Universitat de Girona
Escola Politècnica Superior

Projecte/Treball Final de Carrera

Estudi: Eng. Tècn. Informàtica de Gestió. Pla 2001

Títol: Motor de cerca iSAC (Servei intel·ligent d'atenció ciutadana via web)

Document: Resum

Alumne: Albert Sabrià i Torrent

Director/Tutor: Miquel Montaner Rigall
Departament: Extern a l'EPS
Àrea:

Convocatòria (mes/any)**:** 09/2006

Contingut

Introducció del sistema

El projecte iSAC (Servei Intel·ligent d'Atenció Ciutadana via web) es va iniciar el mes de gener de 2006 amb l'ajut del nou coneixement científic en agents intel·ligents, junt amb l'aplicació de les Tecnologies de la Informació i la Comunicació (TIC) i els cercadors. Actualment, el servei actual d'atenció al ciutadà està composat per dues àrees: l'atenció directa a les oficines i l'atenció telefònica a través del Call Center. Les limitacions de personal i horari d'atenció fan que aquest servei perdi eficàcia.

Es vol desenvolupar un producte amb una tecnologia capaç d'ampliar i millorar la capacitat i la qualitat de l'atenció ciutadana en les administracions públiques, sigui quina sigui la seva dimensió. Tot i això, aquest projecte l'explotaran especialment els ajuntaments, als quals la ciutadania s'acosta amb tot tipus de preguntes i dubtes, habitualment no restringides a l'àmbit local.

Més concretament, es vol automatitzar a través d'un portal web l'atenció al ciutadà per tal d'obtenir un servei més efectiu.

El nou servei web vol trencar les limitacions esmentades mentre que es millora també la informació proveïda i la personalització d'aquesta a les necessitats de cada ciutadà. Com a objectiu més ambiciós del projecte destaca la humanització del servei, és a dir, que el ciutadà se senti atès com si una persona estigués interactuant amb ell i responent a les seves qüestions.

Així doncs, aquest projecte pretén desplegar un software lliure de tecnologia de buscadors de FAQ (en anglès, Freqüent Asked Questions) per oferir-lo a una ciutadania que sol·licita diàriament aquest servei. Aquestes FAQ es crearan des de l'experiència existent als serveis municipals d'Atenció Ciutadana.

D'impacte immediat en l'atenció directa efectiva, la millora de la qual serà percebuda i valorada per la ciutadania, per la reducció dels temps d'espera i per la millora de l'eficàcia en la resposta adequada.

En conclusió:

Comprendre'l quan parla un llenguatge col·loquial, o no utilitza les paraules correctes per no conèixer l'argot de l'administració pública.

Identificar la seva demanda encara que tingui errors d'ortografia.

Atendre al ciutadà en qualsevol demanda, encara que formuli preguntes que no transcendeixin a la responsabilitat de l'administració pública.

Està clar doncs que s'intenta oferir un servei a una ciutadania extensa i molt diferent entre ells. S'ha de crear un sistema capaç d'entendre a tothom, s'expressi com s'expressi, tot i que utilitzi paraules diferents a les que nosaltres podríem utilitzar. L'ambició és que dos usuaris demanin el mateix, utilitzant paraules diferents, i que el sistema sigui capaç de comprendre'ls i retornar el resultat desitjat.

Objectius

L'objectiu del projecte és el desenvolupament del motor de cerca del sistema iSAC.

Quan una persona, hagi de realitzar una consulta a l'atenció ciutadana, no serà un impediment el dia i l'hora que sigui. A través del portal web d'atenció ciutadana, qualsevol persona podrà realitzar la seva consulta. Per fer-ho, només haurà d'accedir al portal i introduir la consulta desitjada, i un cop retornats els resultats per part del sistema, escollir la més adient pels seus interessos.

Però aquest procés simple per l'usuari, es torna complex per part del sistema. La consulta introduïda per l'usuari no es tracta directament, sinó que ha de passar per un seguit de filtres i tractaments per tal d'ajustar al màxim la resposta retornada.

Amb això ens referim, a que tothom s'expressa de la seva manera, utilitzant el seu llenguatge col·loquial, i fins i tot amb paraules típiques i pròpies de la seva regió. També es poden produir errors ortogràfics i això no ha de ser un impediment.

Mitjançant tècniques de llenguatge natural, el sistema ha de ser capaç de proposar i corregir aquests errors, extreure tot tipus de paraules que no aportaran informació (articles, determinants,...), obtenir el lema de les paraules per poder englobar totes les diferents formes, funcions i possibles temps verbals, obtenir els sinònims i el significat d'expressions pròpies de la llengua, així com expressions de temps com poden ser "demà", "el dia de reis",...

Amb tot això, el sistema mostrarà a l'usuari, un llistat de FAQs que s'ajusten a la demanda realitzada, mostrant-les de manera ordenada segons l'afinitat. El càlcul d'aquesta afinitat serà un procés ampli i segurament variant a mida que va avançant el projecte. Es calcularà segons el nombre de paraules coincidents, la importància d'aquestes paraules dins l'estructura, el context de la consulta,...

Altres aspectes com podrien ser la gestió dels diferents diccionaris i FAQs, així com la integració dins l'estructura de bases de dades que estan en funcionament actualment en les oficines d'atenció ciutadana, queden fora de l'abast del projecte.

Un altre objectiu consisteix en la necessitat de generar codi open source. El gran interès del projecte és fer accessible els esmentats desenvolupaments en benefici dels ciutadans que tenen en l'administració pública local la seva primera finestra a l'administració.

També ho serà per tant el fet d'aprofitar-se d'altres aplicacions que ens poden ajudar a reduir temps alhora de desenvolupar el nostre sistema. L'exemple més clar, és la utilització del diccionari català de OpenOffie. Aquest mòdul OpenSource ens ajudarà a simplificar molt les possibles correccions ortogràfiques que l'usuari pugui haver introduït a la seva consulta. També s'ha utilitzat el diccionari FreeLing per tal d'obtenir els lexemes de les paraules.

Amb tot això remarcar, que es desitja desenvolupar i posar en funcionament l'aplicació utilitzant eines gratuïtes per tal que no suposi un cost afegit a l'administració pública posar en funcionament aquest sistema.

El motor de cerca iSAC

PAS 1 – Entrar la cadena de cerca: El pas imprescindible, evidentment, és la cerca que fa el ciutadà i que envia al sistema. Aquesta captura és senzilla, es limita a rebre la cadena de lletres o el conjunt de paraules a través de la xarxa de internet sense fer-hi cap modificació. D'aquesta manera no alterem la forma ni l'estructura i, tot i que acte seguit la tractarem, es conserva per poder fer una nova cerca amb les mateixes paraules que el ciutadà ha escollit.

Aquesta cerca la podem realitzar a totes les FAQs en general, o reduir la cerca cap a un context determinat, reduint així el temps de resposta i el nombre de possibles solucions per part del sistema. La forma de reduir aquesta cerca és escollint el context que té la nostra consulta mitjançant la barra d'eines iSAC.

PAS 2 – Pretractament inicial: Un primer pas a realitzar amb aquesta cadena entrada, independent de la resta, és treure els accents introduïts, així com transformar totes les paraules a lletra minúscula. També s'eliminaran els números, excepte aquells que s'interpreta que poden aportar informació, com podria ser un número seguit d'un mes de l'any, interpretant el sistema que es tracta d'una data.

PAS 3 – Correccions ortogràfiques: Un procés en paral·lel que es realitza amb aquesta cadena entrada, independent de la resta, és la proposta de correccions ortogràfiques a partir de la cerca inicial realitzada per l'usuari del sistema. Aquesta correcció ortogràfica prova d'analitzar tot el text contingut a la cadena de cerca i trobar aquelles paraules que no són acceptades a la llengua catalana.

Aquest procés es limita a fer una proposta de correcció ortogràfica i l'usuari del sistema triarà si l'aplica o si opta per no modificar el text original. En el cas d'haver-hi un error ortogràfic, proposarà sempre la solució més popular, és a dir, aquella paraula que té més opcions segons el diccionari de ser la paraula correcta.

Per tal de millorar en eficiència amb el tema de les correccions, i poder estar a l'últim dia pel que fa a noves paraules incorporades als diccionaris de cada llengua, s'han utilitzat els diccionaris de lliure distribució de l'aplicació OpenOffice.

PAS 4 – Eliminació de stopwords: El tractament pròpiament de la cadena de cerca introduïda pel ciutadà comença amb l'eliminació de les "stopwords" o paraules freqüents. Aquestes paraules freqüents representen totes aquelles paraules que, donat el seu ús massa freqüent o de la seva funció sintàctica, no aporten cap valor afegit al significat global de la cerca. Aquesta discriminació és molt senzilla, es disposa d'un llistat de paraules a bloquejar.

PAS 5 – Identificar el lexema: Eliminar aquella informació que no ens és útil i simplement conservar la que ens portarà a una millor obtenció de resultats. En el cas dels verbs, aquest procés ens extrauria el temps verbal, la persona, el gènere,...

Això ens és molt útil per tal de no haver d'introduir dins els descriptors de les FAQs totes les possibles conjugacions i variants de cada paraula, fet que seria molt costos, tant a nivell d'esforç per part del servei i l'administrador, com per volum de dades al nostra sistema.

Per tal de millorar en eficiència amb el tema del lexema, s'ha utilitzat el diccionari de lliure distribució del FreeLing.

PAS 6 – Cerca de sinònims: La cerca de sinònims és un dels passos interns que realitza el sistema, sense que l'usuari hi pugui intervenir. A partir de la cadena de cerca, el sistema busca sinònims de cadascuna de les paraules entrades per tal d'ampliar i millorar la cerca. Aclarir que la cerca de sinònims dins la taula esmentada es realitza mitjançant la comanda SQL "LIKE". Amb això ens assegurem també que possibles expressions escrites mínimament diferent també apareguin com a sinònim.

Aquesta cerca de sinònims és molt més que això. Dins aquesta taula hi tindrem, a més dels sinònims pròpiament dits, barbarismes, expressions populars, expressions típiques del llenguatge verbal com pot ser parlar de d'un dia sense especificar la data (Nadal , 25 de desembre), informacions locals (prominent de la xarxa de coneixement de l'administració ciutadana), dialectes de la zona,...

El fet d'analitzar sinònims beneficia la llibertat d'expressió i la manera d'entrar cada ciutadà la seva consulta. Dos usuaris poden voler el mateix i poden realitzar la consulta utilitzant paraules diferents, però el sistema respondre el mateix per ambdós casos.

PAS 7 – Identificar el context: Un filtre que intenta aplicar el sistema, per intentar acotar el rang, és intentar identificar el context de la cerca (en el supòsit que l'usuari no l'hagi introduït anteriorment des de la barra d'eines iSAC). En cas que el context de la cerca es redueixi a un, implicarà que el sistema únicament buscarà dins aquell context de FAQ. Si no és així, recordarà el percentatge de correspondència amb cada context per tal d'utilitzar-lo més endavant, i realitzarà la cerca per tot el conjunt de FAQs.

PAS 8 – La cerca: Finalment, ja només ens queda fer la cerca de cadascuna de les paraules (ja siguin les escrites per l'usuari com les afegides pel sistema) dins la taula de descriptors.

Segons el nombre de paraules coincidents a cada FAQ, el pes que tenen aquestes dins d'ella, així com els percentatges de context de les paraules analitzat anteriorment, ens serviran per retornar un llistat de FAQs ordenat per afinitat a la consulta entrada.

Amb aquest sistema es valora més la quantitat de coincidències que no pas la qualitat. Interpretem que és millor que coincideixin moltes paraules de la consulta (tant la inicial com amb les transformacions que ha generat el sistema) amb els descriptors de la FAQ, que no que aquestes siguin exactament les introduïdes únicament per part de l'usuari.

En cas d'empat entre dues o més FAQs a percentatge de possible resultat, el sistema els mostrarà aleatòriament per tal de no afavorir sempre una possible mateixa resposta.

PAS 9 – Aprenentatge:

- Aprenentatge per la correcció: com ja s'ha comentat, en el cas d'haver-hi un error ortogràfic, el sistema proposarà com a correcció ortogràfica la paraula més utilitzada (sempre dins el llistat de possibles solucions que retorna el diccionari). Per tant, aquest aprenentatge, ens servirà per actualitzar les paraules utilitzades i per tant millorar les possibles correccions posteriors.
- Aprenentatge de les FAQs: El fet d'escollir una FAQ, farà que el sistema afegeixi valor a aquella FAQ per a properes consultes, ja que el seu "índex de popularitat" ha incrementat.

Conclusions

El fet de compaginar estudis i treball, em suposava un esforç molt gran realitzar un projecte final de carrera. L'oportunitat que em va donar l'empresa on estava treballant com a becari de realitzar aquest projecte em va permetre aprofitar les hores de feina per tal de realitzar-lo.

Ha estat un repte molt important per mi haver participat en moltes de les tasques que comprèn un projecte d'aquest abast: entrevistes amb el client i entre l'equip de treball, sintetització d'idees, confecció dels requeriments, fer l'anàlisi i el disseny de parts de l'aplicació, programar-la, validar-ne les funcionalitats,...

Analitzant pròpiament el projecte, puc dir que un cop finalitzat el projecte cal valorar tots els punts marcats en els objectius. S'han obtingut satisfactòriament els objectius marcats i s'ha comprovat que es compleixen tots els requisits demanats: s'ha desenvolupat el motor de cerca d'un sistema al qual es poden adreçar tot els ciutadans, entenent el llenguatge col·loquial, essent capaç de preveure errors ortogràfics, millorar la cerca afegint sinònims, obtenint el lema de les paraules per tal d'ampliar la cerca i interpretant expressions populars, regionalismes i barbarismes. El sistema és capaç de respondre a la consulta del ciutadà, tot i que s'expressi d'una manera diferent a la que faríem qualsevol de nosaltres.

S'han posat en pràctica coneixements adquirits al llarg de la carrera sobre mètriques, llenguatges de programació, anàlisi i disseny de bases de dades, enginyeria del software i alternatives en el disseny d'aplicacions.

L'abast d'aquest projecte final de carrera ja s'ha assolit, tot i que el projecte que l'engloba ha fet poc més que començar. Espero en aquests nous passos, seguir aprenent com ho he fet fins ara.