a:

- Concientizar sobre la seguridad de información. Esta concientización debe incluir un plan de capacitación extenso y profundo, acompañado de una campaña de difusión sobre la importancia de los sistemas informáticos en el negocio de la empresa
- Generar una cultura informática en todos los usuarios dirigida a identificar los “activos intangibles” con el mismo valor que “activos tangibles”.
- Definir responsabilidades de todos los usuarios pertenecientes a la organización.
- Analizar, identificar y definir procedimientos y controles en todos los niveles que involucren riesgos a la seguridad de la información.

El Comité de Seguridad no es un equipo técnico-informático, debe ser un grupo de decisión interdisciplinario en el que tengan competencia todas las áreas de la empresa

10. Conclusiones

No se agota el tema con estas consideraciones, por el contrario, se abre un camino de análisis y preguntas que cada empresa o institución deberá andar por sí misma.

Lo importante es comenzar a crear conciencia en los estamentos de decisión, acerca de la necesidad de formalizar acciones de contención y gestión de la información, que hagan que las tecnologías de la información y la comunicación cumplan con el rol fundamental que hoy tienen en la empresa: la alineación con el negocio.

Bibliografía

Subsecretaría de Tecnologías de Gestión, Secretaría de la Gestión Pública, Año 2005, Modelo de Política de Seguridad de la Información para Organismos de la Administración Pública Nacional. ONTI (Oficina Nacional de Tecnologías de la Información).

International Organization for Standarization, Año 2000, Norma 17799– Código de práctica para la administración de la seguridad de la información.

International Organization for Standarization, Año 2005, Norma 27001– Norma para la administración de la seguridad de la información.

Consejo Profesional de Agrimensores, Ingenieros y Profesiones Afines – Universidad Católica de Salta – I-Sec Information Security Education Center, Año 2005, Apuntes "Jornadas de especialización en seguridad de la informacion – curso ISO 17799".

Maestría en Administración de Negocios, año 2006, Apuntes sobre Dirección de Recursos Humanos – Capítulos I, II, III, IV, V, VI - Prof. Edgardo Visñuk.

Vidrios Metálicos Masivos

L. Marta , G. Lavorato , C. Berejnoi , C. Bernal , J. Moya [*]

jmoya@ucasal.net

Resumen:

La primera generación de vidrios metálicos (es decir, aleaciones metálicas con estructura amorfa) fueron desarrollados en la década de 1960 en forma de cintas o chapas con espesores del orden de los 40 μm. Luego, en los años 70, las aleaciones amorfas base Fe o Co encontraron rápidamente su lugar en la industria como núcleos de transformadores, debido a sus excelentes propiedades magnéticas blandas. Un resumen de los mismos y de sus propiedades magnéticas puede encontrarse en el volumen anterior de estos cuadernos. En este artículo haremos hincapié en la nueva familia de los vidrios metálicos, los llamados vidrios metálicos masivos o sus siglas en ingles BMGs (*Bulk Metallic Glasses*) denominados así por obtenerse con espesores mayores al milímetro. Se hará una introducción al problema de la capacidad de amorfización de las aleaciones metálicas, se comentarán sus propiedades mecánicas y corrosivas y sus aplicaciones incipientes como material estructural. Finalmente, se mostrarán algunos resultados experimentales obtenidos gracias al fruto de diversas colaboraciones.

Palabras claves: Vidrios metálicos masivos, aleaciones amorfas, aleaciones metálicas avanzadas.

[*] Leonardo Marta: en Ingeniero Aeronáutico (U.N.L.P) y becario doctoral del CONICET. Su tema de tesis es “Desarrollo de aleaciones amorfas masivas base Fe para aplicaciones estructurales” y trabaja actualmente en el I.Es.I.Ing de la Fac. de Ingeniería e Informática de la UCASAL.

Gabriel Lavorato: es estudiante de Ingeniería Industrial de la Facultad de Ingeniería de la UBA y becario de investigación en dicha institución (Beca Estímulo FI-UBA). Su tema de tesis de graduación es: “Materiales magnéticos amorfos y nanoestructurados masivos para dispositivos eléctricos”.

Carlos Berejnoi: es Ingeniero Metalúrgico (U.N.L.P.), Doctor en Ingeniería (U.N.L.P.), profesor en la F.I.- U.N.Sa. Su actividad principal, actualmente, es la investigación en mecánica de fractura, trabajando en el Grupo Mecánica de Fractura de la U.N.Comahue.

Celina Bernal: es Ingeniera Mecánica (UNMdP), Doctora en Ciencia de Materiales (UNMdP) y miembro de la Carrera del Investigador del CONICET. Su actividad principal, actualmente, es la investigación en materiales en la Facultad de Ingeniería de la Universidad de Buenos Aires.

Javier Moya: es Ingeniero Mecánico (UBA), Doctor en Ingeniería (UBA), miembro de la Carrera del Investigador del CONICET y profesor en la FII-UCASAL. Su actividad principal, actualmente, es la investigación en materiales en el I.Es.I.Ing de la FII-UCASAL.

1. Historia:

Los vidrios metálicos masivos fueron desarrollados en los años 80 en aleaciones base Pd y en los 90 en sistemas metal-metal (base metálica y aleantes metálicos), en aleaciones de base Zr-, Mg-, La-, siendo los metales de transición temprana los primeros que lograron grandes avances (Zr-, Ti-, y Hf). El desarrollo de BMGs de metal de transición tardía fue impulsado por la necesidad de tener aleaciones de menor costo y mayor disponibilidad. Fueron entonces desarrollados a partir de 1995, BMGs en base Fe- en un sistema ferromagnético de aleación Fe-Al-Ga-P-C-B con espesores máximos de 1 mm. Luego, en 2003 fue desarrollado el primer BMG base Fe no magnético a temperatura ambiente en el sistema Fe-Mn-Cr-Mo-C-B con espesores de 4 mm y fue llamado acero amorfo estructural, o *Structural Amorphous Steel* (SAS). Un año más tarde, se estableció que el agregado de tierras raras como Y, Er, Yb, Gd o Dy incrementaba drásticamente el espesor crítico a 12 y 16 mm.

Los BMGs en base Fe- por sus óptimas propiedades mecánicas (resistencia a la fractura 3500-5500 MPa y dureza 850-1370 Hv), magnéticas y químicas despiertan gran interés en el campo tanto científico como industrial, esperando ser aplicados en la industria automotriz, naval, aeronáutica, quirúrgica, deportiva, etc., en un tiempo estimado no superior a 5 años.

Actualmente una de las mayores complejidades en la obtención de estos materiales radica en la gran velocidad de enfriamiento a la que deben ser sometidos en el estado líquido para mantener su estructura amorfa en estado sólido, limitando esto el espesor máximo que puede obtenerse.

2. ¿Qué son los Vidrios Metálicos Masivos (Bulk Metallic Glasses, BMGs)?

El nombre de vidrios se refiere a que la estructura del material fue congelada del estado líquido sin permitirle cristalizar como una aleación metálica convencional, manteniendo una estructura atómica desordenada (amorfa) como la de un vidrio.

Debajo de cierta temperatura los cambios en la estructura atómica de la aleación en estado líquido se vuelven más difíciles y a medida que desciende la temperatura la estructura aparece congelada por una

especie de arresto cinético de los átomos. Esta temperatura es conocida como temperatura de transición vítrea (T_g) (Ver Fig. 1).

Los BMGs son aleaciones amorfas masivas, que a diferencia de los primeros vidrios metálicos de un espesor muy pequeño, tienen mayores dimensiones pudiendo ser utilizados también como componentes estructurales.

Para que las aleaciones de base Fe-, Mg-, Zr-, Ni-, Ti- conserven la estructura amorfa del estado líquido, en el estado sólido, es necesaria una gran velocidad de enfriamiento (Rc), evitando así proceso de cristalización normal, lo que depende de la capacidad de formar amorfo (CFA) que tenga la aleación, es decir, de su resistencia a la cristalización.

La velocidad de enfriamiento y la capacidad de formar amorfo de la aleación, junto con el método de producción que se utilice, determinará el tamaño máximo de la pieza que se obtiene [1].

Cada aleación tiene parámetros propios de la capacidad de formar amorfo (o la resistencia que la aleación opone a formar una estructura cristalina), y de velocidad critica de enfriamiento (velocidad por debajo de la cual la aleación cristaliza y no forma la estructura amorfa).

Otro punto interesante es su comportamiento plasto-viscoso en la región de liquido superenfriado ΔT_{xg} Fig. 1), que es la diferencia de temperatura entre el inicio de la cristalización T_x y la temperatura de transición vítrea T_g, sin que pierda su estructura amorfa, pudiéndose trabajar como vidrios o plásticos y obteniendo piezas metálicas de nanodimensiones de perfecto acabado sin la necesidad de un mecanizado posterior. Esto supone grandes ventajas económicas y ambientales evitando el proceso de mecanizado y sus residuos. Actualmente esta región de comportamiento plastoviscoso es de aproximadamente 50K para las aleaciones base Fe-, al ser tan acotada dificulta el proceso de industrialización.

3. Capacidad de Formar Amorfo (CFA) de las aleaciones

La capacidad de formar amorfo, CFA, puede interpretarse como la facilidad que tiene una aleación en estado líquido, de que al momento de ser enfriada forme una estructura atómica amorfa sin una apreciable formación de fases cristalinas. Entender y predecir la capacidad de formar amorfo CFA es la llave para el desarrollo de nuevos tipos de

BMGs, ya que uno de los mayores tropiezos para el uso de aleaciones no cristalinas es la baja CFA o la gran velocidad de enfriamiento requerida.

En cuanto a este tema la CFA incluye dos puntos fundamentales:

1) Cuán estable es la fase líquida.

2) Cuán resistente es a la cristalización.

Si la fase líquida es estable después del enfriamiento y/o las fases cristalinas son de difícil precipitación, entonces la formación de la estructura amorfa se verá facilitada [2].

Las aleaciones capaces de formar amorfo tienen que ser muy estable en la región de líquido superenfriado. Inoue et al. [3] han propuesto tres reglas empíricas que rigen esta capacidad.

La aleación debe tener al menos 3 componentes.

Tiene que haber una diferencia significativa en el tamaño atómico de los componentes de la aleación (mayor al 12%).

El calor de formación tiene que ser negativo.

La diferencia de tamaño en los átomos desestabiliza la estructura cristalina y genera tensiones volumétricas en la red, siendo esto un punto fundamental en la elección de una nueva aleación amorfa. Hay elementos que se añaden a las aleaciones con este fin, como el Boro y Carbono (radios pequeños comparados con el Fe) y el Itrio, Niobio y Erbio (radios grandes).

Otra regla utilizada es que la aleación se sitúe cerca de su eutéctico, o sea que tenga el punto de fusión más bajo para la aleación. Según Cheney [4] todo buen formador de amorfos situará su aleación cerca del eutéctico y tendrán altas tensiones volumétricas.

La búsqueda de un parámetro o criterio que indique la alta CFA es hoy un punto de gran interés y motivo de esfuerzos científicos. Algunos parámetros simples han sido sugeridos, basados en propiedades físicas y temperaturas características. El más conocido es la temperatura reducida de transición vítrea T_{rg} (la relación entre la temperatura de transición vítrea T_g y la de líquido T_l) propuestos por Turnbull [5]. Basándose en la cinética de la nucleación de cristal y la viscosidad del líquido, cuando una aleación líquida es enfriada hasta T_g la viscosidad

aumenta rápidamente y se forma el vidrio. Otro indicador es la región de líquido superenfriado ΔT_{xg} (la diferencia de temperatura entre el inicio de la cristalización y la temperatura de vitrificación) basado en la consideración de que un mayor ΔT_{xg} está ligado a una menor formación de cristales. Sin embargo, estudios empíricos demuestran que tanto un indicador como el otro no tienen buena correlación en ciertos BMGs dependiendo de su base. El fenómeno de la formación de la estructura amorfa es un proceso muy complicado y no parece susceptible a análisis teóricos. Por otra parte, se han propuesto métodos teóricos como el desarrollado por Senkov y su grupo [6]; allí se evalúa el tamaño de los átomos de los elementos de la aleación, el sitio que ocuparán en la celda atómica (intersticial o sustitucional) y la concentración de cada uno. En base a esto, se calcula la tensión que generarán los átomos en la celda por sus diferentes volúmenes y el lugar que ocuparán. Por encima de una tensión crítica se dice que la aleación tiene capacidad de formar amorfo. Con este método uno puede tener una idea si una aleación va a dar un resultado positivo antes de ensayar la muestra, es decir, sirve para evaluar a priori en una aleación su capacidad de formar amorfo. si tenemos una alta tensión volumétrica (por encima de la crítica) entonces tendremos más probabilidades de obtener un BMG.

Las mayores CFA han sido obtenidas para las aleaciones Pd- y Pt-base, seguidas por Cu-, Ni-, Fe-, y las aleaciones de Co- base. Generalmente la CFA de aleaciones ternarias se ve mejorada con el descenso de la temperatura de líquido T_l.

Los diagramas tiempo-temperatura-transformación, TTT, contienen toda la información que se necesita para predecir la CFA (Fig. 1). La curva Rc no debe cortar la curva de transición cristalina al enfriarse para asegurar una estructura amorfa al solidificar la aleación. En este diagrama se entiende fácilmente la manera de generar un sólido amorfo, el líquido debe ser enfriado lo suficientemente rápido desde la temperatura de líquido hasta la temperatura de transición vítrea T_g sin interceptar la curva TTT

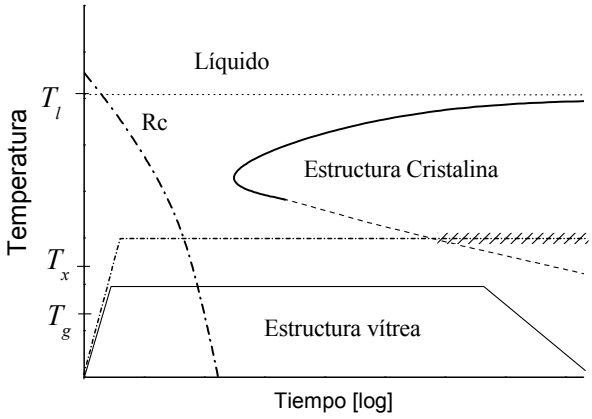


Fig.1: Diagrama Tiempo-Temperatura-Transformación (TTT). La cristalización ocurre entre T_l y T_g y puede ser evitada mediante una rápida velocidad de enfriamiento (Rc). Rc es la velocidad de enfriamiento crítica, más allá de ella ocurre la cristalización. Cuando el sólido amorfo es calentado a ritmo constante, la muestra comienza a cristalizar a la temperatura T_x.

4. Propiedades Mecánicas

Las aleaciones amorfas tienen propiedades mecánicas que se diferencian de sus símiles cristalinas y que las vuelven muy interesantes cuando se las compara con ellas. Por ejemplo la resistencia última puede triplicar la de una aleación convencional. Su comportamiento plasto-viscoso en la región de liquido superenfriado da lugar a deformaciones que superan el 20000%, pudiendo en esta región obtenerse piezas estructurales de un acabado perfecto sin la necesidad de un posterior mecanizado, evitando así residuos y pérdidas. Las propiedades mecánicas en los BMGs están estrechamente ligadas con la estructura atómica y electrónica.

Los BMGs tiene un comportamiento asimétrico en ensayos de tracción y compresión, la fluencia de los BMGs obedece el criterio de Mohr-Coulomb en lugar del criterio de Von Mises. Experimentalmente el límite de elasticidad de los BMGs es típicamente 2%, mientras que el límite elástico de las aleaciones cristalinas es generalmente menor a 0.65%. Por otra, parte la capacidad de deformación plástica se ve afectada por el desorden atómico que dificulta el movimiento de las dislocaciones. La resistencia está estrechamente ligada a parámetros físicos como la temperatura de transición (T_g), el módulo de elasticidad, y los coeficientes de expansión térmica.

Una manera de distinguir un material dúctil de uno frágil es el índice de solidez S definido por:

$$S = \frac{\mu}{B}$$

Donde μ y B son el módulo de corte y el módulo de elasticidad volumétrica respectivamente. Lewandowski extendió este parámetro de los sólidos cristalinos a los vidrios metálicos y encontró una correlación similar con el valor crítico de S. En los metales cristalinos el valor crítico de S es 0.3, por debajo de este valor el comportamiento del metal será dúctil. En los BMGs este valor es 0.41. Esta relación también puede ser expresada en función del coeficiente de Poisson, ν, de manera tal que los BMGs con alto ν suelen resultar dúctiles. [7]

La deformación homogénea en vidrios metálicos siempre ocurre a altas temperaturas ($>0.70\ T_g$) mostrando asombrosa plasticidad en la región de líquido superenfriado.

En la foto de abajo se ve el microengranaje más pequeño del mundo que fue construido en un BMG de base Níquel en la región del líquido superenfriado, sin la necesidad de un acabado posterior. La comparación de algunas propiedades mecánicas significativas, entre aleaciones amorfas y cristalinas puede verse en la Tabla 1.

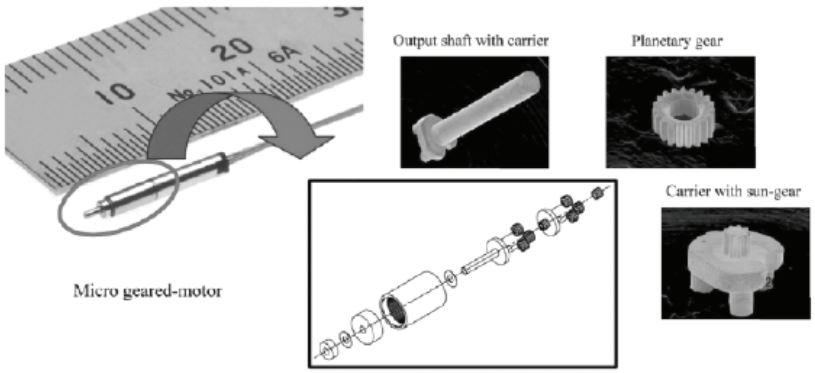


Fig. 2: Microengranaje más pequeño del mundo, 1.5 mm de diámetro, construido en un BMG de base Níquel. [8]

Tabla 1: Hv (microdureza Vickers), E (módulo de elasticidad), σ_f(tensión de fractura), σ_y (tensión de fluencia). [9] [10].

Composición	Propiedades Mecánicas			
	Hv	E(GPa)	σ_f(MPa)	σ_y(MPa)
$(Fe_{0.5}\ Co_{0.5})_{0.75}Bo_{0\cdot2}Si_{0\cdot05})_{96}Nb_4$	1250	210	4210	
$(Co_{0.6}\ Fe_{0.4})_{0.75}Bo_{0\cdot2}Si_{0\cdot05})_{96}Nb_4$	1230	210	4170	
$(Fe_{71.2}\ B_{24}Y_{4.8})_{96}Nb_4$	1120	195	4000	
$Fe_{48}Cr_{15}Mo_{14}Cr_{15}Er_2B_6$	1300	200	4000	
Stainless Steel 316 LVM		187	800	
Stainless Steel 304	129	200	505	
AISI 6150 Steel	384	205	1240	1225
Austenitic Stainless Steel201		197	1276	965

5. Resistencia a la Fatiga

La fatiga es un fenómeno por el cual la rotura de materiales es debido a cargas dinámicas cíclicas, se da a una tensión menor que la máxima del material y de una manera catastrófica y frágil aunque el material sea dúctil. Existe una tensión por debajo de la cual no existe la fatiga, a esta se la denomina limite de fatiga.

Las propiedades de los materiales cambian generalmente con la temperatura, por ejemplo en la mayoría de los metales cristalinos la resistencia a la fatiga disminuye con el aumento de la temperatura. Sin embargo en los BMGs el efecto de la temperatura es limitado: estudios realizados sobre el crecimiento de fisura por fatiga demostraron que no hay cambios significativos con el aumento de temperatura [11].

El comportamiento en fatiga de los BMGs es muy bueno para aleaciones de base Zr- (si se lo compara con aceros de alta resistencia y aleaciones de Titanio) y es pobre en aleaciones base Cu. Por

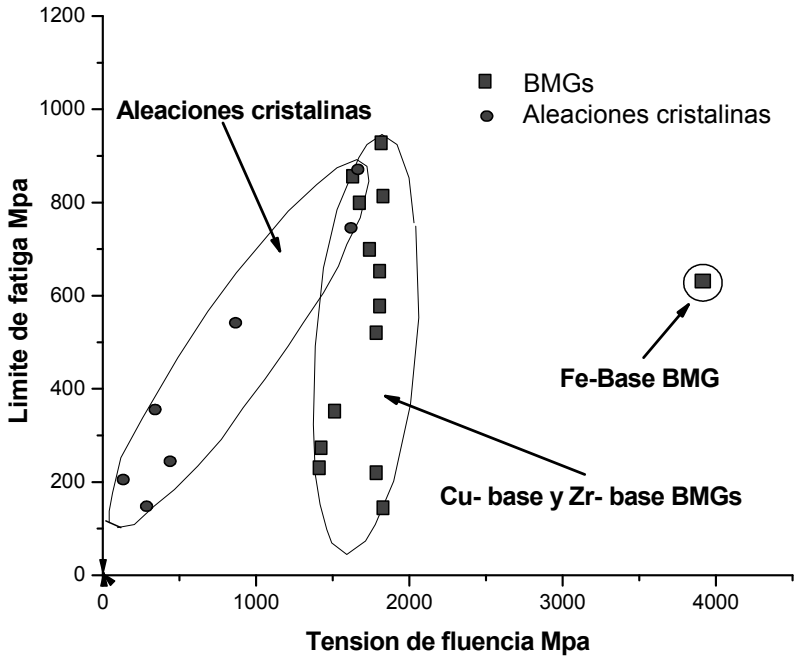


Fig. 3: algunos Fe- base y Cu- base BMG muestran una alta tensión pero bajos coeficientes de fatiga (coeficiente de fatiga es la relación entre el límite de fatiga y la tensión de fluencia) [12].

ejemplo, sometido a un ensayo de fatiga el BMG $Fe_{48}Cr_{15}Mo_{14}Er_2C_{15}B_6$ reportó un límite de fatiga de aproximadamente 680 MPa, mucho mayor que los BMGs base Zr- (aproximadamente 150 MPa) y que una aleación de aluminio 6082 (aproximadamente 72 MPa). Sin embargo, un dato a tener en cuenta es que cuando se supera el límite de fatiga, la vida del material se reduce drásticamente a un número de ciclos mucho menor. En este caso, la vida útil de esta aleación es menor que la de una aleación convencional de aluminio o de un acero de alto nitrógeno. Por otra parte, tal como los cristalinos, los BMGs son muy sensibles al acabado de la superficie en lo que respecta a su resistencia a la fatiga.

6. Corrosión

La corrosión es considerado un fenómeno de superficie mediante el cual masa del material es transferida al medio ambiente por procesos de transporte ya sean químicos, físicos o electroquímicos. Los BMGs tienen buena resistencia a la corrosión por dos motivos fundamentales:

1) Su composición, que no está limitada con límites de solubilidad, puede alearse con elementos que promuevan el pasivado del proceso corrosivo.

2) La ausencia de características microestructurales como borde de grano o dislocaciones, que sirven como inicio del proceso corrosivo en los materiales.

La primera generación de vidrios metálicos, los obtenidos mediante grandes enfriamientos y en espesores del orden de los micrones, suelen ser más factible de un diseño químico que prevenga la corrosión. En este sentido, se han producido aleaciones amorfas con mayor capacidad de pasivado que los aceros inoxidables. Pero el advenimiento de los BMGs junto con un diseño más ajustado de la composición química hace comprometer el logro alcanzado por sus predecesores. Sin embargo, estudios recientes demuestran que los aceros amorfos estructurales presentan un mejor comportamiento a la corrosión que aceros convencionales de alto contenido de Cr (12% peso).

7. Aporte al estudio de los BMGs. Resultados Experimentales

Los primeros resultados exitosos los hemos obtenido en dos aleaciones de composición $(Fe_{0.375}Co_{0.375}Si_{0.25-x}B_x)_{96}Nb_4$, con X= 0.20 (primeramente obtenida por Inoue et al. [13]) y con X= 0.15 (una modificación de la anterior) a las que llamaremos X20 y X15 respectivamente. La aleación se preparó en crisoles de cuarzo fundiéndola con un horno a inducción en una atmósfera inerte de Argón. Luego se procedió a la colada con enfriamiento rápido en atmósfera de aire por el método de inyección en molde de cobre (refrigerado por agua) en forma de cuña, a fin de obtener distintas velocidades de enfriamiento en el lingote y evaluar el espesor máximo de material amorfo que se puede obtener. Fig. 4.

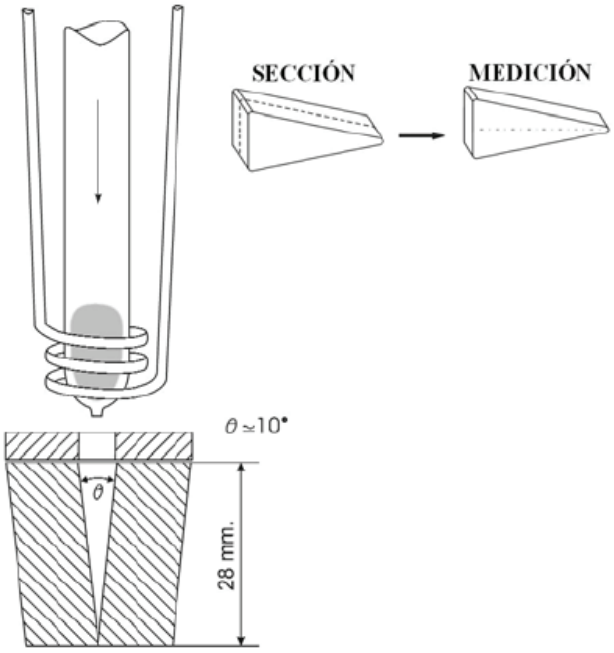


Fig. 4: Esquema de la técnica de colada en molde de cobre y de la forma de la muestra obtenida.

En la aleación X20 se logró un espesor crítico de 2.5 mm y con muy buenas características de material magnético blando.

En la Fig. 5 se muestra una fotocomposición de micrografías ópticas de la muestra X20 donde se pueden apreciar claramente dos regiones: cristalina y amorfa. En la región amorfa no se evidencian bordes de grano. Existe una zona de transición donde se observan

granos inmersos en una matriz amorfa caracterizada por fisuras debidas a tensiones en el material.

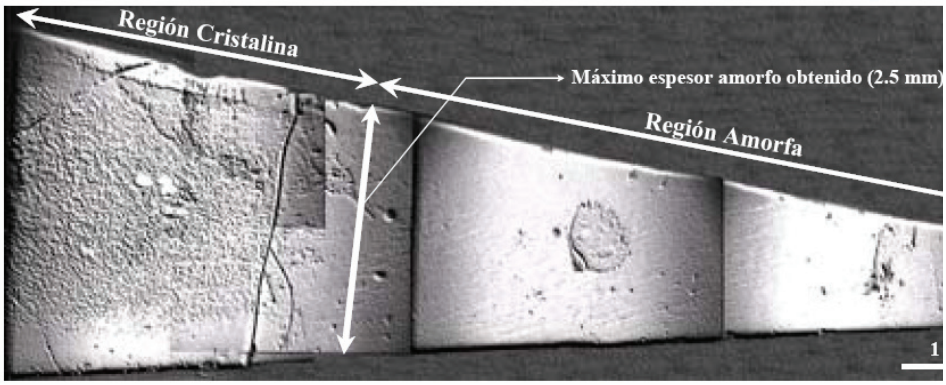


Fig. 5: Fotocomposición de micrografías ópticas de la muestra amorfa X20. Puede verse la región amorfa y cristalina.

La aleación X15 es un composite de dendritas de una solución sólida α-Fe(Co) dispersas en una matriz amorfa de composición similar a la X20.

Presenta un deterioro en sus propiedades magnéticas blandas debido a la mayor anisotropía magnetocristalina que posee la fase α-(FeCo) en comparación con la amorfa. No obstante la presencia de dendritas proporcionaría al material una mayor tenacidad, lo que queda evidenciado en el descenso de los valores de microdureza (en relación con la otra aleación) obtenidos experimentalmente. Las dendritas actuarían como crack stopper durante la propagación de fisuras, por consiguiente este material sería apto para uso estructural con una resistencia mecánica estimada en 2760 MPa.

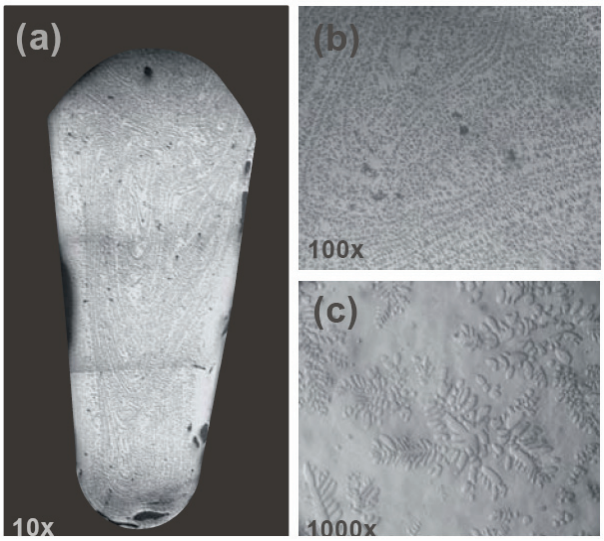


Fig. 6: Micrografía óptica de la aleación con dendritas.

En la Fig. 6 se puede apreciar una única región donde habría dos fases: una matriz amorfa poblada de dendritas que siguen un patrón de líneas de flujo.

Con el objetivo de su empleo como material magnético blando, se trabajó en el estudio de la aleación $Fe_{76}P_5(Si_{0.3}B_{0.5}C_{0.2})_{19}$ (previamente obtenida por Makino et al. [14] mediante dos técnicas de solidificación rápida: i) Se hicieron cintas de un espesor de aproximadamente 40 μm por planar-flow casting y, ii) cilindros de 1 y 2 mm de diámetro por injection copper mold casting [15].

La elección de esta composición química se realizó a partir de un sistema formador de vidrio que no presentara elementos costosos que suelen utilizarse para incrementar la capacidad de formación del amorfo y, a la vez, que no deterioren las propiedades magnéticas blandas. Los estudios realizados sobre el material masivo dieron una imanación de saturación de unos 1.44 T y un campo coercitivo entre 4 y 7 A/m resultando, este material, superior a los materiales magnéticos tradicionales. También, los resultados del estudio de pérdidas magnéticas obtenidas para este material fueron muy competitivos en relación a los materiales tradiciones a frecuencias cercanas a los 60 Hz, siendo dichas pérdidas entre 2 y 3 veces menores que la del Fe(Si) de grano orientado. Estos valores indican el gran potencial en cuanto a la utilización a nivel industrial de estos materiales (aún en etapa de

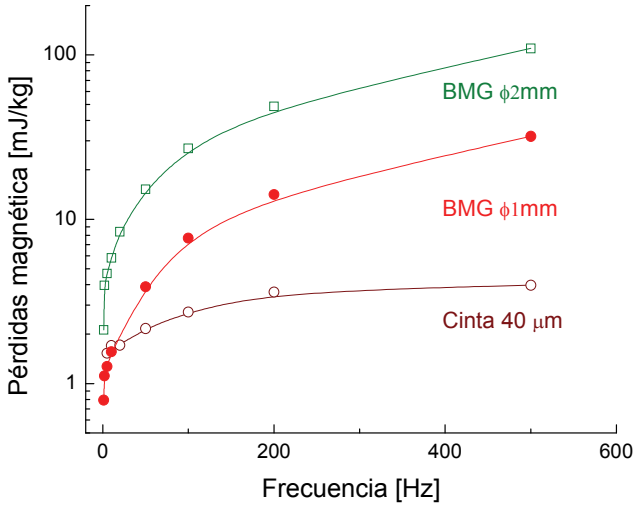


Fig. 7: Estudio de las pérdidas magnéticas en aleaciones amorfas masivas (BMG) y en forma de cintas de composición $Fe_{76}P_5(Si_{0.3}B_{0.5}C_{0.2})_{19}$ en función de la frecuencia.

desarrollo) y se espera que en el futuro sean utilizados en núcleos de transformadores, sensores inductivos y otros dispositivos eléctricos.

Finalmente y, buscando contribuir al desarrollo industrial de estos nuevos materiales, se prepararon tres aceros amorfos estructurales diferentes en aleaciones base Fe-Cr-Mo-C con el agregado de B, Y y/o Gd como elementos amorfizadores [16]. Dos de estas aleaciones fueron coladas también en atmósfera de aire y en moldes en forma de planchuelas de 2 mm de espesor y 20 mm de ancho (y largo libre). La tercera aleación fue una combinación de 56% en peso de un acero comercial AISI430 y 44% de una de las aleaciones anteriores con contenido de Y. En este caso se obtuvo un cilindro de 2 mm de diámetro totalmente amorfo demostrando que es posible la obtención de BMGs a partir de materia prima industrial.

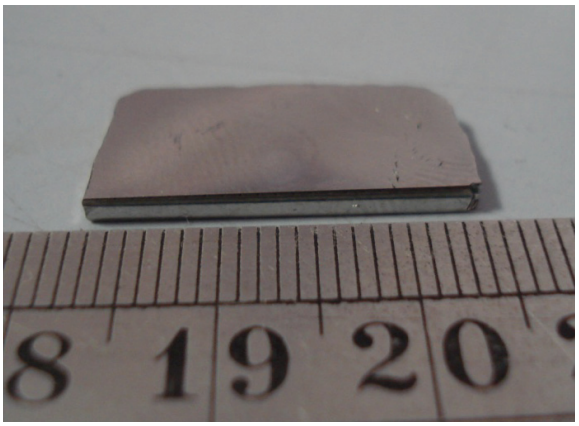


Fig. 8: Aspecto externo de acero amorfo estructural de composición $Fe_{48}Cr_{15}Mo_{14}C_{14}B_6Gd_2$ obtenido en forma de planchuelas

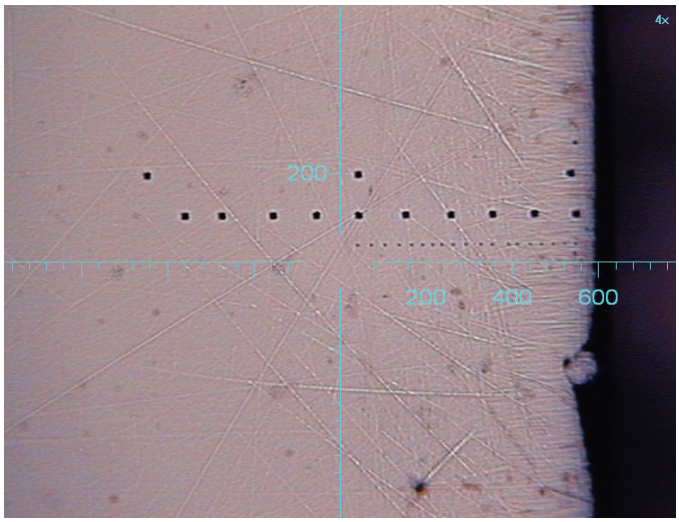


Fig. 9: Micrografía óptica donde se distinguen improntas grandes y pequeñas de nanoindentación, a partir de las cuales que se estudiaron algunas propiedades mecánicas del material (escala en μm).

8. Conclusión:

Se ha pretendido dar una introducción sobre los nuevos materiales llamados Vidrios Metálicos Masivos (o Bulk Metallic Glasses, BMGs) que en el caso particular de las aleaciones ferrosas no magnéticas reciben el nombre de Aceros Amorfos Estructurales (Structural Amorphous Steels). Los resultados presentados, fruto de colaboraciones, muestran el interés por su desarrollo industrial como material magnético blando o como material estructural. Nuevos experimentos se realizarán en este sentido buscando nuevas aleaciones con una buena combinación de capacidad de formar amorfo y de propiedades magnéticas y/o mecánicas.

Referencias

[1] W.Y. Liu, H.F. Zhang, A.M. Wang, H. Li, Z.Q. Hu; New criteria of glass forming ability, thermal stability and characteristic temperatures for various bulk metallic glass systems; Mater. Sci. Eng. A 459 (2007)

[2] Z.P. Lu, C.T. Liu; A New glass forming ability criterion for Bulk Metallic Glasses; Acta Mater. 50 (2002)

[3] A. Inoue, T. Zhang, A. Takeuchi; Ferrous and non-ferrous bulk amorphous alloys; Mater Sci. Forum 269-272 (1988)

[4] J. Cheney, K. Vecchio; Evaluation of glass forming ability in metals using multi-model techniques; Journal of alloys and compounds 471 (2009)

[5] D. Turnbull, Under what conditions can a glass be formed? Contemp. Phys. 10, 473-483 (1969)

[6] O.N. Senkov, D.B. Miracle; A topological model for metallic glass formation; Journal of Non-crystalline Solids 317 (2003)

[7] J.J. Lewandowski, W.H. Wang, A.L. Greer, Intrinsic plasticity or brittleness of metallic glasses, Philos. Mag. Lett. 85 (2005)

[8] Tomado del Libro BMGs, editores M. Miller y P. Liaw, Capitulo 1, Pag. 20.

[9] F. Bassi, G. Lavorato, J. Moya, H. Sirkin, Searching for structural amorphous steels, Proceedings of: New Developments on Metallurgy and Applications of High Strength Steels, May 26 - 28, 2008 Buenos Aires, Argentina. Publisher TMS. USA (2008)

[10] www.matweb.com

[11] P. Hess, R. Dauskart, Mechanisms of elevated temperature fatigue crack growth; Acta Mater 52 (2004)

[12] Tomado del Libro BMGs, editores M. Miller y P. Liaw, Capitulo 7, Pag. 181.

[13] G. Lavorato, F. Bassi, H. de Rosa, J. Moya; Aleaciones metalicas amorfas y compuestas base Fe para uso magnético y estructural; CONAMET/SAM-2008

[14] A. Makino, Ch. Changa, T. Kubota, A. Inoue. Soft magnetic Fe–Si–B–P–C bulk metallic glasses without any glass-forming

metal elements; Journal of Alloys and Comp. 483 (2009) 616–619.

[15] G. Lavorato, M. Baricco, P. Tiberto, J. Moya et al., A ser publicado.

[16] G. Lavorato, G. Fiore, M. Baricco, J. Moya et al., A ser publicado.