MICEI

Mestrado em Informática e Curso de Especialização em Informática

Cased-Based Reasoning (CBR) systems: introducción y aplicaciones prácticas

Florentino Fernández Riverola

Braga, 11 de mayo de 2007

Escuela Superior de Ingeniería Informática

Área de Lenguajes y Sistemas Informáticos
Departamento de Informática

Universidade de Vigo

Índice de la Presentación

- ✓ **Sistemas CBR** — introducción, características, ciclo de vida, tipos, técnicas
- ✓ **HTTPHUNTING** — detección de ataques HTTP
- ✓ **GENECBR** — análisis de datos bioinformáticos
- ✓ **SPAMHUNTING** — detección de correo spam

SISTEMAS CBR (*Case-Based Reasoning*)

• Kolodner (1983a, 1983b). Paradigma para la resolución de problemas en IA

"*Los humanos utilizan lo aprendido en experiencias previas para resolver problemas presentes*"

Joh (1997)

• Un CBR resuelve problemas por medio de la adaptación de soluciones dadas con anterioridad a problemas similares (Riesbeck *et al*., 1989)

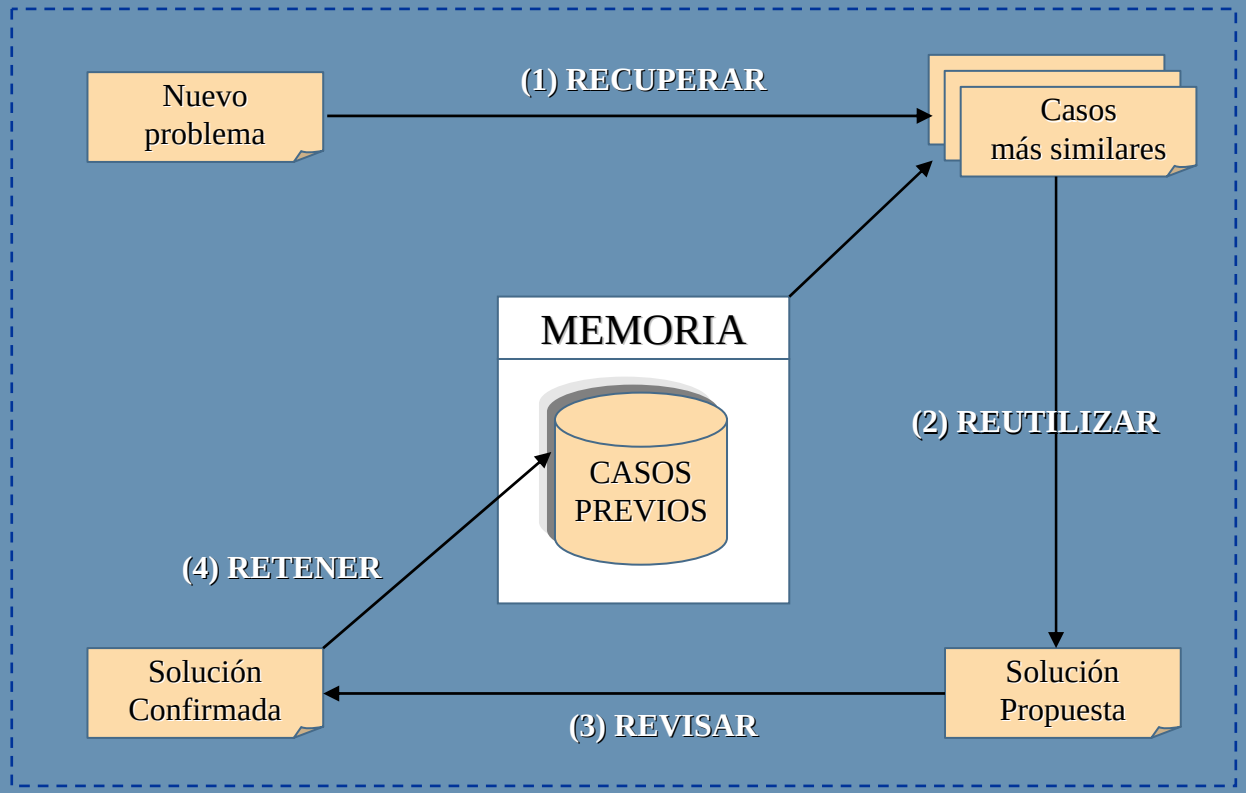
• La base de casos (memoria) del sistema CBR almacena un cierto número de problemas junto con sus correspondientes soluciones:

CASO = *PROBLEMA* + *SOLUCIÓN* [+ *RESULTADO*]

• Cuando surge un nuevo problema, la solución se obtiene recuperando casos similares de la base de casos y estudiando la similitud entre ellos

CICLO DE VIDA DE UN CBR

• 4 etapas secuenciales invocadas cuando es necesario resolver un nuevo problema (Kolodner, 1993; Aamodt y Plaza, 1994; Watson, 1997)



RECUPERACIÓN
Seleccionar los casos más similares al nuevo problema a solucionar

REUTILIZACIÓN
Adaptación de los casos recuperados al nuevo problema

REVISIÓN
Aplicar la solución propuesta a un problema real y determinar su validez

RETENCIÓN o APRENDIZAJE
Integración de nuevo conocimiento en la memoria del CBR

CARACTERÍSTICAS DE LOS SISTEMAS CBR

- Sistemas dinámicos y adaptativos: el número de casos de la memoria cambia, permitiendo la adaptación del sistema a nuevas situaciones
- Permiten la utilización de conocimiento general en la resolución de un problema particular
- Facilitan la organización (indexación) de la información disponible
- Se pueden utilizar casos incompletos (dificultad en la descripción de un problema)
- Los sistemas CBR son conscientes de sus limitaciones (pueden no generar una solución)
- Facilitan el uso de estructuras de datos representativas y flexibles
- La adaptación de casos ayuda a descubrir interrelaciones y estructuras ocultas en los datos
- Los sistemas CBR pueden ser automatizados a diferentes niveles

TIPOS DE SISTEMAS CBR

Diferencias en cuanto a:
- Fases del ciclo de vida implementadas
- Características del dominio
- Tecnologías empleadas en la construcción del sistema

- MBR: **Razonamiento Basado en Memoria**
  - La memoria representa una colección de casos
  - El proceso de razonamiento se corresponde con el proceso de recuperación
  - Utilizan técnicas de procesamiento paralelo
  - Pueden ser utilizados en dominios con fuertes connotaciones sintácticas y semánticas

- ABR: **Razonamiento Basado en Analogías**
  - Resuelven un nuevo problema utilizando casos provenientes de un dominio de conocimiento diferente
  - Centrados en el estudio de mecanismos para la identificación y utilización de analogías entre diferentes dominios

TIPOS DE SISTEMAS CBR

- EBR: **Razonamiento Basado en Ejemplares**
  - Se centran en el aprendizaje de definiciones de conceptos
  - Los casos más parecidos se agrupan en clases
  - La solución de un problema será la solución de la clase a la que pertenece el caso recuperado más similar
- IBR: **Razonamiento Basado en Instancias**
  - EBR centrado en problemas con fuertes connotaciones sintácticas
  - Número elevado de instancias (casos) y falta de conocimiento general sobre el dominio
  - Representación de una instancia utilizando vectores de características
  - En algunos casos, las fases del ciclo de vida pueden llegar a ser automatizadas
- CBR: **Razonamiento Basado en Casos**
  - Engloba el conjunto de los diferentes mecanismos de razonamiento existentes

CBR: TECNOLOGÍA VS. METODOLOGÍA

• Tradicionalmente los sistemas CBR han sido considerados como una tecnología: redes neuronales, genéticos, lógica difusa, ... etc.

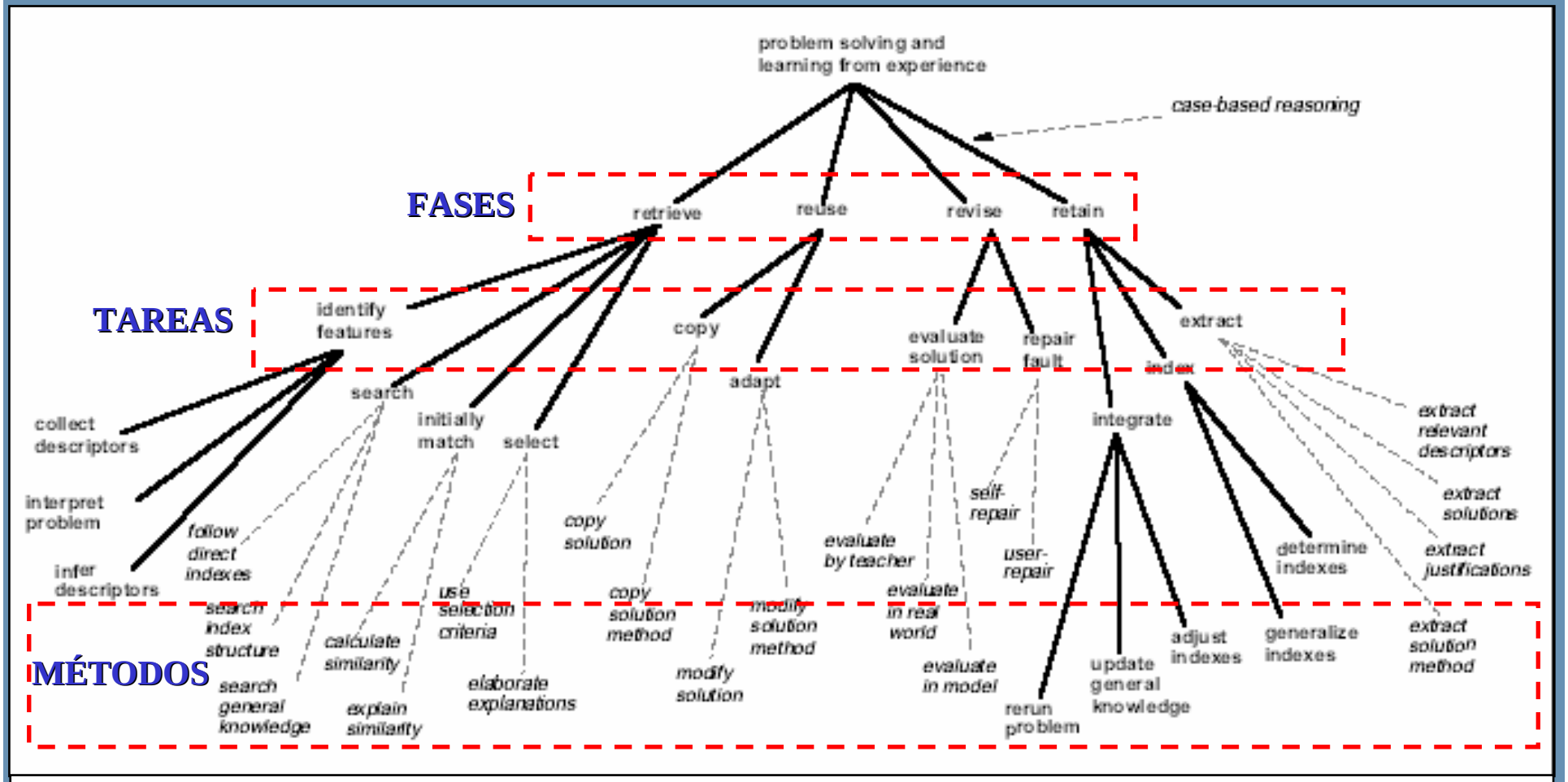
• Conceptualmente, los sistemas CBR se describen como una secuencia cíclica de diferentes etapas formadas por tareas y métodos

Metodología: "*Conjunto de principios organizados, que guían el manejo de situaciones problemáticas del mundo real*"

(Checkland y Scholes, 1990)

• Watson (1998); Kamp *et al*. (1998) proponen los CBR como una metodología que puede utilizar cualquier tecnología existente que respete los principios que define dicha metodología

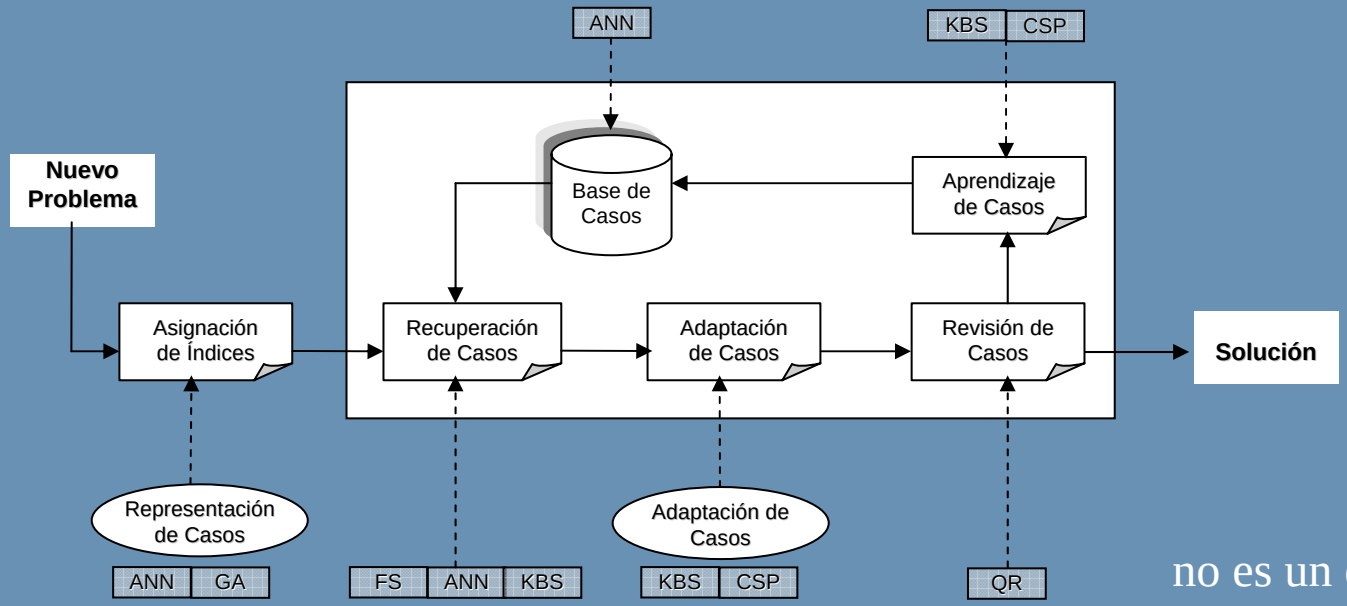
CBR: TECNOLOGÍA VS. METODOLOGÍA



A. Aamodt, E. Plaza (1994); Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications. IOS Press, Vol. 7: 1, pp. 39-59.

TECNOLOGÍAS UTILIZADAS EN SISTEMAS CBR/IBR

• Medsker (1995), realiza una revisión de las diferentes tecnologías utilizadas en la construcción de **sistemas híbridos** basados en el **ciclo de vida de un CBR**



no es un esquema completo...

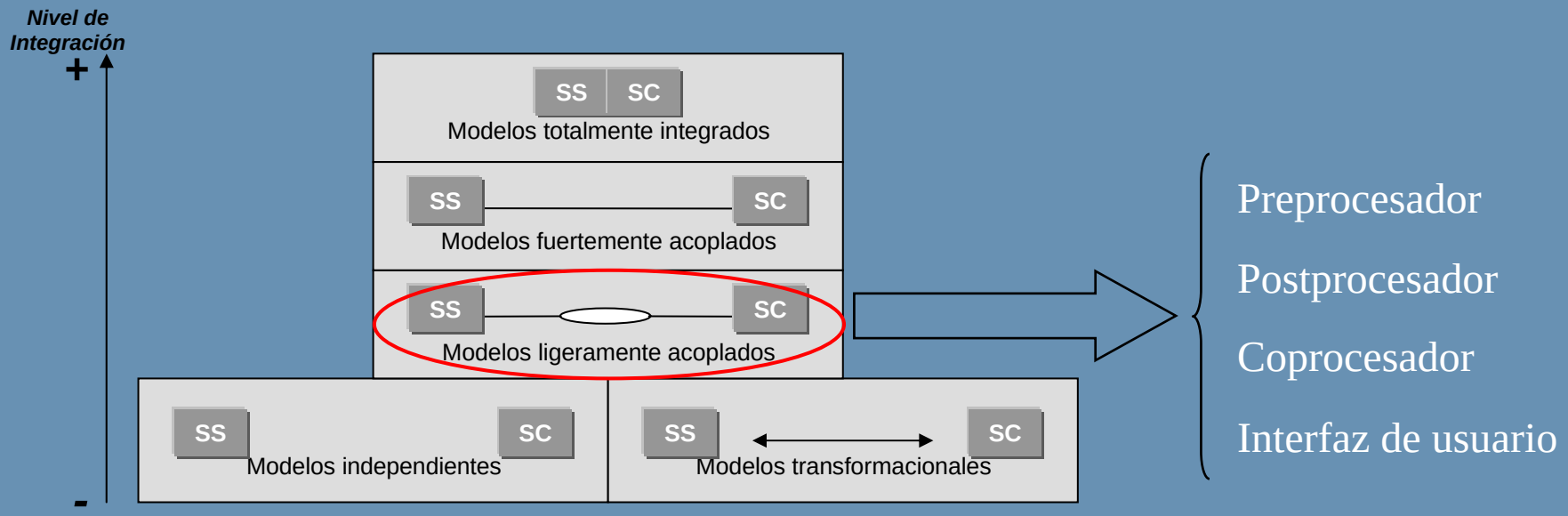
• ¿Cuáles son las distintas posibilidades de interconexión de los mecanismos seleccionados?

SISTEMAS HÍBRIDOS: CLASIFICACIÓN

"El término *híbrido* hace referencia a sistemas compuestos por uno o más subsistemas integrados, cada uno de los cuales presenta un lenguaje de representación y un mecanismo de inferencia distinto"

(Medsker, 1995)

• Medsker y Bailey: *Integration Connectionist-Symbolic*. (5 modos de integración):



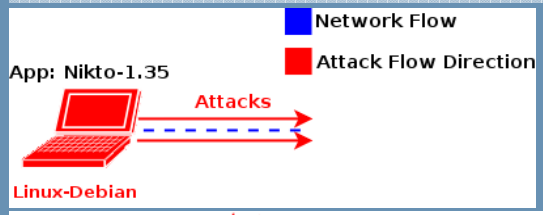
HTTPHUNTING::sistema de detección de intrusos [1/5]

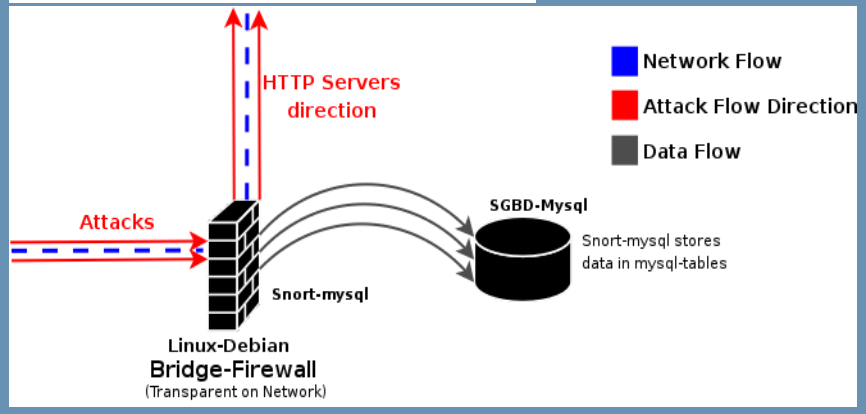
introducción

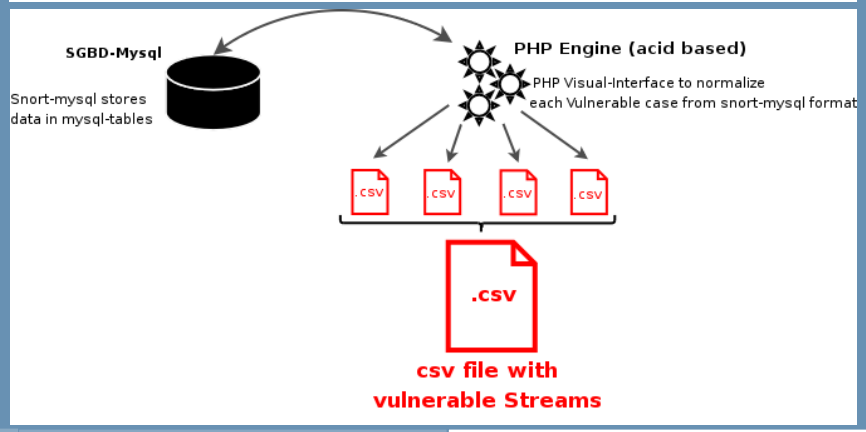
- Intrusión:
  - acceso o utilización de recursos de una máquina de forma no autorizada por un atacante
- Tipos de sistemas IDS:
  - detección de utilización incorrecta (*misuse detection* | *signature-based*)
    - detección de patrones de tráfico de red o datos de aplicación sospechosos
    - sólo detectan ataques previos conocidos
  - detección de anomalías (*anomaly detection system*)
    - reconoce intrusiones identificando contenido *diferente* al tráfico *normal* de la red
    - estado normal en función de: carga de tráfico, protocolos, tamaño de paquetes, ...
  - basados en host (HIDS, *Host Intrusion Detection Systems*)
  - basados en red (NIDS, *Network Intrusion Detection System*)
  - IDS pasivos (*passive IDS*)
    - detectan una posible violación, registran información y generan una alerta
  - IDS reactivos (*reactive IDS*)
    - responden ante actividades ilegales: expulsando al usuario o reconfigurando firewalls

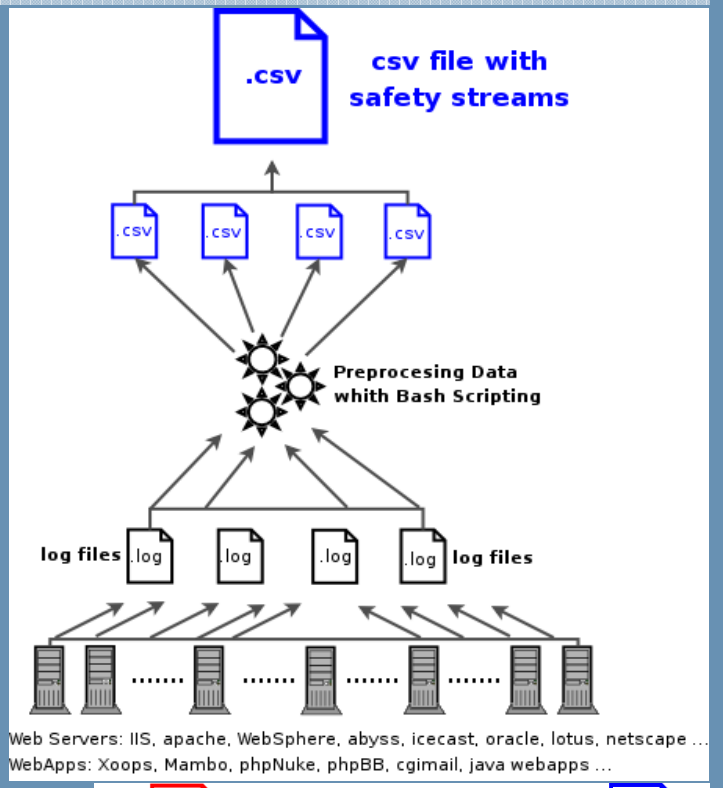
HTTPHUNTING::sistema de detección de intrusos [2/5]

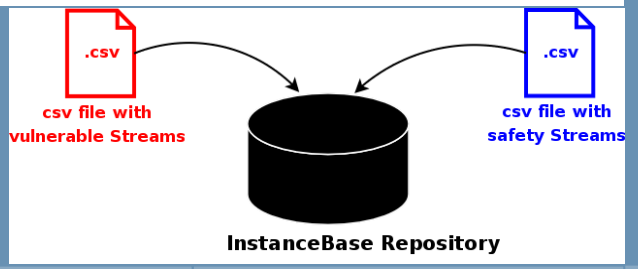
obtención de los datos de prueba







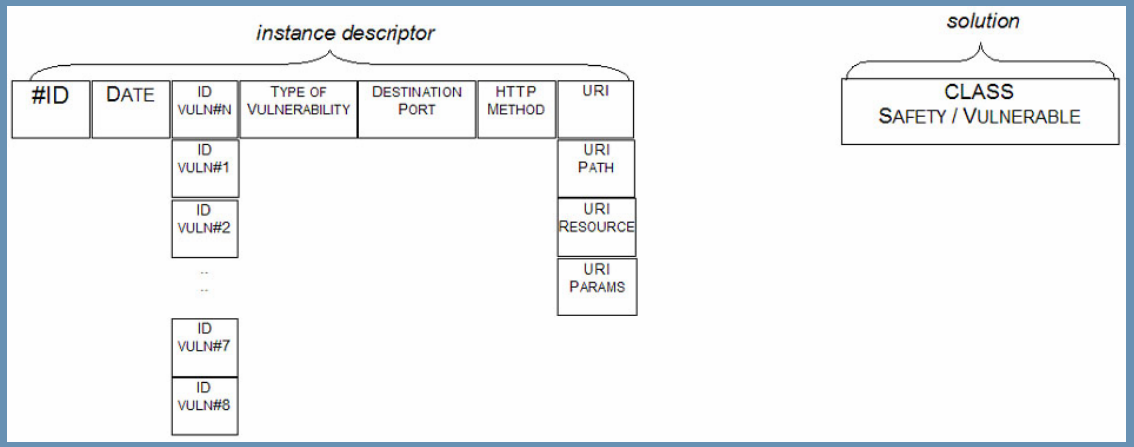




HTTPHUNTING::sistema de detección de intrusos [3/5]

datos disponibles y representación de una trama

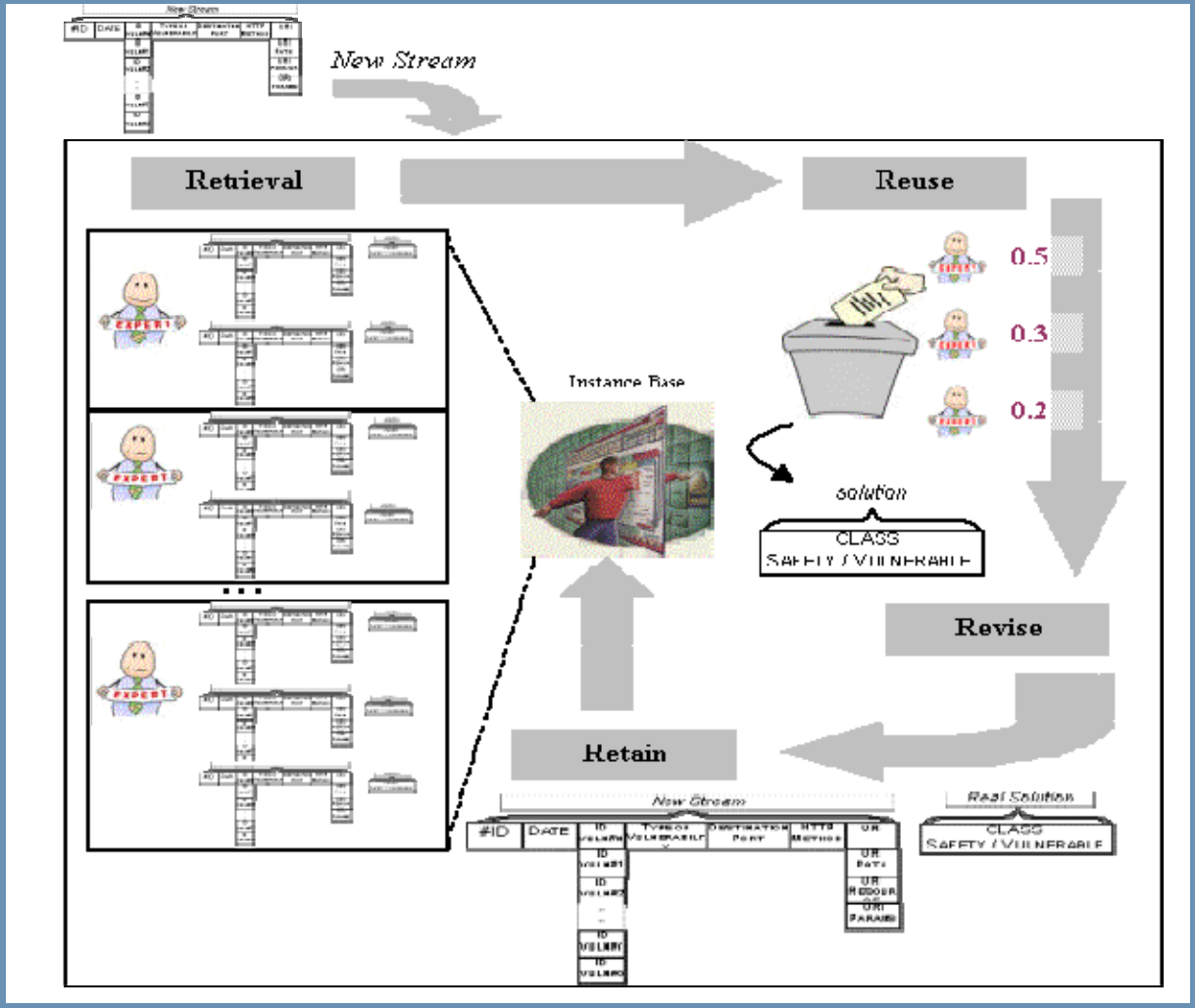
- 256.000 streams
  - 79% tramas vulnerables
  - 21% tramas seguras
- Representación da información de cada stream en HTTPHUNTING



- Medidas de similitud utilizadas
  - URIMatch, PathMatch, ResourceMatch, ParamsMatch
  - URILong, PathLong, ResourceLong

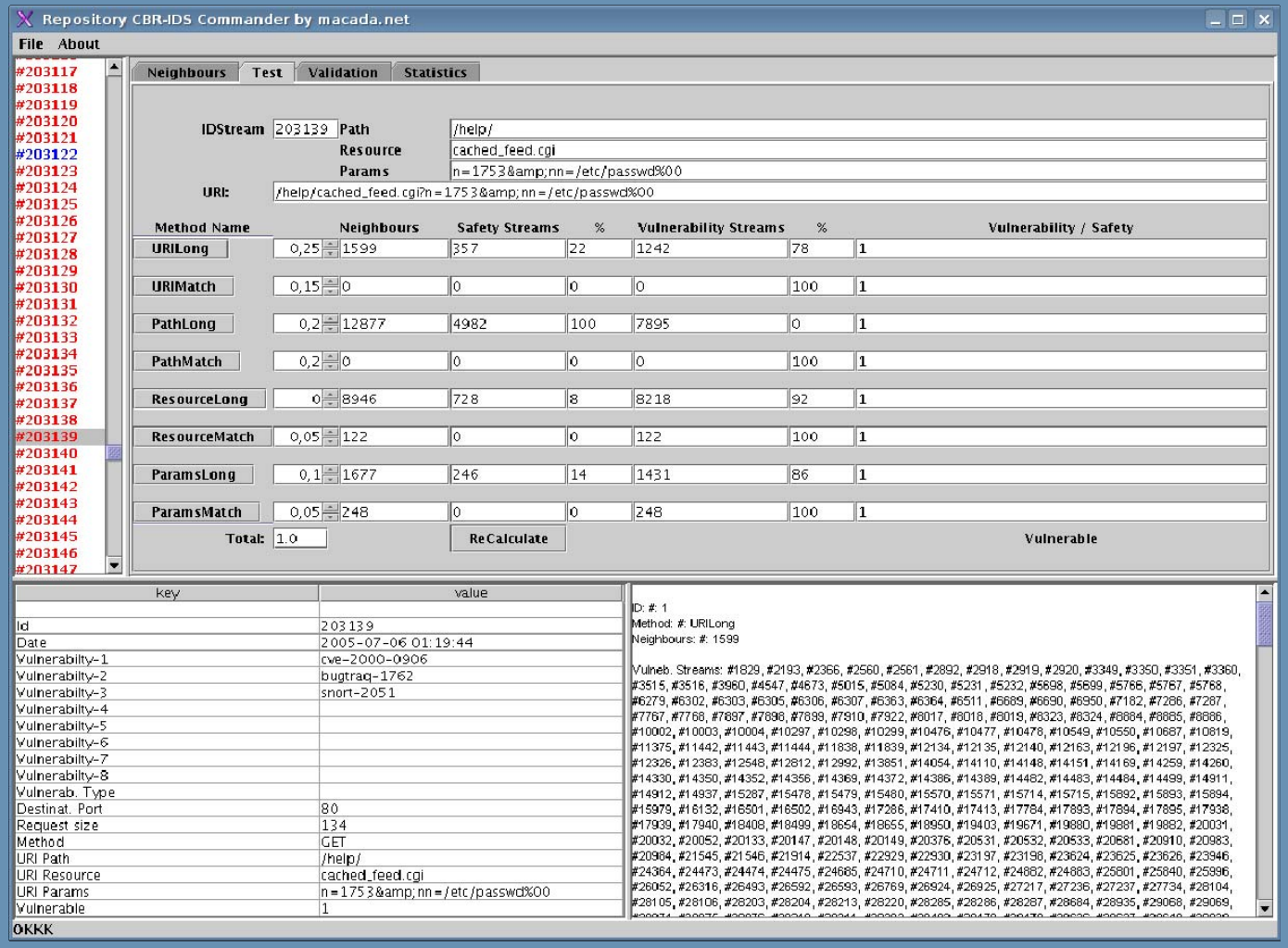
HTTPHUNTING::sistema de detección de intrusos [4/5]

modelo propuesto



HTTPHUNTING::sistema de detección de intrusos [5/5]

la herramienta HTTPHUNTING



GENECBR::análisis de datos de microarrays [1/6]

introducción

- Todos los organismos conocidos están formados por células:
  - Simples: levadura (1 sola célula)
  - Complejos: humanos (trillones de células)
- En el núcleo de cada célula se encuentra el ADN (ácido desoxirribonucleico):
  - 3% genes encargados de la codificación de los distintos cromosomas
  - 97% son secuencias de ADN no-codificante
- Un gen es un segmento de ADN que contiene la secuencia de codificación precisa para cada proteína:
  - Las proteínas determinan el aspecto, metabolismo, conducta, resistencia a infecciones y enfermedades (...) de los organismos
  - En el ser humano se estima que existen entre 30.000 y 45.000 genes
- Prácticamente todas las células de un organismo tienen los mismos genes, pero su expresión genética puede variar por diversos motivos
- El estudio de la variación genética proporciona nuevas fuentes de información para la identificación y el control de enfermedades

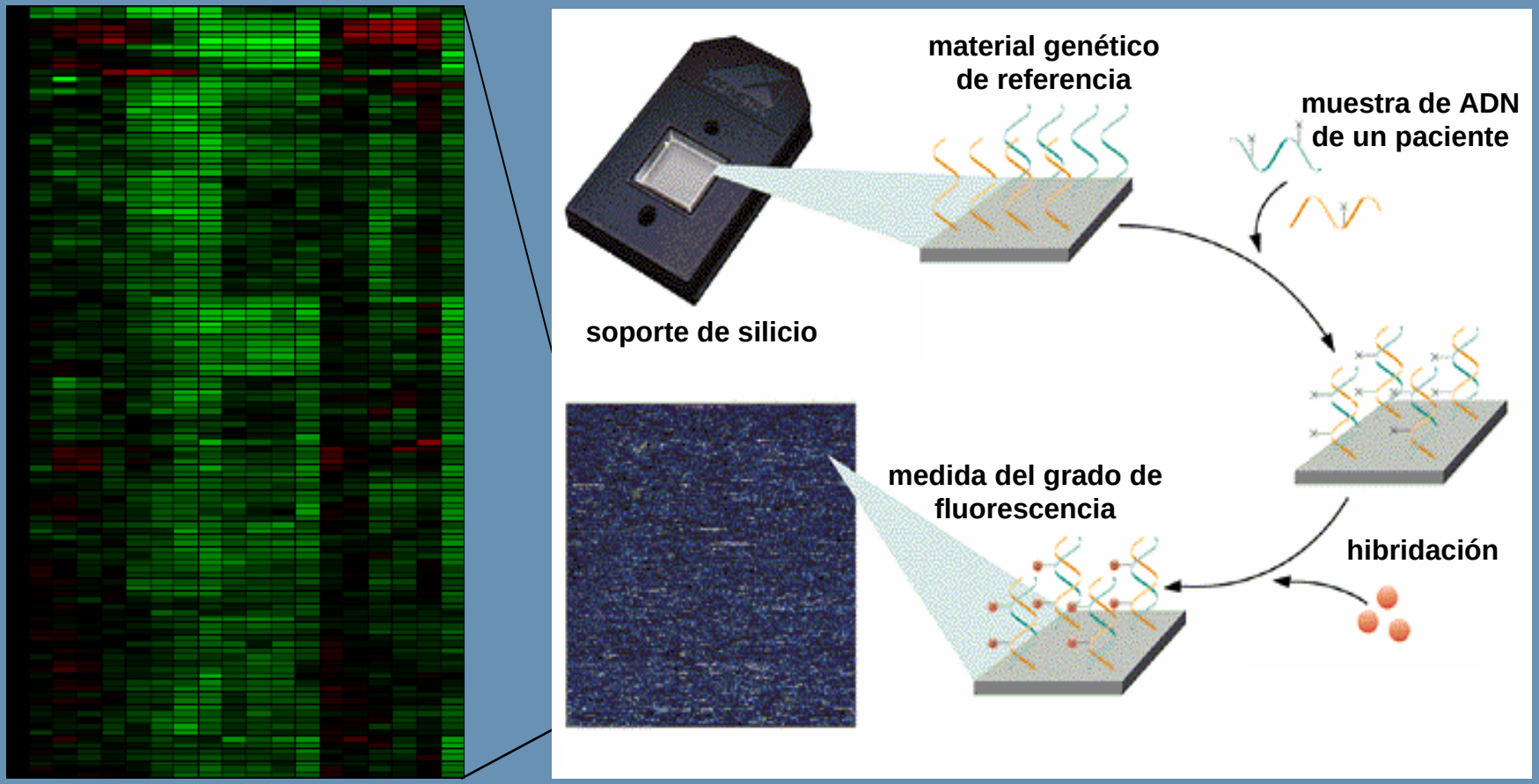
GENECBR::análisis de datos de microarrays [2/6]

microarrays de ADN

- Características:
  - Matriz bidimensional de material genético que permite la automatización simultánea de miles de ensayos
  - Permiten obtener una visión "global" a nivel genético de las células:
    - de diferentes individuos
    - de un mismo individuo en diferentes intervalos de tiempo
    - de diferentes tejidos de un mismo individuo
  - Capaces de medir la expresión genética de decenas de miles de genes en un solo experimento
  - Posibilitan el análisis de funciones e interacciones genéticas complejas a escala global
- Funcionamiento:
  - Soporte de silicio con fragmentos de material genético teñido, correspondientes a aproximadamente 40.000 genes (EST, *Expressed Sequence Tags*)
  - Hibridación con material genético procedente de un individuo concreto
  - Un escáner mide el grado de fluorescencia de cada celda en el microarray

GENECBR::análisis de datos de microarrays [3/6]

human genome U133A GeneChip (Affymetrix)



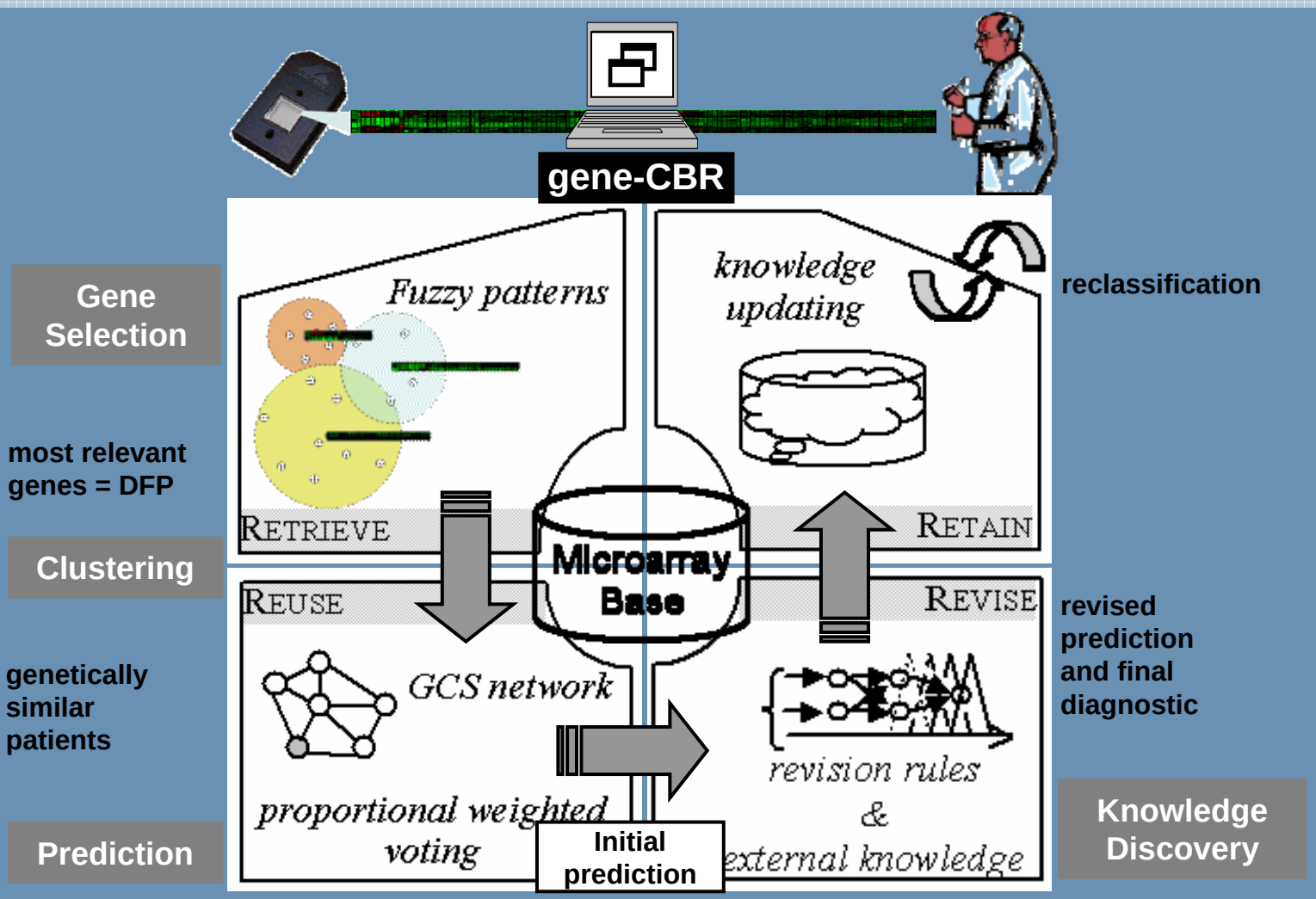
GENECBR::análisis de datos de microarrays [4/6]

bioinformática::áreas de investigación actual

- Tres campos de investigación principales:
  - Selección de genes (≡ selección de características en IA):
    - determinar qué genes son relevantes para identificar una determinada patología o para diferenciar entre varias patologías
  - Predicción (≡ clasificación supervisada en IA)
    - a la vista de los datos de un experimento, identificar si se trata de una persona sana o enferma; discriminar el tipo de patología para aplicar una terapia correcta
  - Clustering (≡ clasificación no supervisada o agrupamiento en IA)
    - descubrimiento de nuevos tipos de patología que no se ajustan suficientemente a una patología conocida
- Tres áreas de investigación paralela:
  - Visualización adecuada de experimentos y resultados
  - Descubrimiento de nuevo conocimiento biológico (interacciones genéticas, ...)
  - Análisis de bajo nivel (tratamiento de imágenes, correcciones de nivel, normalización)
- El análisis de datos procedentes de microarrays presenta importantes retos:
  - Existencia de muchas variables (decenas de miles de genes) y escaso número de observaciones (cientos de experimentos) V.S. *data mining*
  - Gran probabilidad de obtención de falsos positivos con técnicas clásicas de IA

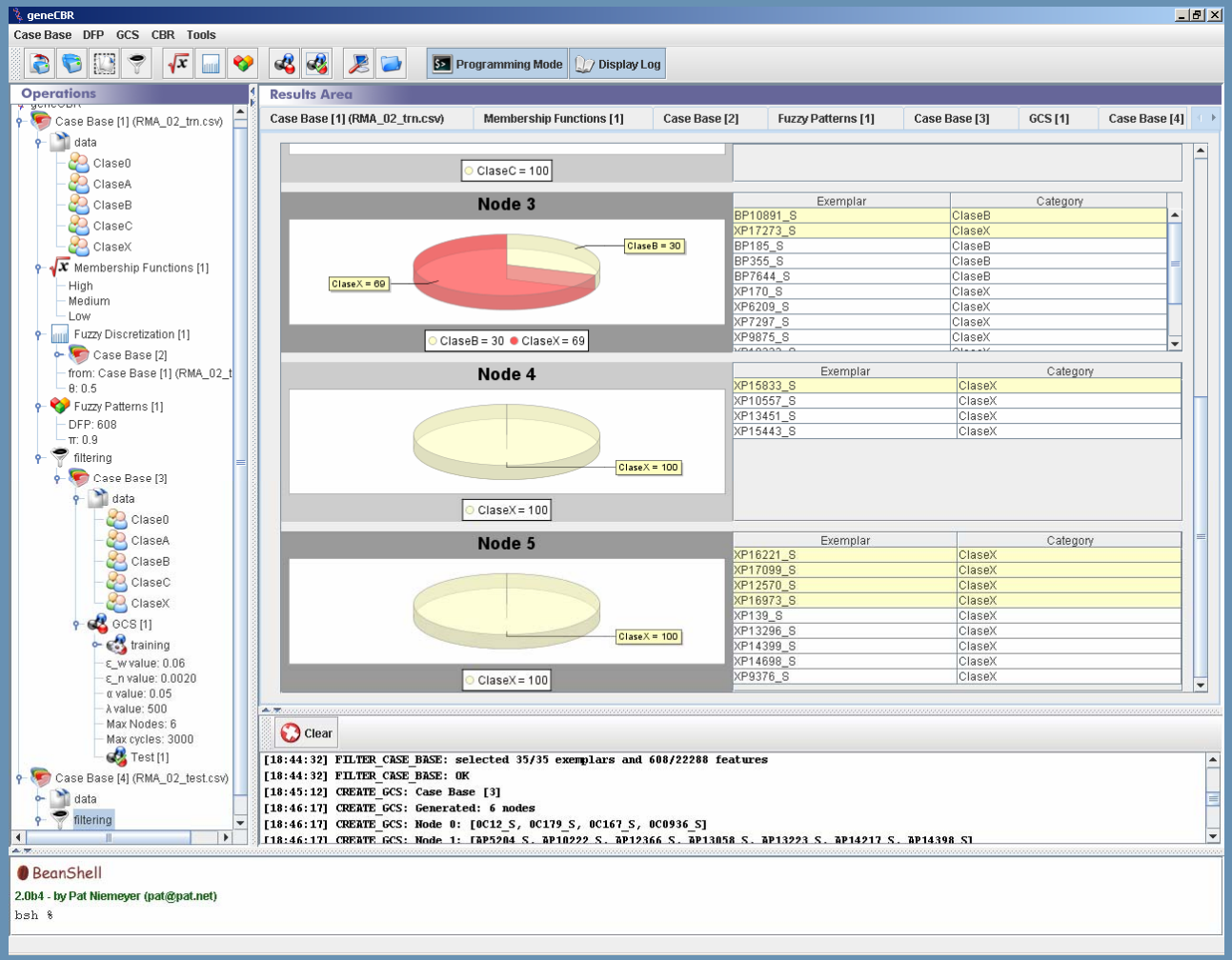
GENECBR::análisis de datos de microarrays [5/6]

arquitectura de GENECBR



GENECBR::análisis de datos de microarrays [6/6]

la herramienta GENECBR



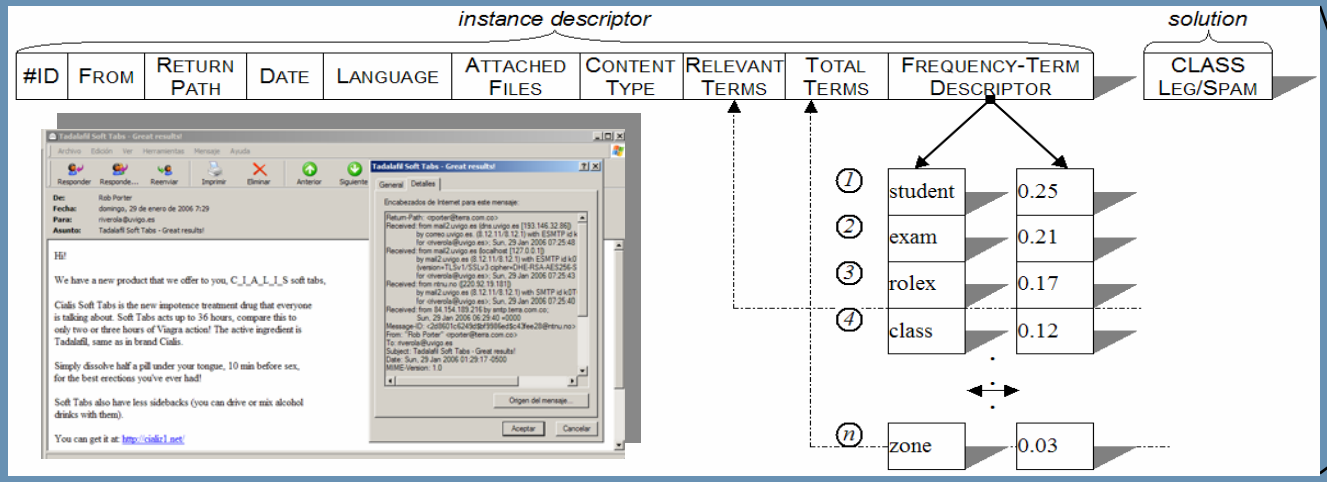
SPAMHUNTING::filtrado de correo spam [1/5]

el problema del correo spam

- Objetivo:
  - Aplicación práctica de los sistemas IBR (*Instance-Based Reasoning*) al problema de la detección de correo basura
- Corpus disponibles:
  - ling spam: 2.869 (t) / 481 (s)
  - junk-email: 2.236 (t) / 1796 (s)
  - bruceg: 171.706 (s)
  - divmov: 1.247 (s)
  - spamassasin: 9.354 (s)
- Generación de una base de instancias unificada (formato xml) y un mecanismo de acceso eficiente (*xml query*)
- Necesidad de corpus propio:
  - Actualmente más de 30.000 mensajes
- Principales retos:
  - Compresión/descompresión de texto (≡ 190.000 mensajes)
  - Análisis de características relevantes ⇒ definición de una instancia en un sistema IBR
  - Necesidad de operación en “tiempo real”

SPAMHUNTING::filtrado de correo spam [2/5]

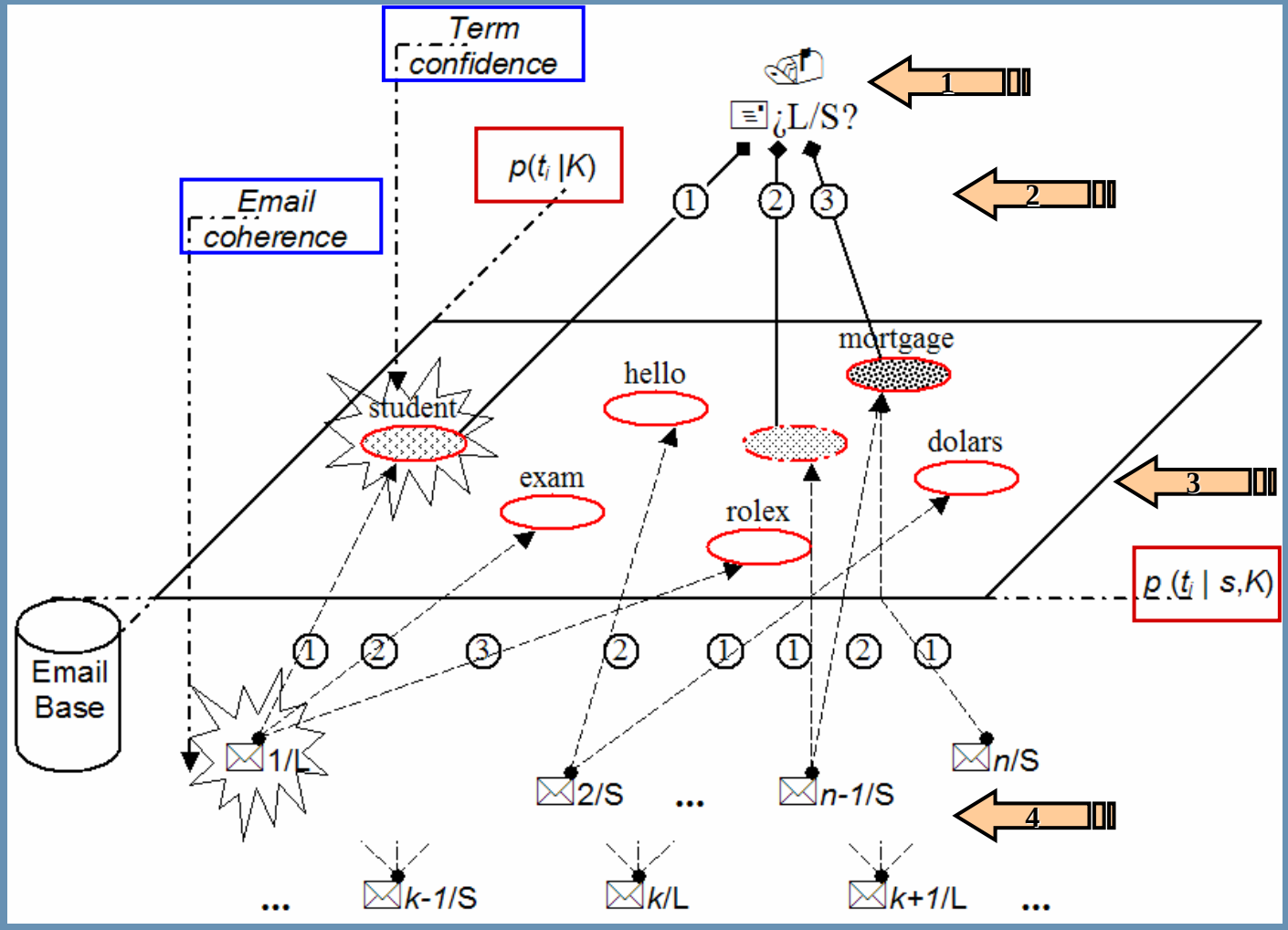
representación de los e-mails



Variable	Type	Description
#ID	Integer	Unique descriptor of the message
From	String	Source mailbox
Return Path	String	Indicates the address that the message will be returned to if one chose to reply
Date	Date	Date in which the message was sent
Language	String	Particular tongue of the message
Attached Files	Integer	Indicates the number of attached files
Content Type	String	MIME type
Relevant Terms	Integer	Number of selected features for cluster the message
Total Terms	Integer	Number of features contained in the message
Terms-Frequency Descriptor	Array of feature-frequency pairs	Storing for each feature a measure of their frequency in the message
Class	Boolean	Class of the message: 0 for Legitimate, 1 for Spam

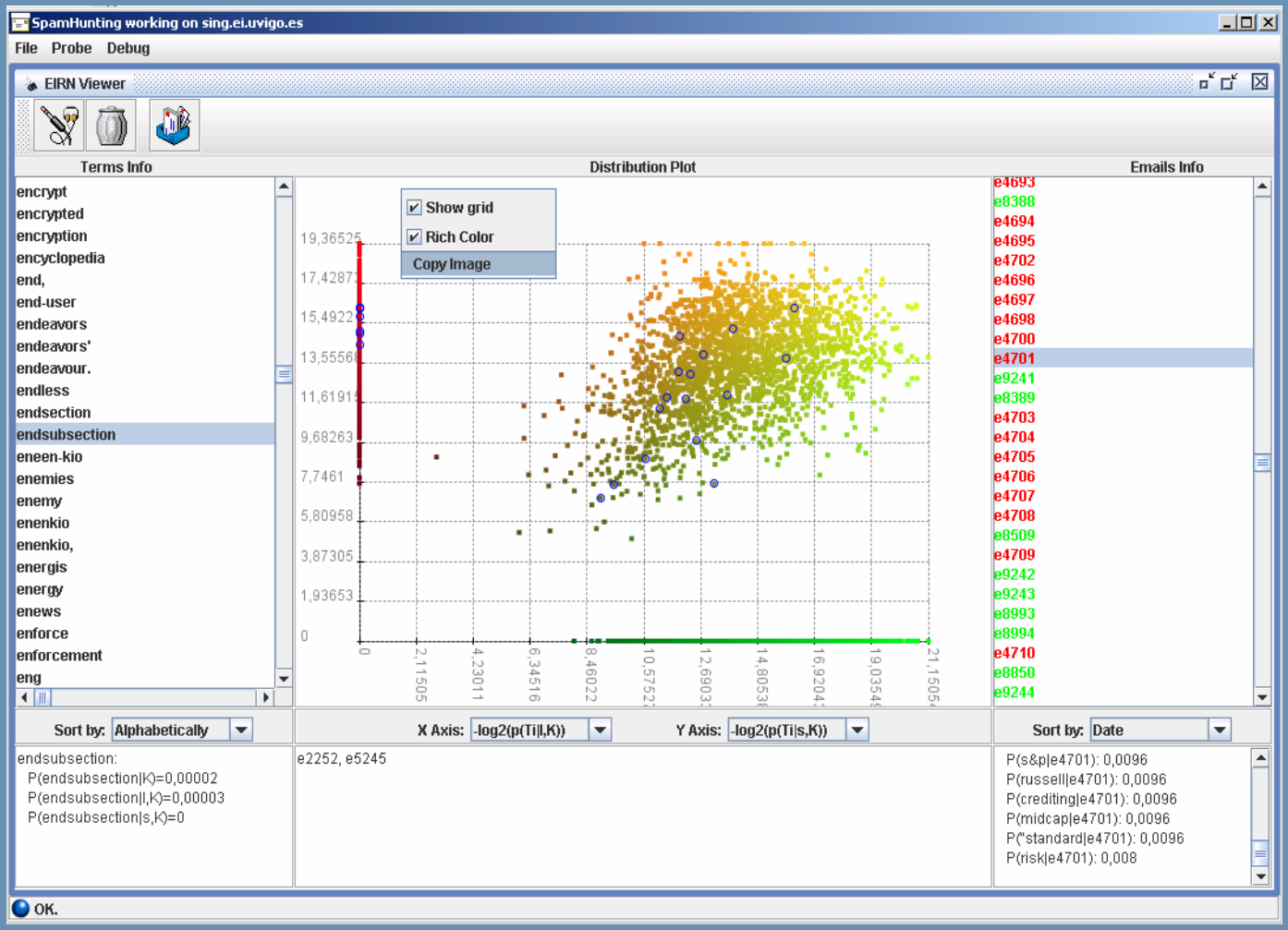
SPAMHUNTING::filtrado de correo spam [3/5]

red EIRN (*Enhaced Instance Retrieval Network*)



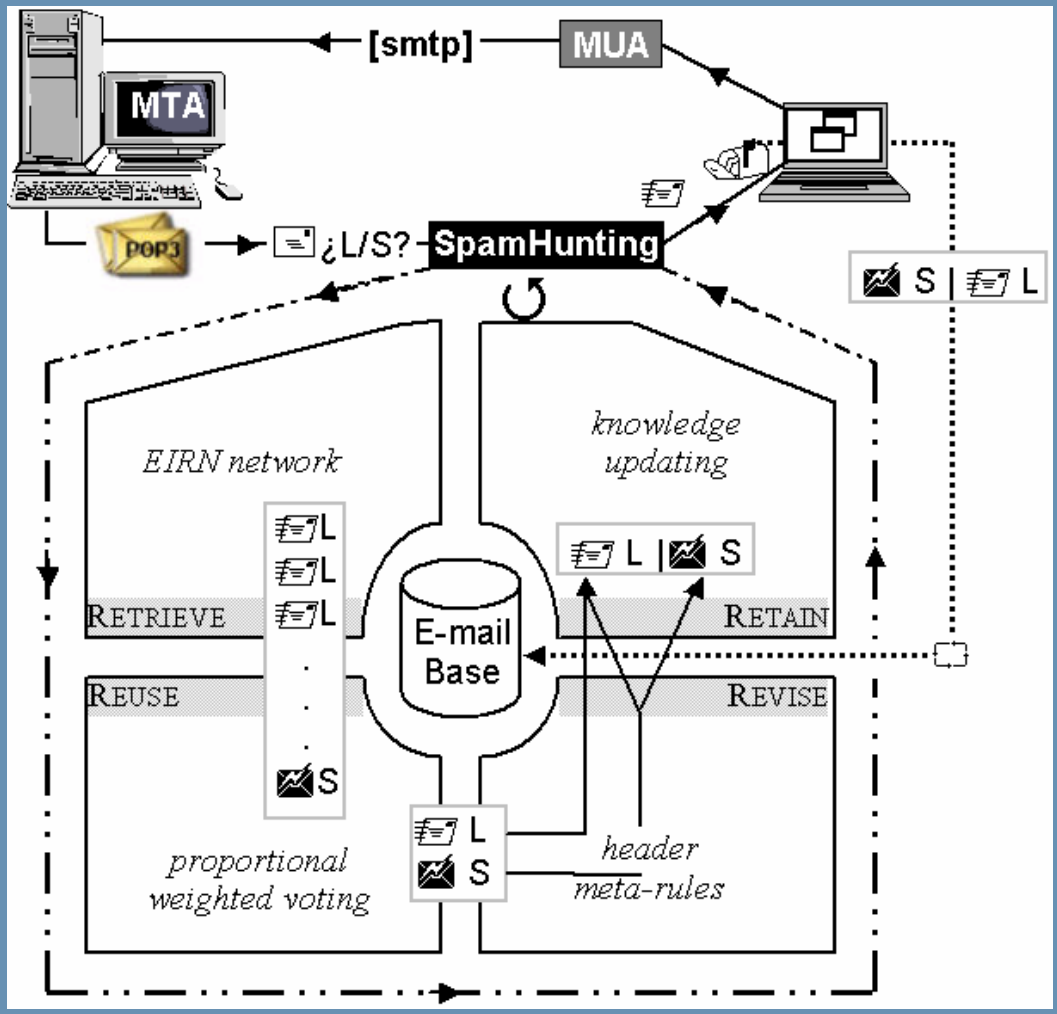
SPAMHUNTING::filtrado de correo spam [4/5]

visor de la red EIRN



SPAMHUNTING::filtrado de correo spam [5/5]

modelo propuesto



cuestiones

Comenta brevemente las fases que componen el ciclo de vida de un sistema CBR

Enumera 3 ventajas acerca de la utilización de sistemas CBR

MICEI

Mestrado em Informática e Curso de Especialização em Informática

Cased-Based Reasoning (CBR) systems: introducción y aplicaciones prácticas

Florentino Fernández Riverola

Braga, 11 de mayo de 2007

Escuela Superior de Ingeniería Informática

Área de Lenguajes y Sistemas Informáticos
Departamento de Informática

Universidade de Vigo