Euskararen presentzia Interneten neurtu nahian

Iñaki Alegria, M. Jesus Rodriguez

i.alegria@si.ehu.es

Sarrera: Internet eta informazioaren bilaketa

Zalantzarik gabe Internet informazio-iturri nagusietako bat da gaur egun. Ezaguna denez, Interneten bidez hainbat zerbitzu eskuratu daitezke, posta elektronikoa, berriketa, urruneko prozesaketa etab. baina informazio-iturri gisa gailendu den zerbitzua Web edo WWW (*World Wide Web* ingelesez, mundu zabaleko armiarma-sarea) izenekoa dugu. Zerbitzu honen irudikapen hedatuena liburutegi erraldoi batena da, liburutegi txiki askoren elkarkonektatzetik sortzen dena.

Paperean euskaraz zenbat argitaratzen den interesatzen zaigun bezala, Webean euskararen presentzia zein den jakitea ere interesgarria da. Webean informazioa jartzeari argitaratzea ere esaten zaio, baina kasu honetan ez dago erregistrorik, ISBN edo bestelako kontrolik. Horren ordez, Webeko informazioari etekina ateratzeko badaude interesgarri egiten zaizkigun beste tresna batzuk: bilatzaileak eta direktorioak batetik eta programatu daitezkeen bestelako informazio-erauzleak bestetik. Ondorengo ataletan tresna horiek eta haien erabilera azalduko dira lehenik, Webean euskarak duen presentzia estimatzeko programatutako behatokiaren deskribapena gero, eta lortutako lehen datuak bukaeran.

Artikulu honetatik kanpo daude, beraz, soziolinguistikoki interesgarri izan daitezkeen beste neurketa batzuk: erabiltzaile euskaldunak, euskararen erabilera posta elektronikoetan, etab. (*Eustat*ek hainbat datu ematen ditu horretaz EAEko biztanlerian oinarrituta, www.eustat.es)

Informazioaren bilaketa eta erauzketa: bilatzaileak eta direktorioak

Informazioaren berreskurapena eta erauzketa

Informatikaren munduan bi arlo bereizi dira tradizionalki [1] [2]: informazioaren berreskurapena (IR, *Information Retrieval*) eta informazioaren erauzketa. (IE, *Information Extraction*). Lehenengoan dagoen informazioa modu zehatz eta azkarrean aurkitzea da helburu nagusia, bigarrenean berriz, oinarria den informazioaren tratamendu bat burutzen da, datu-base bat osatuz gehienetan. Arlo hauek Web mundura eramaten direnean **bilatzaileak** sortzen dira berreskurapenerako eta ***web-mining*** (web-meatzea) izenarekin definitzen den arlo oso bat erauzketa egiteko.

Bilatzaileak aski ezagunak diren bitartean beste erauzketa-sistema horiek ez dira hain ezagunak, hala ere oso praktikoak izan daitezke, programazio-lan handi samarra eskatzen duten arren. Informazioaren erauzketa modu erabat automatikoan edo semiautomatikoan egin daiteke, bigarrenean programa batek laguntzen du datu-basea sortzen baina giza lana egongo da zalantzazko kasuetan erabakiak hartzeko edo lan guztia gainbegiratzeko. Horren ondorioz, lehen kasuan programak sofistikatuafoak izan beharko dira. Ondoren azalduko den *citeseer* da erauzketa automatikoaren adibide bat, eta hainbat direktorio (*yahoo* adibidez) eskuzkoarena.

Citeseer (http://citeseer.nj.nec.com/cs) informatikaren inguruko liburutegi digital bat da, artikuluen erreferentziak, artikuluak beraiek eta beste hainbat datu modu automatikoan

lortzea izanik bere berezitasuna. Beraz, gai horren inguruko artikuluak kontsultatzeko munduko liburutegi onenetako bat da. Gainera, ohiko liburutegian aurkitzen ez diren informazio berriak gehitzen dira, inpaktuaren estimazioa esate baterako.

Adibide horretatik abiatuta tresna hauek eskaintzen dituzten aukeretara hurbil gaitezke: komunikabideetan enpresek duten agerpenen selekzioa eta neurketa, e-posta helbideen katalogoa, enpresen edo enpresen direktiboen bilakaera prentsa ekonomikoan oinarrituta, teknologiaren behaketa automatikoa, etab. Teknika hauen sofistikazioaz eta datu-base arrunten informazioa ere erabiliz ezagutzaren kudeaketa burutzen da. Teknologia hauetan oinarrituta sistema asko sortzen ari dira azken urteetan. Euskal Herrian hainbat ikerketa, proiektu [3] [4] eta enpresa (www.eleka.net, www.diana-tek.com) ere badaude arlo honetan.

Informazioaren metrikaren esparruan inpaktu handia izan dute tresna horiek eta termino berriak sortu dira horren eraginez, *webometrics* adibidez..

Bilatzaileak: *Google* eta *Alltheweb*

Bilatzaileak dira gaur egun IR arloaren teknologiarik sofistikatuenetako bat. Haien konplexutasunaren iturria informazio-kopurua eta sakabanaketa da. Baldintza horietan sistema azkarrak eta zehatzak lortzea lan erraldoia da. Hiru osagai nagusi dituzte bilatzaileek: robota, indexatzailea eta bilatzaile bera.

Robotak Internet sarea miatzen du etengabe, dokumentu berri, eguneratu eta desagertuen bila. Indexatzaileak dokumentua sailkatzen du hainbat parametroren artean, hizkuntza adibidez, eta bere hitzak datu-base erraldoi batean sartzen ditu. Indexatzaileak, aurreratua denean, hitzak indexatzeaz gain lemak, kontzeptuak edo bestelako egitura linguistikoak gordetzen ditu, baina horrek moteltzen du indexatze-prozesua eta gutxitan erabiltzen da. Bilatzailearen azken moduluak interfaze bat eskaintzen du galderak egiteko eta dokumentuen berri aurkezten du erantzun gisa. Dokumentu gehiegi aurkitzearen arazoa aurkezteko ordena mugatzeko erabakitzeko teknika funtsezkoa da. Ordainketaren truke posizio hobetzeko aukerak polemika piztu du hainbat forotan.

Urteekin batera bilatzaile arrakastatsuenak (http://searchenginewatch.com) aldatuz joan dira *Altavista* (www.altavista.com) izan zen lehen ahaltsua, gero *Fast*, gaur egun *AlltheWeb* izenarekin ezagutzen dena (www.alltehweb.com), eta azken urteetan *Google*k (www.google.com) du erabateko nagusitasuna.

*Google*ren arrakasta izugarriaren zergatia hiru ezaugarrietan datza: dokumentu gehien indexatzen duena da, oso azkar erantzuten du eta aurkezteko ordena oso landua du. Euskaldunentzako eragozpen bat du, interfazea euskaraz eskaini arren indexatzean ez du euskara bereizten, eta honek zailtasun batzuk ekartzen ditu hainbat erabilpenetarako[1]. Ordainketaren truke informazioa gailentzeko aukera ematen du baina beti pantailaren beste kokapen batean aurkeztuko du informazio hori.

---

[1] Aipatutako ezaugarri horiek direla eta, ezusteko arlotan ere erabiltzen da *Google*. Adibidez, lexikografian edo idazketan interesantea izan daiteke jakitea euskaraz *webgune*, eta *web gune* terminoen artean zein erabiltzen den gehiago. *Google*k euskara bereizten ez duenez maiztasun handieneko hitza (eta) gehitzen dugu bilaketan, eta bi kontsulta egiten ditugu. Erantzunen arabera 14700 dokumentutan azaltzen da *webgune*, eta 3280tan *web gune*, beti kasu-markarik gabe bilatuta, euskarazko lematizazioa ez baitute egiten bilatzaileek. *Alltheweb* bilatzaileak 13412 eta 3795 itzultzen ditu hurrenez hurren, baina segurtasun gehixeago eskaintzen du euskara ezagutzen baitu (ez beti ondo, hala ere).

Alltheweb oso interesgarria da euskaldunentzat, bigarren bilatzaile ahaltsuena izateaz gain euskara ezagutzen duelako. 49 hizkuntzetan bilaketa egiteko aukera ematen duen bilatzailea da. Dituen gune kopurua hizkuntzetan klasifikatuta dauka eta bilaketa bat egiteko orduan hizkuntza bat aukeratuz eta nahi den terminoa sartuz, hizkuntza horri dagozkion orrietan burutzen du bilaketa. Euskararen behatokia programatzean aukeratu da *Google*ren gainetik ezaugarri hau dela eta.

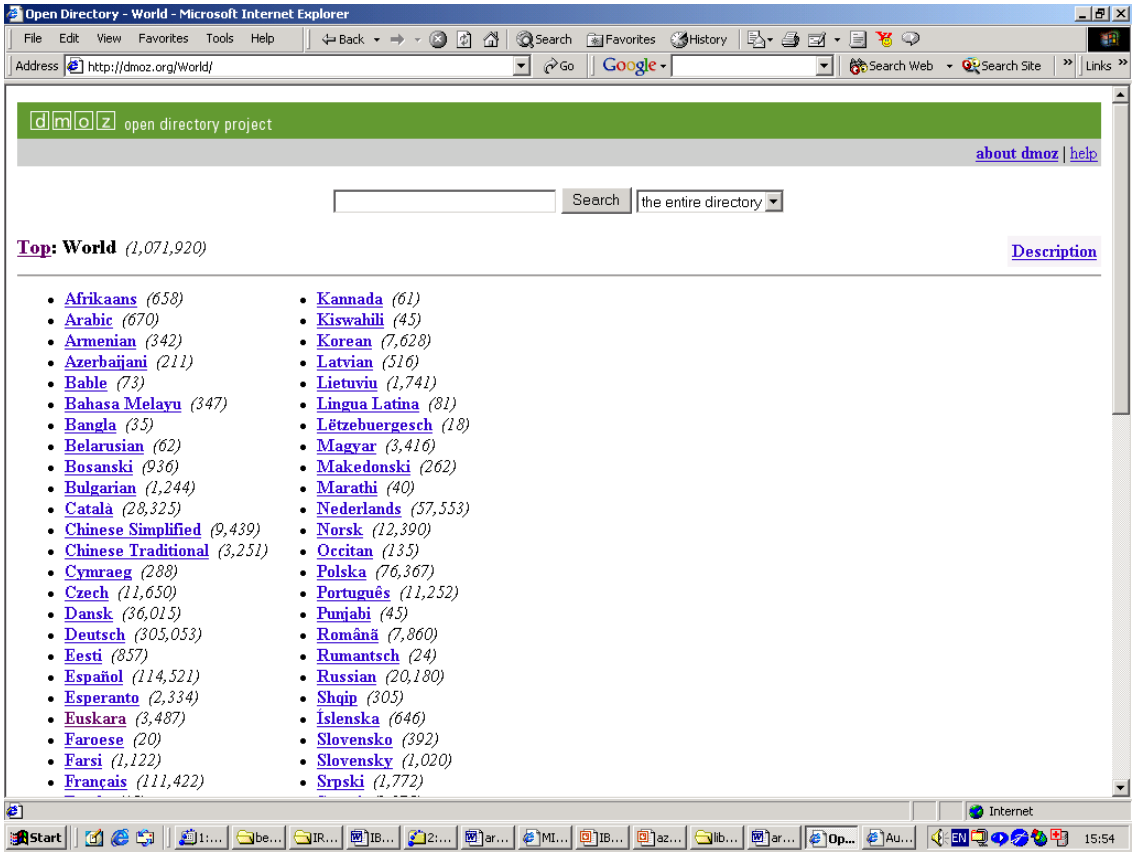
Direktorioak: *Yahoo*, *DMOZ*, ...

Direktorioetan informazio gutxiago aurkitzen da, baina informazio hori kualitatiboki hobea da, hierarkia batean ondo antolatuta dagoelako eta kalitate minimoa ziurtatzen duen onespen bat duelako. Horretarako lana automatikoa izan beharrean eskuzkoa izan ohi da, erabat edo programa baten proposamenen azterketaren ondoren.

Beraz, direktorioetan dokumentuak hainbat ataletan banatzen dira, eta atalen barruan azpiatalak bereizten dira. Hierarkia horretan nabiga daiteke interesatzen zaiguna aurkitu arte. Bilaketa errazteko asmotan sistema osoan, kategorian edo azpikategorian katalogatutako dokumentuen artean bilaketa arruntak egin daitezke.

Direktorio ezagunenak bi dira *yahoo* (www.yahoo.com) eta *dmoz* (http://dmoz.org), bietan eskuzko lana nagusia izanik. Lehena famatuena izan arren bigarrenak dituen bi ezaugarrik oso interesgarri bihurtzen dute:

- Ingelesa nagusia izan arren beste hizkuntza guztietara irekita dago, *World* azpikategoriaren barruan hizkuntzak gehitzeko aukera eskaintzen baitu (http://dmoz.org/World). Gaur egun 71 hizkuntzetako dokumentuak aurki daitezke bertan. 1. irudian pantaila nagusia ikus daiteke.
- Edozein pertsona libre da bere webgune kutunak gehitzeko, sistema irekia eta librea baita. Horren ondorioz direktorio bera librea da, kopiatu daiteke eta nahi dugun webgunean integratu. Horrela, beste askoren artean, *aurki*k (www.aurki.com) euskarazko hierarkia integratzen du, eta *Amfibi*k (http://directory.amfibi.com/c) osoa eta katalanezkoa.



1. irudia.- *dmoz* hainbat hizkuntzatan

Direktorioak interesgarriak dira informazioaren iturburu gisa, baina kalitate handiago duten arren bilatzaileekin alderatuta askoz ere dokumentu gutxiago indexatzen dute, eta muga horrek galarazten du iturri gisa erabiltzea aplikazio askotan.

Euskaraz: *aurki*, *jalgi*, *kaixo*, ...

Euskal Herrian egindako bilatzailerik ez dago, oso garestiak baitira, nahiz eta euskarazko dokumentuak baino ez bilatu, sarea osoa miatu behar baita abiadura handiz. Horren ordez geroago azalduko diren direktorio batzuk sortu dira eta horien barruan bilaketak eskaini, baina ez dute sare osoa usnatzen duen robotik.

Direktorio batzuk, berriz, badaude. Aipatutako *Aurki*, *Jalgi* (www.jalgi.com) eta *Kaixo* (www.kaixo.com) erabilienak dira, eta lehen biak euskara hutsezkoak diren bitartean azkena elebiduna da. Dena den, sailkatutako dokumentu kopurua urria da guztietan. *Aurki* proiektua sendotzea beharrezkoa da, beste gauzen artean, euskararen presentzia ebaluatu ahal izateko.

Webeko informazioaren izaera, dimentsioa eta hizkuntzak

Berreskuratu edo erauzi egin nahi den informazioaren izaera eta bolumena aztertzea ezinbesteko urratsa da sistema (semi)automatikoak eraikitzeko.

Datu-baseetan dagoen informazioa egituratutzat hartzen den bitartean Webean dagoen informazioa desegituratua edo erdi-egituratua da. Gainera, inolako kontrolik ez dagoenez, hizkuntzaren kalitatearen aldetik oso aldakorra da. Dena den, oso interesgarria da informazio-iturri gisa, informazio kopuruari zein aberastasunari begira. Informazioa desegituratua delako eta edukien kalitatea zalantzazkoa delako, lortutako informazioaren fidagarritasuna beste iturburuena baino txikiagoa izan daiteke.

Webean dagoen dokumentuen kopurua zehazteko garaian hainbat faktore hartu behar dira kontuan:

- Dokumentu ikusezin asko daude Webean arrazoi desberdinengatik, babestuta daudelako, web dinamikoak direlako (informazioa datu-baseetan dago eta galderak egitean baino ez dira dokumentuak sortzen) eta bere informazioaren formatua ezin delako interpretatu besteak beste.
- Dokumentuak eskuragarriak eta ulergarriak izanda ere, aurkitu behar dira, eta horretarako bilatzaileak dauden arren hauen estaldura nekez iristen da %50era.
- Internetek duen dinamikotasuna dela eta, gaur aurkitzen dugun dokumentu bat bihar aurkituko dugula ez dago ziurtatzerik, beraz, bilatzaileetan dauden hainbat dokumentu dagoeneko desagertu dira. Are gehiago, desagerpen horiek behin-behinekoak izan daitezke, zerbitzarian edo sare lokalean gertatutako akats batengatik adibidez.

Bilatzaileek dira, hala eta guztiz ere, Webaren dimentsioa emateko iturburu fidagarriena. Webaren analisietan espezializatuta dagoen www.notess.com/search webguneak 2002ko abendurako ematen dituen datuak 2. irudian azaltzen dira.

Bilatzailea	*Showdown*en estimazioa (milioitan)	Bilatzaileak zehaztutako dimentsioa (milioitan)
Google	3.033	3.083
AlltheWeb	2.106	2.112
AltaVista	1.689	1.000
WiseNut	1.453	1.500
Hotbot	1.147	3.000
MSN Search	1.018	3.000
Teoma	1.015	500
NLResearch	733	125
Gigablast	275	150

2. irudia.- Bilatzaileen estaldura estimatua

Datu horietan oinarrituta ondorio zuzenena hauxe da: 3 mila milioi dokumentu inguru berreskura daitezke! Dena den, oso inportantea da dinamikotasuna kontuan hartzea hazkundea oso azkarra baita. Aurreko iturriaren datuak azterturik 2002 urtean edukiak %50ean hazi direla estima daiteke.

Hizkuntzen presentziaren aldetik, argi dago ingelesa nagusia dela Webean, teknologia honekin duen loturarengatik eta *lingua franca* izateagatik. Hala ere, Webaren %80 baino gehiago izatetik %65 baino gutxiago izatera pasa da urte gutxitan.

Gainontzeko hizkuntzen artean begi bistan dago bi faktore nagusi hartu behar direla kontuan, hiztunen kopurua eta garapen ekonomikoa, bigarren faktorea askoz garrantzitsua izanik beste arlotan, liburugintza esaterako, baino.

Datu batzuk ematearren 1999ko irailean egindako bilaketetako datuak aldatuko ditugu hona. *Altavista* bilatzailean egindako azterketan oinarrituta hauek ziren hizkuntzen araberako kopuruak milatan[2]: guztira 137.500, ingelesez 103.500, frantsesez 2.700, gaztelaniaz 2.400, suomieraz 480, islandieraz 40, euskaraz 4. Gaur egun, 2003ko uztailean, *AlltheWeb* (www.alltheweb.com) bilatuta datu hauek lortzen dira[3]: ingelesez 874.855, frantsesez 43.535, gaztelaniaz 45.589, euskaraz 58. Datu guzti hauen zehaztasuna mugatua da, hizkuntzaren detekzioa hurbilpen batez egina baita, eta gainera neurketa bilatzaile desberdinetan egina dela kontuan hartu behar da.

dmoz direktorioan oinarrituta, berriz, beste datu hauek lor daitezke (ikus 1. irudia): frantsesez 111, gaztelaniaz 114, suomieraz 10, islandieraz 0,6 eta euskaraz 3,4. Lehen bezala kopuruak milatan emanda daude. Datu hauen estimazioa eskasagoa da, esan bezala direktorioetan editoreen lanaren eragina erabatekoa baita. Hala ere, datuak interesgarriak dira, denboran zeharreko bilakaerari begira batez ere.

Hizkuntza latinoetan zentratutako azterketa sakonagoak eta interesgarriak egin dira 1995, 1998, eta 2001erako [5]. Bertatik atera dira bigarren irudiko datuak, denak *Al theWeb* bilatzailean oinarrituta. Web osoarekiko eta ingelesarekiko portzentajeak ematen dira irudi horretan hainbat hizkuntzatarako[4].

	ingelesa	frantsesa	gaztelania	suomiera	islandiera	euskara
hizkuntza/guztiak	64,94	3,14	2,62	0,39	0,04	0,01
hizkuntza/ingelesa	100,00	4,83	4,04	0,60	0,07	0,02

3. irudia.- Hizkuntzen arteko konparaketa

Aurreko bilaketa horiek bilatzaileen *bilaketa aurreratua* modua erabiliz lor daitezke edozein momentutan, bilatzaileak hizkuntza bereizten badu.

*Softcatalà*k berriki argitaratutako azterketan[5] [6] beste datu batzuk lortzen dira. *AlltheWeb* bilatzailean ere oinarrituta 48 hizkuntza hartzen du kontutan, eta euskara, 155 mila orri inguruan esleiturik, 40. tokian geratzen da zenbaki absolutuetan,. Hiztun kopurua kontuan hartuta, 0,2 inguruko koefizientea edukiko luke euskarak eta 24. toki inguruan leudeke gaztelaniaren eta portugesaren pare, baina Europako herri aurreratuetatik urruti (alemaniera 1,82; frantsesa, 1,39; italiera 0,67, katalanera 0,45).

Beste kontu bat da orri horien azterketa kualitatiboa. Gai hori artikulu honetatik kanpo geratzen da baina Andoni Sagarnak egindako azterketa oso gomendagarria da [7].

Euskararen presentzia: Software-katalogoa eta behatokia

Aurretik azaldutako datuen arabera Webean euskararen presentzia oraindik txikia da, horregatik software-katalogoa egitean ez genuen software-bilketa soila egin nahi, beste tresnak eskaini nahi genituen, gauden egoera aztertzeko eta ahal den neurrian dauden hutsuneak betetzeko. Tresna hauen artean berriak publikatzeko aukera eta behatokia

---

[2] kontuan hartu bilatzailean jasotako kopuruak direla

[3] *Altavistan* hizkuntzen arabera bilatzeko modua aldatu da eta arazoak izan ditut. Geroago behatokian azaltzen den metodoa erabili da neurketa honetan.

[4] Azterketan hezkuntza gehiagotarako datuak daude: http://www.funredes.org/LC/english/L5/L5appendix_6.html

[5] Euskarazko laburpen bat aurki daiteke helbide honetan: www.sustatu.com/1062598690

sartzen dira. Berrien bidez, martxan dauden proiektuak, hauen arazoak edo produktu berriei buruzko informazioa eman nahi da, eta behatokiak euskararen presentzia sarean aztertzeko gunea izan nahi du.

Software-katalogoa (http://softkat.ueu.org) UEUk bultzatutako proiektua da, Bizkaiko Foru Aldundiaren laguntza duena. Softwarean espezializatutako direktorio bat da, euskaraz sortzen diren programak eta zerbitzu informatikoak bilatzen laguntzeko. 2002 urtean hirugarren bertsioa diseinatu zenean behatokia barneratzeko ideia sortu zen, euskararen presentzia kuantifikatzeko modu sistematikoan.

Interneten edukiak etengabe aldatzen dira, hori dela eta, berari buruzko estatistikak lortu nahi izanez gero datuen bilketa eta jarraipena egin behar da, bestela, datu hauek betiko gal daitezke. Esate baterako, ezinezkoa izango da jakitea euskarazko zenbat webgune zeuden aurreko hilabetean aurretik datu hori nonbait gorde ez baldin bada.

Bestalde, periodo batean egon diren eboluzioak edo tendentziak neurtzeko aurretik ere datu-bilketa bat egin behar da eta datuen kopuru bat edukitzean estatistika aberatsagoak eskaintzea posible izango da. Hau da, hain zuzen ere, behatokiaren lehenengo helburua, datuen bilaketa bat egitea, etorkizunean estatistika aberatsagoak eta osatuagoak eskaintzeko.

Aurreko hori guztia kontuan hartuta, informazioaren erauzketaren arloan dagoen tresna bat eraiki da, behatoki izena duena, martxan dagoena eta informazio interesgarria eskaintzen duena: http://softkat.ueu.org/hizkunkonp.php

Iturriak eta irizpideak

Datu soilak baino konparaketak eta bilakaerak interesgarriagoak direnez, beste bost hizkuntzaren egoera ematea ere erabaki zen, kontsultatzen duenak beste erreferentzia batzuk eduki ditzan. Hasteko euskaratik gertutasun geografiko eta erabilpenagatik hurbilen dauden hizkuntzak hautatu dira, hizkuntza hauek katalana, galegoa, gaztelania, frantsesa eta ingelesa izanik. Dena den, erreferentzia interesgarri eta eredu izan daitezkeelako, suomiera eta islandiera sartzeko beharra detektatu da.

Iturriak aukeratzeko orduan, hizkuntza ezberdinetan bilaketak egiteko aukerak ematen duten bilatzaileak eta direktorioak hartu dira kontuan, eta proiektua martxan jartzeko bi iturri hauek hautatu dira: *AlltheWeb* bilatzailea eta *dmoz* direktorioa.

Bilaketarako irizpideak ezberdinak izan dira bi iturrietan. Fidagarritasuna gehitzeko asmotan *AlltheWeb* bilatzailearen kasuan “eta” hitzaren agerpena eskatzen zaie dokumentuei euskarazkoak direla onartzeko, guk egindako probetan konturatu baikinen euskarazkotzat hartzen zituen hainbat dokumentu beste hizkuntza batean zeudela[6]. Kontsistentzia ziurtatzearren hizkuntza guztietan hitz arruntena (maiztasun handienekoa) ere gehitu da bilaketetan; beraz, katalanez “i” bilatu da, galegoz “e”, gaztelaniaz “y”, frantsesez “et” eta ingelesez “and”. Murriztapen hau dela eta, behatokian azalduko diren neurriak txikiagoak izango dira aurreko batzuk baino, baina zehaztasuna irabazi delakoan gaude. *dmoz* direktorioan, berriz, ez dugu zertan bilaketa egin behar, bakarrik nahi dugun hizkuntzaren dokumentu kopurua begiratu.

---

[6] Horren ondorioz behatokian lortutako kopuruak aurreko zenbakiak baino txikiagoak izango dira, baina hizkuntza guztietarako modu berean eginda, uste dugu datu fidagarriagoak direla.

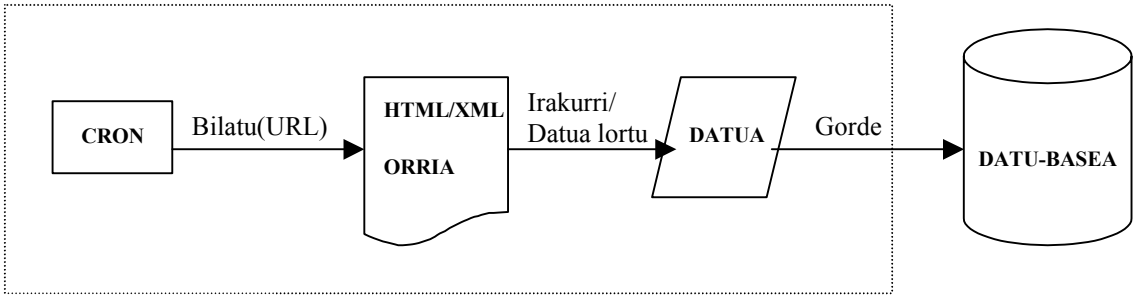
Teknologia

Behatokia gauzatzeko orduan, egin beharreko datu-bilketa ahalik eta automatiko egitea erabaki zen, hau da, behatokiaren eguneraketa lan periodiko automatizatuen bidez egitea.

Hau burutzeko *Unix* sistemaren *cron* izeneko tresnan oinarritu gara.C*ron* atazen planifikatzailea da, zeinek minuturo aztertzen duen ea ordu horretarako programatutako prozesuak dauden, eta horrela izanez gero bere exekuzioa burutzen du.

4. irudian jarraitutako prozesuaren azalpena errepresentatzen da. Urratsak honako hauek dira:

- *cron* tresna programatzen da guk nahi dugun maiztasunaz buru dadin, kasu honetan hilero.
- Unea heltzen denean prozesu bat jartzen da martxan zeinek URL[7] bati dagokion dokumentua irakurtzeko eta HTML edo XML formatuan egongo den web-orria lortu ondoren fitxategi batean gordetzen da. Gure kasuan *AlltheWeb* edo *Dmoz*eko informazioaren helbide zehatza hartuta Interneten bilaketa bat burutzen du automatikoki. Hizkuntza bakoitzerako URLa aldatzen da, hizkuntzaren aipamen bat barneratzen baita URLan.
- Gordetako fitxategia lerroz lerro irakurtzen da bilatzen den datua topatu arte. Aplikazio honetan funtsezko datua lortutako webguneen kopurua izango da.
- Behin datua lortu ondoren aplikaziorako definitutako datu-base batean gordetzen da erabiltzeko prest.



4. irudia.- Prozesu automatiko urratsak

Behin datu-basean datuak edukita, beste modulu independente batek datuak kontsultatzeko aukera eskaintzen du. Datuak eskaintzeko taulak eta grafikoak erabiltzen dira, berauen sorkuntza guztiz dinamikoa izanik. Hau da, erabiltzaile batek datuak kontsultatzeko eskaera egiten duen momentuan, eskatutako datuak datu-basean bilatzen dira eta momentuan dagokion taulak eta grafikoak sortzen dira modu erabat automatikoan.

Azaldutako prozesu guzti hau egiteko, lan informatikoarekin batera beste lan bat egin behar izan da; iturburu diren webguneen azterketa sakona, dokumentu kopurua lortzeko zehaztu behar den URL zehatza jakiteko eta lortzen den orritik dokumentu-kopurua ondo eraucteko datu-basera. Aplikazioaren bizitza osoan zehar iturburu horien jarraipena egin beharko da, gune horietan aldaketak gertatuz gero behatokian aldaketa batzuk egin beharko dira. Dena den, diseinatutako sistemaren malgutasuna dela eta, aldaketa horiek burutzea edo bilatzaile, direktorio zein hizkuntza berriak gehitzea oso lan erraza da, programa ukitu gabe egiten baita.

---

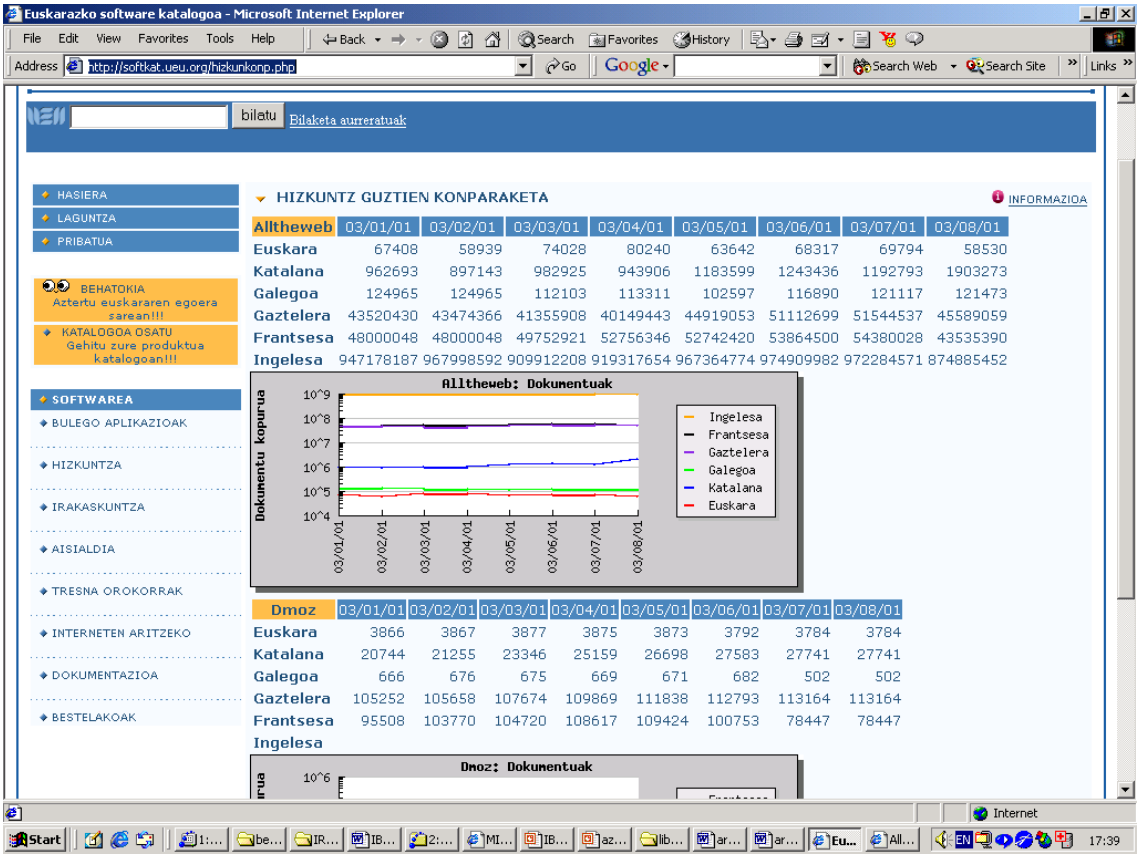
[7] URL esaten zaio web-orri baten helbide zehatzari. Testuan zehar agertzen diren web-orrien helbideak URLak dira.

Azpimarratu behar da deskribatutako bi prozesu nagusiak, datuen bilketa eta kontsulta, erabat automatikoak direla. Horrekin batera esan behar da denbora pasa ahala datu gehiago lortuko direla, hilero datu-basea elikatzen baita automatikoki.

Lehen emaitzak

5. irudian agertzen dira gunearen itxura eta hizkuntza guztietarako lortu diren aurtengo zortzi hiletako zenbakiak. Dena den artikulua irakurtzen duzunerako datu gehiago egongo da, beraz, http://softkat.ueu.org/hizkunkonp.php helbidera jo, kontsultatu eta zure helbide kutunetan sartu.

Datu horiek aztertzean bitxikeri bat agertzen da, hainbatetan kopurua jaisten da eta beste hainbatetan emendatu egiten da bat-batean. Honen arrazoia bilatzaileen lan egiteko modua da, noizean behin garbiketa egiten dute desagertutako dokumentuak ezabatuz eta modu independentean gune berriak gehitzen dituztelarik. Hori dela eta, bilakaera modu fidagarriagoan aztertzeko epeak hiruhilabetekoak izan behar dira gutxienez.



Alltheweb	03/01/01	03/02/01	03/03/01	03/04/01	03/05/01	03/06/01	03/07/01	03/08/01
Euskara	67408	58939	74028	80240	63642	68317	69794	58530
Katalana	962693	897143	982925	943906	1183599	1243436	1192793	1903273
Galegoa	124965	124965	112103	113311	102597	116890	121117	121473
Gaztelera	43520430	43474366	41355908	40149443	44919053	51112699	51544537	45589059
Frantsesa	48000048	48000048	49752921	52756346	52742420	53864500	54380028	43535390
Ingelesa	947178187	967998592	909912208	919317654	967364774	974909982	972284571	874885452

Dmoz	03/01/01	03/02/01	03/03/01	03/04/01	03/05/01	03/06/01	03/07/01	03/08/01
Euskara	3866	3867	3877	3875	3873	3792	3784	3784
Katalana	20744	21255	23346	25159	26698	27583	27741	27741
Galegoa	666	676	675	669	671	682	502	502
Gaztelera	105252	105658	107674	109869	111838	112793	113164	113164
Frantsesa	95508	103770	104720	108617	109424	100753	78447	78447
Ingelesa								

5. irudia.- Lortutako lehen emaitzak

Ondorioak eta etorkizuneko helburuak

Orain arte proiektuak jaso duen harrera oso positiboa izan da, proiektu berria izanik interes handia piztu du eta etorkizun handia aurreikusten diogu.

Etorkizunean proiektua zabaltzeko asmoa dugu, hizkuntza gehiago sartuz, aipatutako suomiera eta islandiera lehenak, eta hizkuntza bakoitzeko webgune kopuruen bilakaera eskainiz, orain ematen diren datuak esanguratsuagoak bihurtuz (hiztun kopuruarekiko tasa sartuz adibidez).

Gainera, datuak gehitzea aurreikusten da, orain ematen diren baino datu esanguratsuagoak eskaintzeko asmoz. *Google*ko datuak ere integratzea oso garrantzitsua litzateke fidagarritasuna handitzeko, batetik dokumentu gehien sailkatzen duelako, bestetik *AlltheWeb* bilatzaileak eskaintzen dituen datuak ezegonkorrak direlako.

Bibliografia

[1] Strzalkowsky T (ed.). *Natural Language Information Retrieval*. Kluwer, 1999.

[2] Pazienza M. T. (ed.) *Information Extraction.* 1997. Springer.

[3] Garatea, J. Berbagune proiektua: euskararen erronka teknologia berrietan.. *erabili.com* 2003. http://www.erabili.com/zer_berri/muinetik/1056729701

[4] Alegria I., Balza I., Ezeiza N., Fernandez I., Urizar R. 2003 Named Entity Recognition and Classification for Texts in Basque. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1061807449/publikoak/Enti.pdf

[5] Funredes. Languages & Cultures. 2001. http://funredes.org/LC/english/L5

[6] Mas i Hernàndez, J. 2003. La salut del català a Internet. http://www.softcatala.org/articles/article26.htm

[7] Sagarna, A. 2003. Euskarazko edukiak dituzten web guneak (agertzeko). http://www.erabili.com