EPS
Escola Politècnica Superior
UdG

Projecte/Treball Fi de Carrera

Estudi: Enginyeria Informàtica. Pla 1997

Títol: ARI: Agent Recaptador d'Informació. Desenvolupament d'una aplicació que reculli inforació de portals webs dedicats a la gestió de premsa.

Document: Resum del Projecte

Alumne: Alejandra Gómez Pérez

Director/Tutor: Gustavo Patow
Departament: Informàtica i Matemàtica Aplicada
Àrea: LSI

Convocatòria (mes/any): 09/07

1 Introducció i Objectius

Avui en dia ens trobem davant d'una gran necessitat d'informació. Això és a causa de la gran explosió tecnològica que ha viscut el món en els últims anys. De fet, no fa gaire temps, la gent podia viure sense haver de tenir un mòbil o sense una connexió d'Internet d'alta velocitat... Aquestes noves necessitats, provocades per la societat a la que vivim, fan que hagin sorgit nous horitzons empresarials.

Així doncs, enfocant la visió cap a Internet es pot observar com han sorgit nous portals d'informació on els grans mitjans de comunicació espanyols han reflectit el que fins al moment feien en paper, radio o televisió. Fent un cop d'ull a aquests mitjans, han evolucionat cap l'apropament en tot moment de la informació més actual a l'usuari. En un primer moment, aquest portals es limitaven a transmetre el que ja havien donat prèviament en paper o per la ràdio o televisió, però clar, si quelcom aporta aquesta nova plataforma és la immensa capacitat per tenir la informació actualitzada al moment que es vulgui.

D'aquesta manera sorgeix el recull de notícies en RSS (Really Simple Syndication –RSS 2.0) que va atorgar als *webmasters* una eina per a publicar continguts d'una manera ràpida i senzilla. Aquest format per distribuir el contingut dels portals pels seus clients habituals va ser un gran esdeveniment, que va fer que les empreses ho veiessin com una nova font de consum i van passar de facilitar-los gratuïtament a fer que els usuaris s'haguessin de subscriure (i per tant, pagar quelcom) per poder rebre les notícies a les seves bústies de correu electrònic.

Veient aquesta evolució de la informació a Internet, va sorgir la idea d'aquest projecte, un motor de cerca orientat a la recaptació de notícies dispersades per les diferents pàgines web dels gran mitjans de comunicació espanyols, per tal que es pogués obtenir informació sobre "descriptors contractats"[1] pels usuaris del portal facilitat per aquest PFC.

Com a idea sobre el paper va semblar suficientment bona com per poder desenvolupar un Projecte Final de Carrera interessant i amb el gruix necessari d'investigació i innovació com per correspondre a una enginyeria superior.

En el seu origen, fa ja any i mig, es va començar a desenvolupar en una empresa on estava realitzant unes pràctiques de becària. Però degut a canvis a la legislació (referents a la Llei de Protecció de Dades i de la Propietat Intel·lectual), l'empresa en qüestió va decidir deixar de banda el projecte perquè no els resultava viable pels seus recursos de petita empresa. A pesar d'això, van decidir cedir el codi per poder continuar el desenvolupament del present Projecte Final de Carrera (PFC). Així doncs, encara que les expectatives empresarials amb les que va néixer el projecte no eren viables pel desenvolupament dins del marc de l'empresa en que s'estaven fent les pràctiques, es va canviar la visió original per una visió més acadèmica i d'investigació sobre el món dels portals de comunicació i els *spiders*[2] recaptadors d'informació.

Finalment, cal dir que el que es desenvoluparà a continuació és un projecte de caire investigador que vol esbrinar les possibilitats de recaptació d'informació que reporta un món tan extens i tant ple de possibilitats com és Internet, per portar a terme aquest PFC a continuació es desglossaran els objectius que es volen assolir i el corresponent anàlisi i disseny que el faran possible. L'estructura de la documentació que es podrà veure, vindrà donada per la tècnica Extreme Programming, que donarà un caire d'evolució permanent al document.

2 Objectius

Els objectius a assolir al final del desenvolupament d'aquest PFC són molt clars: l'obtenció d'un sèrie de notícies extretes de diferents pàgines web, dels diferents portals dels mitjans de comunicació, seguint uns paràmetres preestablerts.

Per poder arribar a desenvolupar la idea cal definir les fites a les quals es vol arribar. Tal com s'esmenta, l'objectiu principal està clar, obtenir notícies. Però, com arribar fins a elles? Com seleccionar les que es volen guardar al sistema? Per això calen uns passos previs d'investigació sobre el mode de procedir.

El primer objectiu és l'anàlisi de les necessitats que es volen cobrir per un hipotètic client de l'aplicació. D'aquesta manera s'estableix que es necessita un portal des d'on aquests clients es puguin registrar i accedir a una zona reservada per escollir sobre un llistat predeterminat uns mitjans a on es vol fer el seguiment i una

[1] **Descriptors contractats:** expressió simbòlica, ja que al ser un PFC experimental no es realitza cap relació contractual amb els usuaris de la web dissenyada per aquest projecte.
2 Spider: "aranya cercadora", motor de cerca on-line.

altra secció a on es pot configurar el llistat de *descriptors*[3] que serviran per fer la selecció de notícies als mitjans escollits.

El segon objectiu és a l'àmbit algorítmic. Cal obtenir una metodologia de treball que permeti l'obtenció de la notícia. Per aconseguir això s'ha estudiat el següent: la llibreria *cUrl*[4], la utilització de patrons de cerca, i l'emmagatzemament a Base de Dades.

Aquest estudi permetrà obtenir de les pàgines web, mitjançant les funcionalitats de la llibreria cUrl, el contingut per poder fer el posterior anàlisi per poder determinar l'optimitat i decidir si cal o no guardar-ho al sistema. Per poder fer un bon anàlisi, s'aplicaran els conceptes extrets de la investigació sobre el món dels patrons i les expressions regulars[5], per finalment emmagatzemar-ho tot a la Base de Dades escollida.

Així doncs, els objectius a l'àmbit de la programació passen per tres etapes: descarregar les pàgines web necessàries, que es farà mitjançant les eines que proporciona la llibreria esmentada (cUrl). Amb aquesta eina es facilita la feina d'investigació d'obtenció de la informació, base des d'on es localitzaran els enllaços a les notícies que contenen les pàgines principals de cada portal.

Un cop es té descarregat el contingut de la pàgines, el següent objectiu és l'anàlisi. Hi ha tres tipus d'anàlisis: Primer, obtenir tots els enllaços que corresponen a notícies, segon, filtrar els descriptors que es tenen a la Base de Dades per decidir si s'ha de guardar la notícia en qüestió o no i tercer i últim, un cop s'ha decidit si es necessita la notícia, analitzar la seva estructura interna, per guardar només les parts preestablertes (titular, entradeta i cos[6]) de la notícia.

Com a últim objectiu es troba la Base de Dades. Es necessari una estructura organitzada que permeti tenir tot totalment estructurat per poder obtenir les notícies i saber per quin o quins descriptors s'han escollit, de quin mitjà són o a quin o quins clients pertanyen.

3 Primera i Segona Versió del desenvolupament

Tant la primera com la segona versió tenen molt en comú. A continuació es farà un breu resum de totes les coincidències i diferencies que es troben al llarg de les dues versions i el perquè d'una i l'altre.

Per començar s'especifiquen els requisits que ha de complir el sistema en general:

- *Obtenir notícies*: aquest és el fi bàsic del projecte, el motiu pel qual es desenvolupa i es vol investigar la Xarxa, per esbrinar la dificultat que comporta el voler donar aquest servei a l'usuari.
- *Gestionar els mitjans de comunicació*: per poder portar una gestió i ordre de les notícies que obté cada usuari, cal facilitar-li la gestió dels mitjans de comunicació que té al seu abast per fer els seguiments de premsa que desitgi.
- *Gestionar els descriptors:* així com succeeix amb la gestió anterior, aquesta és una conseqüència lògica del primer punt, aquí esmentat, ja que per poder fer seguiments de premsa calen uns descriptors, que seran els ítems a localitzar dins de les notícies per poder filtrar-les com a bones per a l'usuari.

Un cop es tenen clars els requisits, es passarà a dividir en subsistemes per a concretar amb més detall quins processos ha d'abordar cada un.

Descriptors

A aquest subsistema es gestiona tot el relacionat amb els ítems o descriptors que l'usuari vol utilitzar per fer els seguiments de premsa. Així, les dades s'emmagatzemen a la Base de Dades fent de la gestió de l'àrea un recurs senzill i pràctic per l'ús de l'aplicació en qualsevol moment que l'usuari desitgi.

Mitjans de Comunicació

Tal com succeeix al subsistema de Descriptors, aquest subsistema també està encarat per a que s'utilitzi per orientar la cerca per la Xarxa. L'usuari administra els mitjans de comunicació, dels quals vol extreure les notícies que li interessen.

D'aquesta manera, a la primera versió del desenvolupament, el sistema té prèviament a establert a la seva algorísmica els patrons necessaris per extreure la informació referent a tipologia d'enllaços, titular, entradeta

[3] **Descriptor:** paraula o paraules que es cercaran dins de les notícies i que la seva ocurrència determinarà si aquesta s'emmagatzema al sistema o no.
[4] **cUrl:** llibreria PHP que permet la descàrrega del contingut d'una plana web a memòria.
[5] **Expressions regulars i patrons:** veure la secció 3.2.3 de la memòria adjunta, a on s'expliquen amb detall.
[6] **Titular, entradeta i cos:** veure la secció 3.1.1 de la memòria adjunta, on s'explica el detall de cada part de la notícia

i cos, i buscarà per les pàgines d'aquests mitjans les notícies que coincideixin amb els descriptors que l'usuari a configurat amb antelació.

En canvi, a la segona versió, els patrons no estan encapsulats estàticament a la programació de l'aranya cercadora, sinó que per cada mitjà del qual es vol recuperar notícies, el sistema recuperarà de la Base de Dades els patrons que hi tingui estipulats a la taula pertinent.

Notícies

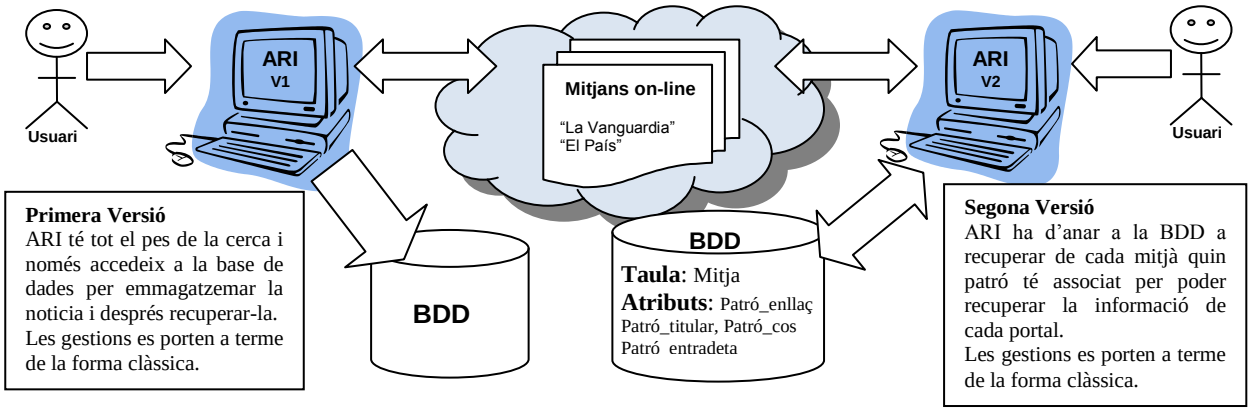
El subsistema de notícies és el més complex dels tres, ja que és l'encarregat de recopilar tota la informació de la Xarxa, analitzar-la i emmagatzemar-la (si cal) a la Base de Dades. D'aquesta manera, l'usuari pot trobar a l'àrea reservada de l'aplicació tot el recull de notícies que el sistema a recaptat per a ell, seguint els paràmetres que ell mateix a establert.

Cal tenir en compte que al subsistema de notícies hi ha dos àrees independents on una està dirigida pel sistema, essent l'encarregat de la cerca per la xarxa i exercint d'*spider* per les diferents pàgines dels mitjans de comunicació i l'altre àrea del subsistema vindrà regida per l'usuari, on podrà visualitzar de diverses maneres tota la informació que el sistema recopili per a ell.

Tal com s'ha descrit al subsistema anterior, la diferència entre la forma en que la primera versió i la segona versió del desenvolupament recapten les notícies dels mitjans de comunicació, radica en l'algorísmica que utilitzen darrera. En una primera versió el sistema de patrons està encapsulat al codi, en canvi, a la segona, la taula de la Base de Dades que emmagatzema la informació sobre el mitjà de comunicació també té quatre camps nous que guarden els patrons corresponents a les zones a localitzar dins de la informació que es descarrega de cada pàgina web.

A continuació es mostra el diagrama de context que reflecteix la manera en que les dues versions actuen per arribar a emmagatzemar la informació necessària de la Xarxa.

Diagrama de Context del Sistema (v1. i v.2)



4 Tercera Versió del desenvolupament

Al finalitzar la segona versió del desenvolupament es va detectar com el procés de recaptació de notícies s'havia automatitzat molt més que a la primera versió, que només funcionava pel primer mitjà de comunicació analitzat.

De totes maneres, es van detectat petits errors de captures en mitjans de comunicació que s'havien analitzat al principi del procés. Això va provocar que es tingués que analitzar un altre cop el codi font dels mitjans de comunicació que fallaven, per poder localitzar on radicava l'error de funcionament. Després de fer aquest anàlisi es va observar com els mitjans de comunicació havien anat canviat la seva estructura interna que utilitzen per la programació de les seves pàgines web.

Tot això va provocar que es tornés a dissenyar l'aplicació, (la tercera versió) per poder superar aquest nou inconvenient trobat. Aquesta nova versió del desenvolupament afectarà als mateixos punts crítics que s'han canviat de la primera a la segona versió, ja que els processos que es veuen afectats són els mateixos: l'alta de Mitjans de Comunicació i l'Obtenció de Notícies.

En resum, aquesta nova versió dels requisits ha d'afrontar el sistema, a part del superat de la primera a la segona versió (la NO homogeneïtat de l'entorn web), és la versalitat de codi, és a dir, la possibilitat de que un

mitjà de comunicació canviï la seva estructura de codi font de les seves pàgines web cada una quantitat de temps no conegut.

A continuació es passen a detallar els dos processos diferencials d'aquesta versió:

Donar d'alta un Mitjà de Comunicació.

En aquesta versió es troba que l'administrador del sistema ha de donar d'alta el mitjà de comunicació a la Base de Dades per a que l'aplicació pugui fer la cerca, al igual que succeeix amb la primera versió i la segona. D'igual manera, aquest procés el segueix tinent que portar a terme l'administrador del sistema, que haurà d'incloure a la Base de Dades la informació necessària del nou mitjà de comunicació que es vulgui donar d'alta: nom, descripció, enllaç i si està actiu o no. La diferència radica que ara caldrà realitzar l'anàlisi de l'HTML, de com està estructurat el codi font a la web i extreure del seu anàlisi els patrons declarats a segona versió. La diferència amb la segona versió és el que l'administrador del sistema ha de portar a terme un cop ha localitzat aquests patrons. Cal que s'emmagatzemin a les noves taules i fer els càlculs necessaris per establir les probabilitats que necessitarà el procés d'obtenció de notícies que s'explicarà en el següent punt.

Recopilar notícies d'Internet

Aquest procés està dirigit pel sistema i és l'encarregat d'anar per les pàgines web dels mitjans de comunicació i buscar les notícies que corresponguin amb els paràmetres establerts a la Base de Dades en cada moment.

1. *Recuperar els patrons MÉS PROBABLES de la Base de Dades associats a cada Mitjà de Comunicació.*
2. *Recuperar enllaços dels Mitjans de Comunicació i descarregar-ne el contingut per filtrar-ho mitjançant els patrons recuperats*.
    a. Si s'han pogut recuperar el enllaços es passa al punt 3.
    b. Si no s'han pogut recuperar els enllaços es passa a buscar per la Base de Dades un patró de més a menys probabilitat d'encert per mirar si encaixa.
        i. Si encaixa un patró del mitjà de comunicació, es recalculen les probabilitats de tots els patrons associats al mitjà, a part de recuperar els enllaços i seguir amb el fil normal d'execució.
        ii. Si encaixa un patró que no està associat al mitjà de comunicació, se li associa i es recalculen les probabilitats, i es segueix amb el fil d'execució recuperant els enllaços i seguint amb el procediment.
        iii. Si no encaixa cap patró s'avisa a l'administrador del sistema que cal tornar a analitzar l'estructura de la pàgina web del mitjà de comunicació
3. *Descarregar el contingut de l'enllaç*.
4. *Recuperar els descriptors que s'han d'utilitzar per a la selecció*.
5. *Descarregar el contingut dels enllaços trobats*: ha d'anar enllaç per enllaç i filtrar pels descriptors.
6. *Coincidències a la Cerca:*
    a. No hi ha coincidència entre els descriptors i el contingut de la notícia: es passa a l'enllaç següent i el sistema guarda l'enllaç per no consultar-ho en posteriors cerques.
    b. Hi ha coincidència entre els descriptors i el contingut de la notícia: localitzar les parts principals de la notícia. Per poder fer això, el sistema haurà d'utilitzar els patrons pertinents:
        i. Si encaixa el patró més probable recuperat de la Base de Dades inicialment es continua pel pas 7.
        ii. Si encaixa un dels patrons següents en probabilitat, es continua el procés i es recalculen les probabilitats per posteriors recaptacions.
        iii. Si encaixa un dels altres patrons, caldrà associar-lo al mitjà de comunicació, recalcular les probabilitats i continuar amb el procés del punt 7.
        iv. Si no encaixa cap patró, caldrà avisar a l'administrador del sistema de que cal tornar a analitzar el codi font de la pàgina web del mitjà de comunicació.
7. *Emmagatzemar a la Base de Dades el resultat de la manera que convingui*.

Cal tenir en compte que en tot moment que s'utilitza un patró per fer la cerca, tan sigui dels enllaços de notícies, com després de les parts que està formada aquesta, pot ser que cap dels patrons que es tingui associats al mitjà de comunicació funcioni, per això cal tenir preparat el sistema per a que avisi a l'administrador de que en el mitjà de comunicació en qüestió ha ocorregut una excepció i que cal tractar-la.

Els paràmetres probabilístics que s'utilitzaran per calcular els pesos dels patrons seran dos: l'antiguitat al sistema del patró i la duració que aquest ha estat habilitat al sistema.

5 Ampliacions, millores i conclusions

Un cop es té la tercera versió es dóna per finalitzat el present PFC, ja que es considera que s'ha arribat a aconseguir els objectius que s'especificaven al inici de la documentació.

Aquest projecte es deixa en una fase de proves on s'ha d'anar veient l'efectivitat de l'algorisme probabilístic implementat i valorar les diferents opcions de disseny del mateix, si en un futur es veiessin caigudes importants del sistema. Per això cal tenir en compte que a dia d'avui es deixa un sistema en funcionament per mitjans de comunicació en el que no ha hagut canvis en la seva estructura interna des de que s'ha posat en funcionament el projecte, de tal manera no s'ha pogut valorar la capacitat de l'algorisme probabilístic dissenyat per donar solucions als possibles problemes al llarg de la recaptació de notícies.

Si es centra la visió a en les funcionalitats que l'aplicació dóna a l'usuari, es veu una línia clara d'ampliació, noves possibilitats a oferir a l'usuari, com són la possibilitat d'unes interfícies més gestionables, a mode de carpetes que facilitessin l'administració tant dels descriptors com de les notícies, podent així repartir la informació d'una manera forma i intuïtiva. Això oferiria a l'usuari la possibilitat de tenir un magatzem de notícies administrat tal com ell mateix decideixi.

Per un altre costat, com a possible millora, durant tot el projecte es diu que l'administrador del sistema ha d'anar realitzant inspeccions periòdiques a l'estructura web de cada mitjà de comunicació, per poder detectar els patrons que sorgeixen per poder captar les diferents informacions que es volen emmagatzemar a la Base de Dades. Aquest procés manual es podria automatitzar gràcies a mètodes d'Intel·ligència Artificial que detectessin automàticament els patrons al text, es necessitarien tècniques de reconeixement de text i algorismes preparats per la confecció automàtica d'expressions regulars.

Totes dues vessant, tant la millora del sistema com la posterior ampliació, impliquen l'estudi d'un món nou i que només s'ha deixat entreveure al llarg del desenvolupament d'aquest projecte: la Intel·ligència Artificial. Aquesta ciència s'intuïa que seria necessària en algun moment del desenvolupament donat el caire desconegut del problema al qual es volia donar solució al inici del projecte. La aplicació d'Intel·ligència Artificial al desenvolupament futur del projecte obriria els horitzons de les possibilitats d'un cercador de notícies a Internet.

Com a conclusions es vol destacar l'evolució del projecte a nivell de responsabilitats que s'han tingut al llarg del seu desenvolupament, posteriorment s'exposaran les implicacions que ha tingut el desenvolupament d'una aplicació d'aquest estil al llarg de tot el temps que ha durat el PFC, a nivell acadèmic i personals .

Fa un any i mig, quan es va començar a desenvolupar el projecte, havia entrat a treballar a una empresa que es dedica a donar servei de notícies als seus clients. Aquest servei, a l'origen de l'empresa només es feia de pont entre els clients i una altra empresa que sí realitzava les cerques a la Xarxa. Però al veure oportunitat de negoci, van decidir implementar la seva pròpia aranya cercadora i no haver de dependre d'un proveïdor de notícies. En aquest punt vaig arribar a l'empresa, just presa la decisió de començar el disseny d'aquesta aranya. Vaig poder participar en el procés de selecció de la plataforma, el llenguatge de programació, la Base de Dades i el seu corresponent gestor. I el responsable de l'empresa em va donar total llibertat per investigar sobre el món d'Internet i els portals dels mitjans de comunicació. Aquesta responsabilitat va suposar un gran incentiu per portar a terme el desenvolupament del projecte i quan a mitjans de la segona versió es va decidir no acabar d'implementar-ho, per les raons ja comentades durant la documentació, vaig demanar poder utilitzar el desenvolupat fins al moment com a Projecte Final de Carrera.

Després de l'acceptació per part de l'empresa de poder prosseguir amb el projecte fora de l'empresa vaig tenir la primera toma de contacte amb el director (Gustavo Patow) que ha acabat supervisant l'execució del projecte. Li vaig exposar la naturalesa del projecte i en quin punt estava i vam decidir continuar-ho per presentar-ho com a Projecte Final de Carrera d'Enginyeria Informàtica.

Per un altre costat, les implicacions a nivell acadèmic que ha provocat el desenvolupament del projecte han desembocat en l'obtenció de nous coneixements, tant sobre el llenguatge PHP utilitzat a la programació, com en tècniques d'optimització d'algorismes de captació i gestió de taules a Base de Dades. També s'ha començat a entreveure la necessitat d'adquirir nous coneixements en tècniques d'Intel·ligència Artificial com són els sistemes d'agents i multiagents.

Com a conclusió final destacar que a nivell personal aquest projecte a suposat un gran salt a nivell de coordinació de projectes ja que m'ha donat la possibilitat de veure la meva capacitat organitzativa, quins són els meus punts forts de treball i en quines àrees necessito adquirir més pràctica.