Projecte Final de Carrera
Àrea Compiladors

GestNews:

Servei de Cerca Automatitzada de Notícies

Estudiant: Carlos E. García Gómez
Consultor: Jordi Ferrer Duran
15 de Gener del 2012

Índex

- Problemàtica Plantejada.
- Arquitectura de la solució.
- Components de l'aplicació.
- Component GUAP.
- Interfície WEB.
- Procés sincronització de noticies.
- Subprocés tractament de documents RSS.
- Problemes trobats amb l'extracció de noticies.
- Subprocés extracció de noticia.
- Regles de Selecció.
- Execució de les Regles de Selecció.
- Generació d'avís.
- Seguiment de la Programació.
- Conclusions

Problemàtica Plantejada

- Gran quantitat d'informació a Internet.
- Pèrdua de temps en cercar la informació realment d'interès.

Solució que es demana:

- La creació d'una aplicació web que llegeixi dels servidors de contingut.
- Seleccioni les noticies potencialment d'interès per a l'usuari.
- Li avisi de l'existència d'aquestes noticies.

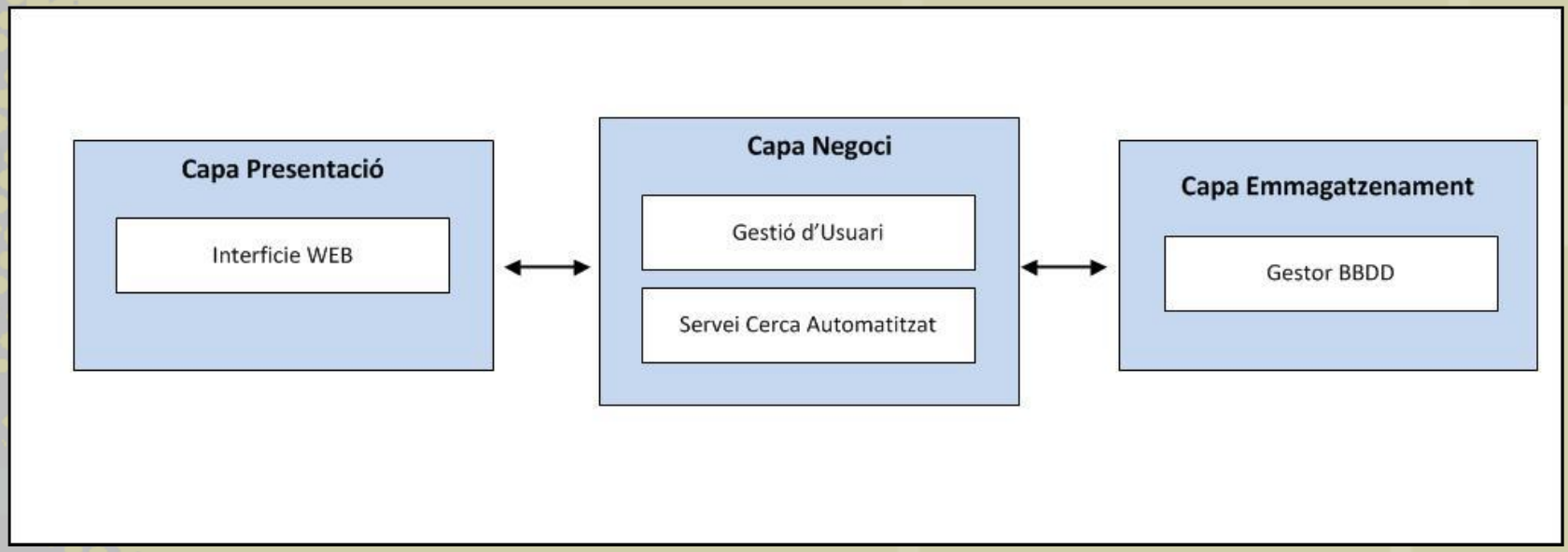
En definitiva:

- L'usuari no ha de perdre temps cercant informació.
- GestNews avisa a l'usuari quan aparegui informació publicada del seu interès.

Arquitectura de la Solució

S'ha triat un disseny d'arquitectura en Capes:

- Capa de Presentació. Interfície Web.
- Capa de Negoci:
  - Subcapa de Gestió d'Usuari i Administració de Preferències (GUAP)
  - Subcapa del Servei de Cerca Automatitzada (SCA).
- Capa d'Emmagatzemament. Base de Dades GestNews.



Components de l'aplicació

A l'aplicació s'identifiquen quatre components:

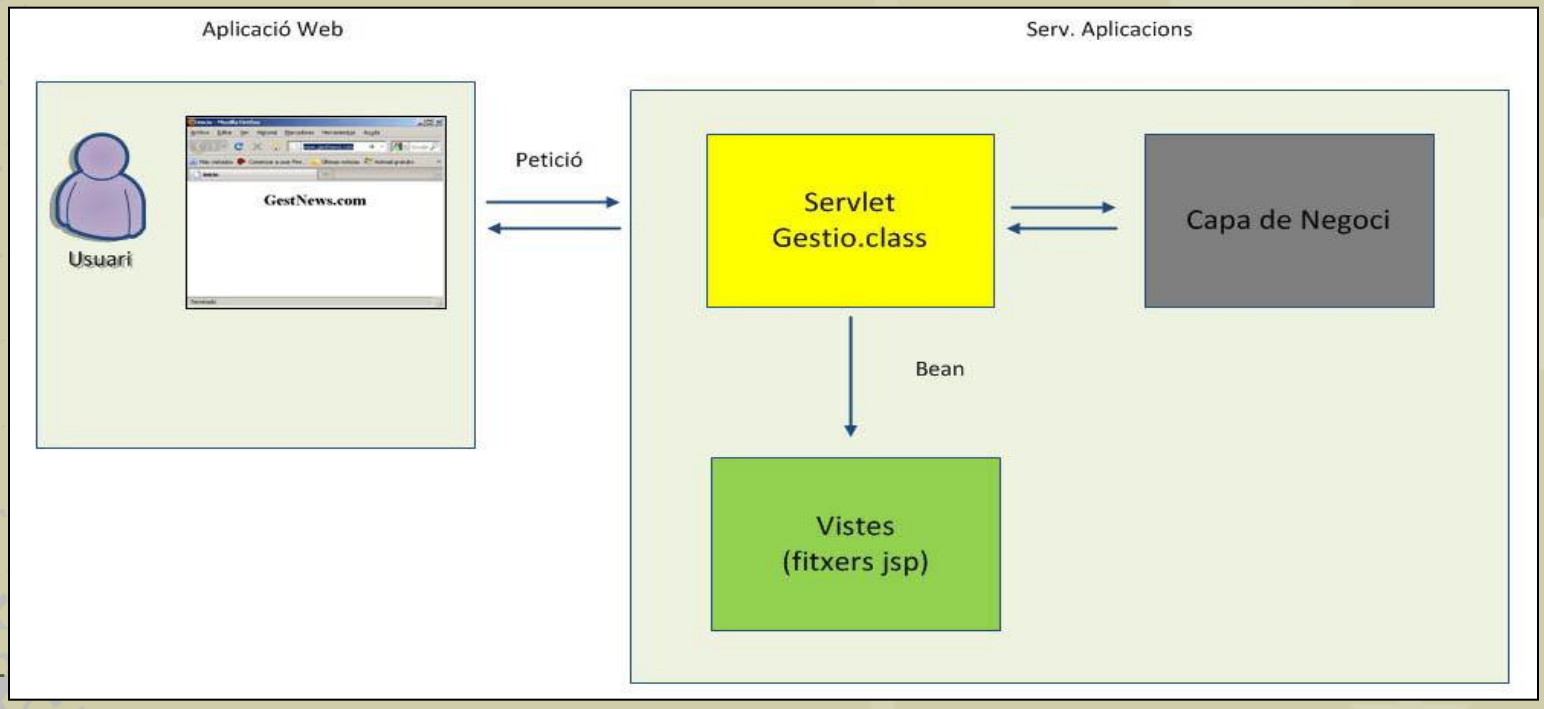
- Gestió d'Usuari i Administració de Preferències (GUAP):
  - Part visible de l'aplicació.
  - Executat des del Servidor d'Aplicacions (jboss)
  - Funcionalitat realitzada per la subcapa que rep el seu nom.
- Servei de Cerca Automatitzat (SCA):
  - Procés de Sincronització de Noticies.
  - Procés d'Execució de les Regles de Selecció.
  - Procés de generació i enviament d'avisos.
- Gestor de Base de Dades.
  - Persistència de l'aplicació.
- Canals d'informació.
  - Servidors de continguts: Canals RSS.

Component GUAP

- Executat al servidor d'aplicacions i controlat per una Servlet.
- Funcionalitat implementada a la subcapa de Gestió d'Usuaris i Administració de Preferències (GUAP).
- Funcions que realitza:
  - Gestionar als usuaris del sistema.
  - Creació i Administració de la subscripció als canals de contingut.
  - Creació i Administració de les regles de selecció.
  - Anàlisi i Compilació de les regles de selecció.
- Resta de processo que requereix de més temps d'execució són realitzats per component SCA en background.

Interfície Web

- Representa la part visible del component GUAP.
- Utilitza el patró de disseny MVC on el controlador és la servlet i les vistes pàgines JSPs.
- La Servlet envia objectes Bean a les Vistes per personalitzar el resultat.
- Les Vistes utilitzen JSLT per accedir al contingut dels objectes Bean.



Procés Sincronització Noticies (SCA)

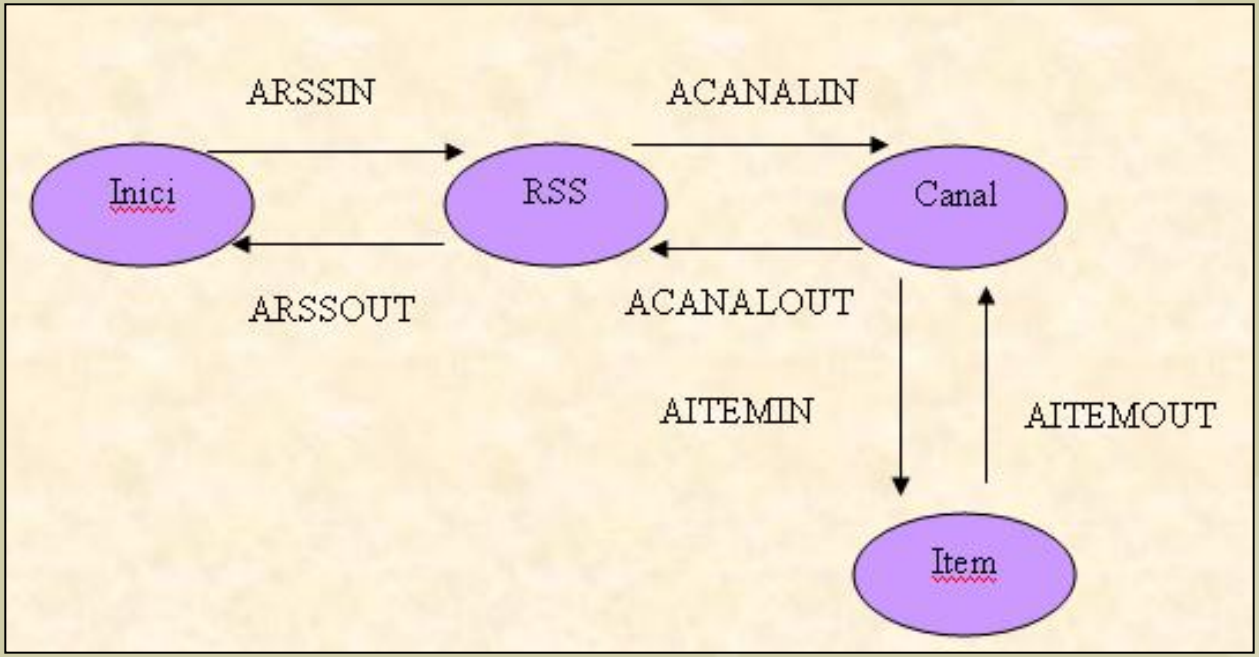
- Els Servidors de Contingut publiquen un document en format RSS amb l'objectiu de facilitar la distribució dels seus continguts als lectors RSS o altres servidors web.
- El procés de sincronització accedeix als servidors de continguts per conèixer les noves noticies i descarregar-les.

Accions realitzades:

- Descàrrega document RSS dels canals subscrits per l'usuari.
- Analitzar document RSS mitjançant un analitzador SAX i un Autòmat d'estats Finits.
- Extracció dels diferents ítems d'informació.
- Emmagatzemament de la informació a la Base de Dades.

Subprocés tractament documents RSS

- La classe **ParserRSS** implementa un analitzador SAX.
- L'analitzador utilitza un autòmat d'estats finits per poder conèixer l'element que està analitzat.
- El mètode **startElement** de l'analitzador s'activa quan identifica un nou element i avisa a l'autòmat perquè canvi d'esta.
- El mètode **characters** de l'analitzador consulta l'estat de l'autòmat per emmagatzemar la informació que rep a la propietat adient.



Problemes trobats en l'extracció

- El format RSS no inclou l'estructura de la noticia publicada.
- Extracció del contingut d'una noticia al document HTML publicat.
- Diferents tipus de codificació a les noticies publicades.
- Documents HTML mal formats.

Solució Aplicada:

- Divisió dels canals utilitzats pel sistema en:
  - Canals Coneguts: Es coneix l'adreça XPATH del contingut de la noticia al document HTML publicat.
  - Canals No Coneguts: No es coneix l'estructura del document HTML publicat.

Canals Coneguts:

Les regles de selecció de noticies per contingut s'apliquen únicament al contingut de la noticia publicada.

Canals No Coneguts:

Les regles de selecció de noticies per contingut s'apliquen en tot el document publicat. Poden produir falsos positius.

Subprocés extracció noticia

- Als canals coneguts es coneix l'adreça **XPATH** de la noticia al document HTML publicat.
- La noticia es descarregada i s'eliminen els elements innecessaris (scripts, comentaris, urls, etc.)
- La classe ParserNoticiaHTML implementa l'analitzador SAX.
- Utilitza un autòmat d'estats finits per conèixer l'adreça XPATH de l'element analitzat (tag).
- Quan l'element analitzat coincideix amb el XPATH, s'ha trobat el contingut de la noticia i finalitza el procés.
- L'autòmat crea un arbre amb els elements que va analitzant. Aquest determina el seu estat.

Avantatges respecte ús DOM:

- L'estructura d'arbre que es crea, ocupa menys espai.
- En el moment en què es trobat el contingut s'atura el procés. No es crea una estructura de tot el document HTML com faria DOM.

Regles de Selecció

- Una regla de selecció és una expressió lògica que l'usuari introdueix al sistema per determinar el tipus de noticia que li interessa.
- La gramàtica que valida les regles de selecció admet funciones de cerca i operadors lògics (Operadors de l'àlgebra de Boole: &&, ||, !)

Exemples:

- begin(títol, "amèrica")
- begin(títol, "amèrica") && begin(descripció, "amèrica")
- (begin(títol, "amèrica") && begin(descripció, "amèrica")) || contains(noticia, "amèrica")
- contains(noticia, "picasso") && !contains(noticia, "dalí")

- Les regles són analitzades i traduïdes a una expressió postoperacional.
- S'ha utilitzat LEX i CUPs per implementar els analitzadors i el compilador.

Exemples:

- _titol__;_amèrica_;__beg_
- _titol__;_amèrica_;__beg__;__descrip__;_amèrica_;__beg__;__&&_
- _titol__;_amèrica_;__beg__;_descrip_;_amèrica_;__beg__;__&&__;_noticia_;_Amèrica_;__cont__;__||__
- _noticia__;_picasso_;__cont__;__noticia__;_dalí_;__cont__;__!__;__&&_

Execució Regles de Selecció

- Les regles de selecció són executades per una Màquina Virtual.
- La MV determina si la regla es compleix per una determinada noticia.

Implementació:

- La MV està implementada a la classe VMachine.
- Amb l'expressió postoperacional es crea un array amb els seus components.
- L'array es tractat:
  - Operands són emmagatzemats a una pila auxiliar.
  - Operadors són executats traient els operands necessaris de la pila i retornant a la pila el seu resultat.
- El resultat de l'execució és el valor final que hi ha a la pila.
- Si el resultat és positiu és crea un avís.

Generació de l'avís

- L'usuari pot determinar tres tipus d'avisos:
  - Succés en format ICALENDAR.
  - Registre successos en format ICALENDAR.
  - Registre successos en format ODT

Procés Generació d'avís:

- La generació de l'avís es realitza en dues parts:
  - Crear un document xml amb la informació de l'avís.
  - Aplicar una plantilla xslt al document xml crear per generar el tipus d'avís triat per l'usuari.
  - Llançar subprocés d'enviament de missatge.

Enviament de l'avís:

- Per a l'enviament de l'avís s'utilitza la llibreria JavaMail:
  - Es crea el missatge i se li adjunta el fitxer generat.
  - S'envia a l'adreça de l'usuari utilitzant un servidor smtp establert a la configuració de l'aplicació.

Seguiment de la Planificació

- Seguiment de la Planificació referent a les fites externes (lliuraments de les PACS):
  - Primera PAC: Les tasques es van realitzar sense massa problemes.
  - Segona PAC: Es van detectar diversos problemes i es van posposar per després del lliurament. La documentació va quedar un poc endarrerida.
  - Tercera PAC: La solució del problemes comentats requereix de la creació d'una nova tasca. Es va necessitar replanificar la programació i augmentar el recursos humans per poder finalitzar el projecte amb garanties.
  - Lliurament Final: Amb la nova planificació es va aconseguir finalitzar les tasques marcades pel projecte.

Conclusions

- La publicació de documents RSS és un bon mecanismes per donar a conèixer les noticies dels Servidors de Continguts.
- Importància dels analitzadors SAX per extraure continguts de documents de documents XML.
- Necessitat d'un estàndard en la publicació de noticies per facilitar la creació d'eines per a l'extracció d'informació.
- L'ús d'eines com GestNews podem facilitar a l'usuari molt les coses: Subscripció a diversos canals d'estrenes de pel·lícules. Establir regla de selecció per conèixer quan s'estrena una pel·lícula que apareix un determinat actor. Moltes possibilitats.
- La implementació de GestNews m'ha permès ampliar coneixements en l'àrea dels compiladors:
  - Possibilitats dels autòmats finits.
  - Anàlisis i traducció d'expressions amb JLEX i CUPS.
  - Generació de documents XML amb JAXB.
  - Els analitzadors SAX.
  - Transformació de documents XML amb xslt.

Breu demostració de l'aplicació

- Posada en funcionament del servidor d'aplicacions.
- Accés al sistema.
- Subscripció a dos canals:
  - Canal conegut: tv3.cat.
  - Canal no conegut: ELPAIS.com
- Activació del component CSA per veure com descarrega les noticies.
- Alta de tres regles de selecció:
  - Selecció per descripció:
  - Selecció per contingut en canal conegut.
  - Selecció per contingut en canal no conegut.