# Datuetatik ezagutzara. Web orrietan nabigatzean utzitako aztarna abiapuntu

*Olatz Arbelaitz*, Aizea Lojo, Javier Muguerza, Iñigo Perona*

Konputagailuen Arkitektura eta Teknologia Saila. ALDAPA ikerkuntza taldea.
Euskal Herriko Unibertsitatea (UPV/EHU)

*olatz.arbelaitz@ehu.es

Jasoa: 2013-06-01 Onartua: 2013-10-15

**Laburpena:** Teknologia berriak direla medio informazio asko metatzen da gaur egun eta gainera, gehiena formatu digitalean. Askotan, informazio hori kontzienteki gordetzen da eta beste hainbatetan berriz, gure ekintzen albo ondorio gisa. Metatutako informazio hori guztia, zergatik ez erabili datuetan bertan ez dagoen ezagutza sortzeko? Hauxe da datu-meatzaritza eta ikasketa automatikoko tekniken helburua. Webguneetan nabigatzen dugunean uzten dugun aztarna izan liteke datu-meatzaritzak zukua atera diezaiokeen datu multzoetako bat. Lortutako ezagutzak erabilera anitz ditu: baliabideak egokitzea edo webgunea pertsonalizatzea, gomendio sistema baten oinarri izatea edo zerbitzu-emaileari bere webgunean nabigatzen duten erabiltzaile moten berri ematea. Ezagutza hori lortzeko erabil litezkeen tresnak eta prozesua deskribatzea da artikulu honen helburua.

**Hitz gakoak:** datu-meatzaritza, web-meatzaritza, ikasketa automatikoa, erabiltzaile profilak.

**Abstract:** Nowadays' technologies allow the storage of large amounts of information, most of it in digital format. That information is often stored consciously. However, many other times the information is stored as a side effect of our actions. Why shouldn't we use all the stored information to extract knowledge that data does not show explicitly? The track we generate when navigating in websites can be one of the data sources to use data mining for extracting knowledge. The knowledge extracted can have different uses: to adapt the resources or web personalisation, to be the base of a recommender system or to inform the service provider about the type of users navigating in the website. The aim of this paper is to describe the tools and processes that can be used to extract that knowledge.

**Keywords:** data mining, web mining, machine learning, user profiles.

## 1. **SARRERA**

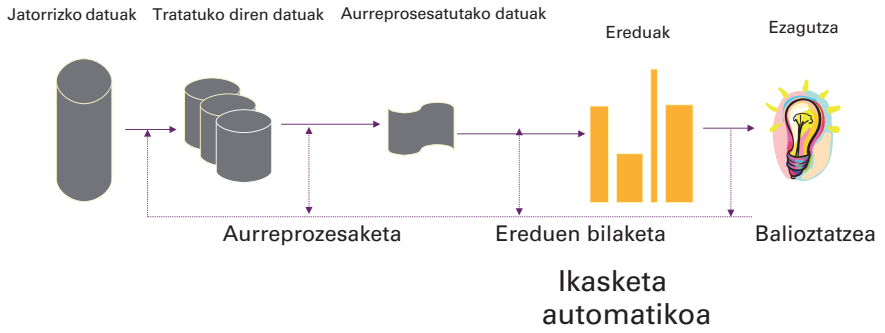
Digitalizazio garaiaz geroztik informazio asko pilatzen da gure inguruan eta gainera, teknologia berriei esker informazio hori ugaritzeaz gain edonorentzat atzigarriago bihurtu da. Informazio/teknologia konbinaketak eskaintzen dizkigun aukerak zabalak dira. Oinarrizkoak datuak eskuratzea, estatistikak ateratzea, bilaketa anitzak modu errazean egitea, datuak erlazionatzea, historiaren parte gisa gordetzea, etab. izan litezke baina badaude urrunago doazen aukerak; informazio hori datuetan agerian ez dagoen ezagutza ateratzeko erabil liteke. Horretarako, beharrezkoa izango da datu-meatzaritza prozesu bat aurrera eramatea eta tartean datuetatik ezagutza ateratzea ahalbidetzen diguten ikasketa automatikoko teknikak erabiltzea. Ezagutza horrek erabilera anitz izan ditzake baina beti ere lagungarria izango da.

Bestalde, informazioa kontzienteki eta intentzio osoz gordetzea da ohikoena baina gaur egun eta berriz ere teknologia berriei esker, guk bururutako ekintza asko automatikoki dokumentatzen dira. Hau da, automatikoki, erabiltzaileok oso kontziente izan gabe, informazio ugari metatzen da. Horren adibide izan litezke txartel elektronikoekin egiten ditugun eragiketei buruzko informazioa edo web orrietan nabigatzean egindako mugimenduen aztarna. Nahita gordetako informazioarekin gertatzen den moduan, informazio hori ere erabilgarria izango da datu-meatzaritza (DM) prozesu baten ondoren ezagutza lortzeko eta lortutako ezagutza erabili ahal izango da baliabideak erabiltzaileari egokitzeko zein zerbitzu-emaileari haren helburuetarako erabilgarri izango den informazioa helarazteko. Lan honetan, datu-meatzaritza prozesua eta ikasketa automatikoko tekniken ezaugarri orokorrak azalduko dira lehenengo eta jarraian, web ingurune batean metatutako datuetatik abiatu eta ezagutza lortzerainoko prozesua deskribatuko da.

## 2 **DATU-MEATZARITZA ETA IKASKETA AUTOMATIKOA**

### 2.1. **Datu-meatzaritza prozesua**

Datu-meatzaritzako (DM) prozesu batek beharrezko ditu hainbat urrats datu gordinetatik abiatu eta ezagutzaraino iristen den arte (ikus 1. irudia). Lehenik eta behin, tratatuko diren datuak aukeratu behar dira, gerta bailiteke jatorrizko datuak gehiegi izatea, haietako batzuk baliagarri ez izatea, etab. Tratatuko diren datuak aukeratzeko, laginketa erabiltzen da. Hurrengo urratsa datuak aurreprozesatu, informazioa gehien ematen duten aldagaiak edo datuen partea aukeratu eta ikasketa automatikoko tresnentzako egoki diren modu eta formatuan prestatzea izango da.




**1. irudia.** Datu-meatzaritza prozesu osoa deskribatzen duen eskema.

Datuak prest izanik ebatzi beharreko problemarentzat egokia den ikasketa automatikoko algoritmoa aukeratu eta hura erabiliz, ezagutza emango digun eredua sortzea izango da hurrengo urratsa. Eredu hau balioztatu egin beharko da azkenik bertatik ateratako ezagutza fidagarria izan dadin. Datu-meatzaritza prozesuak hainbatetan berrelikadura eskatzen du. Hau da, posible da lortutako lehenengo eredua balioztatzerakoan akatsak detektatzea eta berriro atzera buelta egin behar izatea eredu berri bat lortzeko; datuak egokituz edo/eta ikasketa automatikoko algoritmoa edo parametroak aldatuz.

## 2.2. **Ikasketa automatikoa**

Ikasketa automatikoko teknikak Adimen Artifizialeko (AA) teknikak dira, datu-baseetatik ezagutza erabilgarria lortzea helburu dutenak. Hau da, datu-basean galdera soilak eginez lortu ezin den informazioa lortzea helburu dutenak, oro har, estrategikoki garrantzitsua den informazioa.

Teknika hauen oinarria batez ere estatistika da; ondoko ideia hain zuzen ere: «probableena, lehenago ala beranduago, beti gertatu egin ohi da». Haien erabilera berriz, esparru anitzetan zabaldu da. Medikuntzan adibidez diagnosirako laguntza tresna gisa erabili izan dira, edo farmaziako laguntzaile gisa botika eraginkorrenak aukeratzeko. Oso ohikoa den beste aplikazio arlo bat irudiak oinarri gisa hartuz patroien errekonozimendu automatikoa egitea da (*pattern recognition*); objektuak, aurpegiak, hatzmarkak, edukiak eta testuak ala karaktereak izan daitezke adibide zabalduenetako batzuk. Meteorologian ere eguraldi iragarpenetarako zein klima aldaketaren azterketarako ohikoa da horrelako tekniken erabilera. Gaur egun hainbeste kezka sortzen dizkigun finantzen munduan ere erabili izan dira iruzurra detektatzeko edo maileguen arrisku-maila iragartzeko adibidez. Datu-meatzaritzaren beste alor bat testu-meatzaritza (*text mining*) da. Honetan, ikasketa automatikoko teknikak erabiltzen dira testuen azterketa

eta haietatik ezagutza ateratzea helburu izanik. Adibidez: dokumentuak sailkatzea, sentimenduen analisia, testuak laburtzea. Azken adibide gisa, azken urteetan zabaldutako web-meatzaritza [1,2] aipatu liteke, ikasketa automatikoko teknikak erabiltzen dira web ingurunean dauden eta sortzen diren datuetatik ezagutza lortzeko. Azken kasu honetan mota desberdinetako ezagutza lor liteke; edukien ingurukoa eta erabiltzaileen ezaugarrien ingurukoak adibidez.

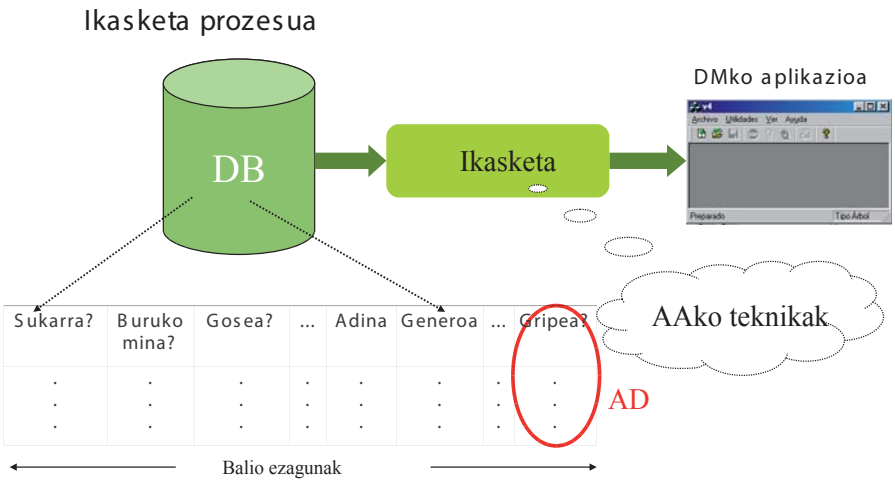
Ikasketa automatikoko teknikek bi urrats nagusi dituzte: ikasketa-prozesua eta esplotazio prozesua. Ikasketa-prozesuan, datu-basetik abiatu eta ikasketa automatikoko teknikak aplikatuz ereduak sortzen dira. Esplotazio prozesuan aldiz, eredu horiek testuinguru jakinetan erabiltzen dira.

Ikasketa automatikoko tekniken artean bi mota nagusi bereizten dira: gainbegiratutakoa eta gainbegiratu gabekoa. **Gainbegiratutako ikasketa**n ezagutzen da eredua eraikitzeko erabiltzen den adibide bakoitza aurrez definitutako zein kategoria edo klasetakoa den eta prozesuaren helburua, objektu edo fenomeno fisiko bati kategoria edo klase horien artean bat esleitzea da. Klasea adierazten duen aldagaiari Aldagai Dependente (AD) esan ohi zaio. Klase desberdinetako adibide kopurua oso desorekatua badago ikasketa prozesua zailagoa izan ohi da eta baliteke datuen aurreprozesaketa egiterakoan haietako batzuk ezabatu behar izatea. Gainbegiratutako ikasketako algoritmoek sortzen dituzten ereduen konplexutasuna problemaren zailtasunaren araberakoa izango da.

**Gainbegiratu gabeko ikasketa**ko problemetan aldiz, ez da aurrez ezagutzen eredua eraikitzeko erabiltzen diren adibideak zein klasetakoak diren eta helburua egitura edo klase »naturalen» bilaketa da. Gainbegiratu gabeko tekniken artean nagusiena *clustering*a edo taldekatzea da eta bere helburua taldeak topatzea da talde berean dauden kasuak haien artean antzekoak izanik eta beste taldeetakoekin aldiz, ahalik eta desberdinen. Mota honetako algoritmoetan beharrezkoa izango da antzekotasun edo desberdintasun neurri bat kasuak haien artean konparatu ahal izateko.

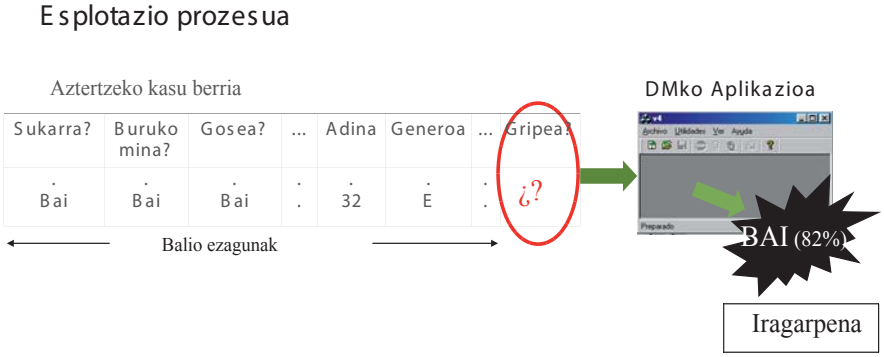
Gainbegiratutako ikasketako adibide bat erabiliko dugu ikasketa prozesua azaltzeko. Kasu horretan, ikasketa automatikoko teknikek sarrerako aldagaien eta Aldagai Dependente (AD) izeneko aldagai bereziaren arteko erlazioak bilatuko dituzte eta erlazio horiek sortutako ereduetan islatu. 2. irudian medikuntzan gripea diagnostikatzeko ikasketa automatikoko balizko aplikazio baten ikasketa-prozesuaren eskema ikus daiteke.

Abiapuntua pazienteak artatzerakoan medikuak egindako neurketak eta haien historia, aldagai-multzoa hain zuzen ere, eta kasu bakoitzaren diagnosia, aldagai dependentea, dituen datu-basea da. Ikasketa-prozesuaren ondoren, paziente berrien diagnosian lagunduko duen datu-meatzaritzako aplikazioa izango dugu. Aplikazio honek barnean edukiko ditu ikasketa automatikoko prozesuari esker sortutako ereduak.




**2. irudia.** Ikasketa-prozesua: diagnosia medikuntzan.

Sortutako datu-meatzaritzako aplikazioa etorkizunean erabiliko du medikuak paziente berriak artatzean. Medikuak sistemari adieraziko dizkio pazientea artatzean egindako neurketatan lortutako balioak eta haren historia eta sistemak aldiz, diagnosia egiten lagunduko dio medikuari. Esplotazio prozesuaren adibidea 3. irudian ikus daiteke. Paziente berria artatu ondoren, lortutako datuak datu-meatzaritzako aplikazioari sarrera gisa eman, eta honek emaitza gisa, pazienteak gaixorik egoteko duen probabilitatea itzuliko du. Erantzun hau noski, medikuaren esperientziarekin konbinatu beharko da azken erabakia hartzeko.

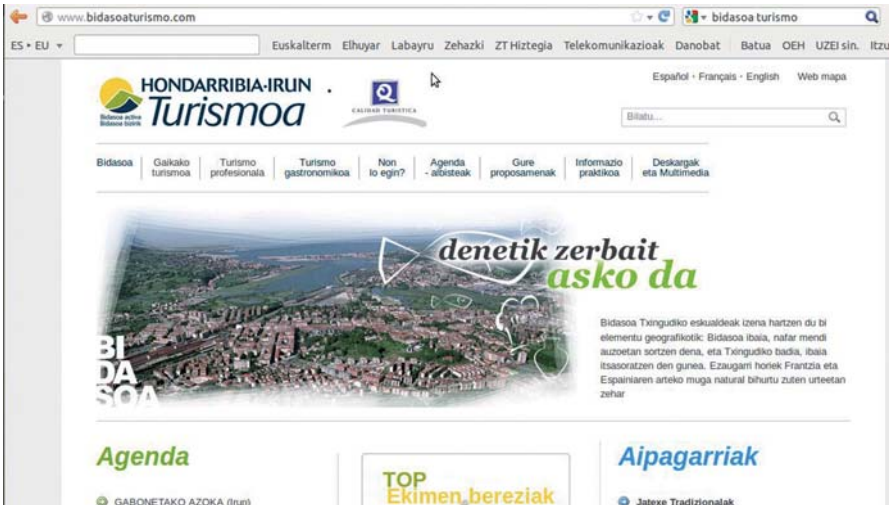



**3. irudia.** Esplotazio prozesua: diagnosia medikuntzan.

## 3. **ADIBIDE BAT: BIDASOA TURISMOKO WEBGUNEA**

Sarreran aipatu den moduan esparru askotan gerta daiteke interesgarri datu-meatzaritza prozesu bat ezagutza lortzeko bide gisa; web ingurune baten erabilera ere horietako bat da. Testuinguru honetan bi mota desberdinetako datuak erabili ahal izango ditugu: alde batetik, web orriek eta bertako edukiek osatutako multzoa eta bestetik, erabiltzaileek webgunean nabigatzean uzten duten aztarna.

Adibide gisa, Bidasoa Turismoko webgunean egindako datu-meatzaritza prozesua deskribatuko da (ikus 4 irudia). Bertako turismo bulegoko kideek 10 hilabetetako erabilera datuak helarazi zizkiguten eta Aldapa ikerkuntza taldean datu horiez gain webgunearen edukia, sarean atzigarri dagoena eta automatikoki eskuratu daitekeena, erabili genuen ikasketa automatikoko prozesu bat aurrera eramateko.




**4. irudia.** Bidasoa Turismoko webgunearen lehen orriaren irudia.

Datu horiek abiapuntu, ikasketa-prozesu bat baino gehiago da posible baina beti ere testuingurua gainbegiratu gabeko ikasketarena izango da; ez baitugu aurrez ezagutzen erabiltzaileen profilik ez eta webgunearen gaikako egitura.

Artikuluan aipatzen den kasuan, lehenik eta behin, erabilera datuetatik abiatuz, hau da, web zerbitzarian metatutako *log* fitxategietatik, erabiltzaileen nabigazio profilak eraiki genituen [3]. Horrez gain, testu-meatzaritzako teknikak erabiliz, webgunearen gaikako egitura lortu genuen eta azkenik, au-

rreko biak konbinatuz erabiltzaileen interesen profilak ezagutzea lortu genuen [4]. Lortutako ezagutza guztia Bidasoa Turismoko arduradunentzat oso interesgarri suertatu zen haiek proiektuan zehar adierazi zutenaren arabera.

### 3.1. **Erabilitako ikasketa automatikoko tresnak**

Jarraian Bidasoa Turismoko webgunearekin egin dugun datu-meatzaritza prozesuan erabili ditugun ikasketa automatikoko metodo nagusiak deskribatuko ditugu; taldekatze algoritmo bat, sekuentzia maizenen meatzaritzako algoritmo bat eta dokumentuak sailkatzeko beste bat azkenik. Hauek ez dira deskribatuko dugun prozesuan erabiliko diren teknika bakarrak, *K-means* izeneko taldekatze algoritmoa eta ikasketa gainbegiratuko K-NN algoritmoa ere erabiliko baitira, baina azken hauek askoz ezagunagoak izanik, haien deskribapena lan honetatik kanpo utzi da.

#### 3.1.1. *Taldekatze algoritmoa: PAM*

PAM (*Partitioning Around Medoid*s) [5] algoritmoa, Kaufman eta Rousseeuw-k (1990) proposatutako *clustering* edo taldekatze algoritmoa da eta datu-base bat hainbat taldetan zatituko du talde bereko kasuen gertutasuna (kohesioa) eta talde ezberdinetakoen urruntasuna bilatuz, hau da, talde berdineko adibideen arteko distantziak minimizatzen eta talde ezberdinen artekoak maximizatzen saiatuko da. Gisa honetako *clustering* algoritmoei algoritmo partiziogile deritze.

Edozein taldekatze algoritmo egikaritu ahal izateko beharrezkoa da kasuen arteko distantzia bat definitzea, kasuak taldetan banatzeko euren arteko gertutasuna eta urruntasuna ezagutu behar baita. Beraz, algoritmoari erabili behar duen distantzia zein den adierazi behar zaio. Lan honetan datuak sekuentzia mduan adierazi ditugu eta beraz, beharrezkoa dugu sekuentziak konparatzen dituen distantzia bat eta edizio distantzia (*edit distance* [6]) erabili dugu. *Edit* distantziak bi sekuentzien arteko distantzia kalkulatzen du, sekuentzia bat bestean bihurtzeko egin beharreko eragiketa (txertatu, ezabatu eta ordeztu) kopuru minimoan oinarrituz. Adibidez, «bidean» karaktere katea «kideak» sekuentzian bihurtzeko eragiketa minimoa 2 da, zehazki 2 ordezkatze, batetik hasierako b-a k-z ordeztu behar da eta bestetik azken n-a k-z ordeztu behar da. Artikulu honetan aztertzen ari garen adibidean arreta jarriz, *edit* distantziak erabiltzaile-saioei dagozkien klik sekuentziak hartu eta bata bestean bihurtzeko kostua adieraziko digu.

PAM algoritmoak beharrezkoa duen beste parametro bat datu-basea zenbat zatitan banatu behar duen da, K parametroa alegia. K parametroaren balio egokiena datuen egituraren menpe egon ohi da, K handiegiekin talde orokorregiak lortzeko arriskua dago eta K txikiegia bada aldiz, berariazkoegiak, beraz K-ren balioan oreka bilatzea garrantzitsua izango da.

PAM algoritmoa datu-baseko K kasu-ordezkari aukeratuz hasten da (hasieraketa). Lehenengo partizioa lortzeko gainerako kasuak gertueneko ordezkariari esleitzen zaizkio. Partizio baten kostua kasuek euren gertueneko ordezkarira duten distantzien batura izango da. Ondoko urratsetan (trukea) partizioa hobetzen saiatuko da algoritmoa, hau da, partizioaren kostua minimizatzen. Horretarako ordezkarien multzoa aldatuz joango da kostua gehiago txikitzea lortzen ez den arte (ikus 1. taula).

**1. taula.** PAM taldekatze algorimoa.

HASIERAKETA: K kasu ordezkari hautatu.

1. Hautatu lehen ordezkaria: datu-baseko kasu guztietara distantzien batura minimoa duena.
2. Hautatu gainerako ordezkariak ordezkarien arteko distantziak maximizatuz.
3. Adibide bakoitza gertuen duen ordezkarira lotu.

TRUKEA: K ordezkarien multzoa hobetu.

4. *Cluster* bakoitzeko adibide guztietarako aztertu ea adibide hori ordezkari gisa hautatuz gero, *cluster*reko adibide guztien ordezkariarekiko distantzien batura txikiagotuko ote litzatekeen. Existitzen bada, hautatu batura minimizatzen duen adibidea.
5. Gutxienez *cluster* bateko ordezkaria aldatu bada, joan 3. urratsera.

### 3.1.2. *Sekuentzia maizenen erauzketa: SPADE*

SPADE (*Sequential PAttern Discovery using Equivalence classes* [7] - sekuentzietan patroiak aurkitzea baliokidetasun klaseak erabiliz) Sekuentzia Maizenen Meatzaritza (*Frequent Sequence Mining*) arloko algoritmoa da eta sekuentzia multzo batek elkarren artean konpartitzen dituzten azpi-sekuentziak aurkitzea du helburu; hauen errepikapen kopuruan oinarritzen delarik. Algoritmo honi adierazi behar zaion parametroa sostengu minimoa (*minimum support*) da.

SPADE algoritmoa erosketa-orgatxoaren arazoa (*shopping cart problem*) ebazteko diseinatutako algoritmoa da, hau da, bezeroen erosketa sekuentzietatik haien ohiturei buruzko informazioa ateratzeko problematikari egiten dio aurre. Problema honetan, bezero bakoitza erosketa-orgatxo sekuentzia batez adieraziko da non orgatxo bakoitza produktuez beterik egongo den. Bezero baten erosketa-orgatxoen ordena kontuan hartzen du algoritmoak, denboran zehar bezeroak izan duen bilakaera erakusten duelako, orgatxoko produktuen ordena aldiz, ez du kontuan hartzen, problema honetan garrantzirik ez duelako.

Ondorengo adibidea SPADE algoritmoak egiten duenaren azalpena izan liteke. Datu-baseak 4 sekuentzia ditu. Sekuentzia horiek lau beze-

roren erosketen garapena adierazten dute. Sekuentziak orgatxo segidak izango lirateke eta orgatxo bakoitza item-multzo baten bidez adierazten da. Datu-basean zortzi item ageri dira (A, B, C, D, E, F, G, H). Horrela, [(C,D) ->(A,B,C)->(A,B,F)->(A,C,D,F)] sekuentziak zera adieraziko luke. Bezero bat lau bider joan dela erosketetara eta lehenengoan C eta D produktuak erosi zituela, bigarrenean, A, B eta C produktuak, hirugarrenean, A, B eta F eta azkenik laugarrenean A, C, D eta F. Parentesi artean dauden item edo produktuen ordenak ez du garrantzirik baina erosketen arteko ordenak aldiz bai.

Demagun sekuentzien datu-basea ondokoa dela,

[(C,D)->(A,B,C)->(A,B,F)->(A,C,D,F)]
[(A,B,F)->(E)]
[(A,B,F)]
[(D,G,H)->(B,F)->(A,G,H)]

eta SPADE algoritmoari eskatzen dioguna sekuentzia hauetatik gutxienez % 50eko sostengua (*minimum suppor*t) dutenak aukeratzea dela. Kasu honetan, 4 sekuentzia daudenez 2 sekuentzia edo gehiagotan errepikatzen diren sekuentzia atalak itzuliko ditu algoritmoak. Atal horiek, luzeraren arabera sailkatuta ondokoak izango lirateke (sekuentzia bakoitzaren alboan dagokion errepikapen kopurua ageri da):

1 luzerako sekuentziak: [(A)], 4; [(B)], 4; [(D)], 2; [(F)], 4;
2 luzerako sekuentziak: [(A,B)], 3; [(A,F)], 3; [(B)->(A)], 2; [(B,F)], 4; [(D)->(A)], 2; [(D)->(B)], 2; [(D)->(F)], 2; [(F)->(A)], 2;
3 luzerako sekuentziak: [(A,B,F)], 3; [(B,F)->(A)], 2; [(D)->(B,F)], 2; [(D)-> (B)->(A)], 2; [(D)->(F)->(A)], 2;
4 luzerako sekuentziak: [(D)->(B,F)->(A)], 2;

Algoritmoak ohikoenak diren azpi-sekuentziak erauzi ditu datu-basetik.

Gure lanean SPADE algoritmoa erabili dugu *cluster* edo saio multzo bakoitzeko URL esanguratsuenak ateratzeko. Horretarako SPADE algoritmoari sarrera gisa eman diogu *cluster* bakoitzeko sekuentzia multzoa. Problema ez da erosketa-orgatxoaren problemaren erabat berdina, kasu partikular bat dela esan genezake, baina algoritmo hau bera erabil liteke saio bakoitzeko osagai diren URLak item bakarreko erosketa gisa adieraziz.

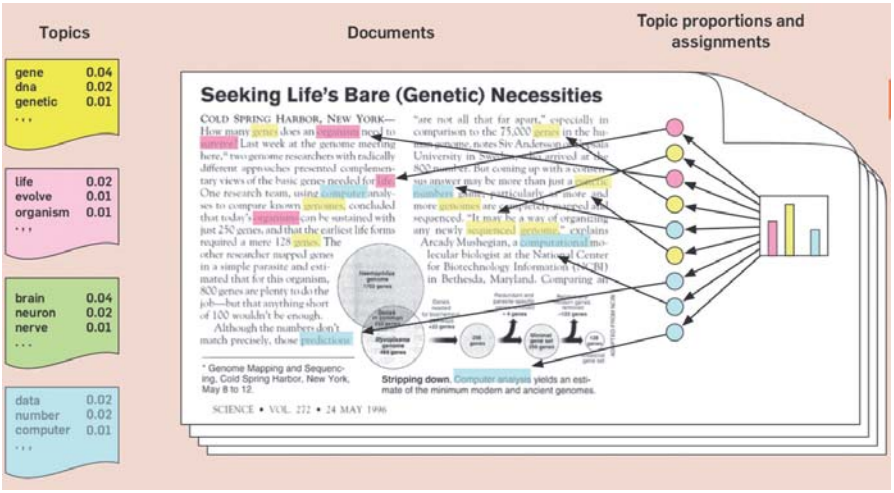
### 3.1.3. *Topic modeling*

*Topic modeling*-a testu-meatzaritzako arlo bat da eta korpus batean patroiak identifikatzeko modu bat. Azken aldian eredu estatistiko ugari garatu dira automatikoki dokumentu-multzo handien gaikako egitura lortzeko [8]. *Topic modeling*-a ikasketa automatikoan eta lengoaia naturalaren prozesa-

menduko tekniketan oinarritzen den eredu estatistiko bat da dokumentu-multzo handi batean agertzen diren gai abstraktuak aurkitzen dituena.

*Latent Dirichlet Allocation* (LDA) deritzona, *topic modeling* eredu erabiliena da. LDA dokumentu multzo baten gaikako egitura bilatzen duen eredua da, estatistikan oinarritzen dena. Dokumentuok hainbat gai edo gai bakarra jorratu ditzakete. Algoritmo hauek dokumentuak gai nahasketa moduan interpretatzen dituzte, non gai bat, hemendik aurrera *topic* deituko dioguna, testuetako hitz-gako (*keyword*) multzo batez osatzen den. Algoritmo mota honek, dokumentu sorta bat hartzen du sarrera gisa eta irteera moduan *topic* zerrenda bat osatzen du. *Topic* bakoitza hitz-gako multzo baten bidez adierazten du eta horrez gain algoritmoak irteera gisa ematen du testu bakoitzak *topic* bakoitzarekin duen erlazioaren pisua edo afinitate bektorea. Behin dokumentu-*topic* afinitate bektorea lortuta, dokumentu bakoitzaren edukiaren gaia zein den ondoriozta daiteke (hainbat *topic* edo bakarra). Informazio honek hain zuzen dokumentu ezberdinak euren artean alderatu eta gaikako egitura batean biltzeko balioko du.

*Topic modeling*-ak nola funtzionatzen duen azaltzeko adibide argigarria, artikulu batekin eta markatzaile batzuekin lan egiten ari garela imajinatzea izan liteke. Artikulua irakurtzen goazen heinean, kolore desberdineko markatzaileak erabiltzen ditugu artikuluaren gai desberdinen hitz-gakoak nabarmentzeko. Irakurtzen bukatzen dugunean, gai izango ginateke kolore desberdinetan markatu ditugun hitzak multzokatzeko. Kolore desberdineko hitz multzo bakoitza gai zehatz baten ingurukoa izango da eta kolore bakoitzak *topic* desberdin bat adieraziko du beraz (ikus 5. irudia).




**5. irudia.** *Topic modeling*-en adibide baten irudia. *«Probabilistic topic models»* [8] lanetik ateratakoa.

### 3.2. **Nabigazio profilak**

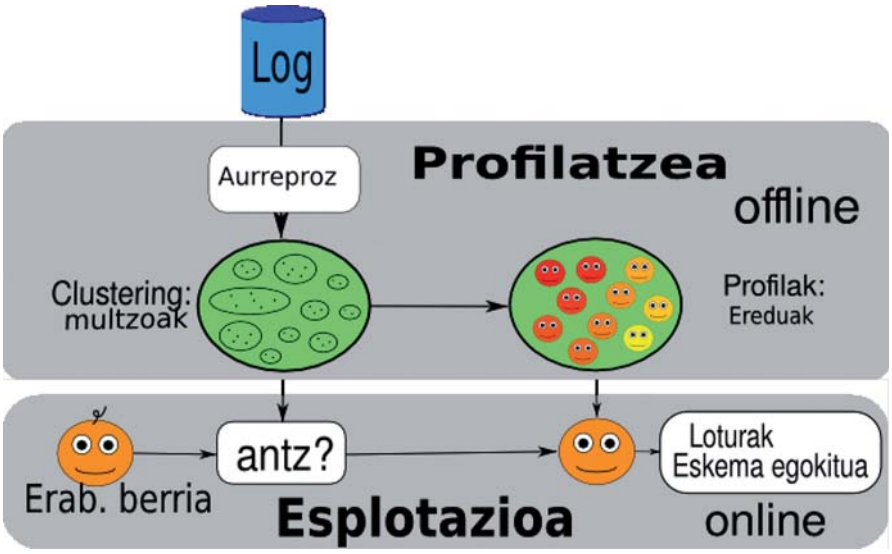
Abiapuntua web zerbitzariko *log* fitxategiak dira, zeinetan erabiltzaileek zerbitzariari egiten dioten orri eskaera orori buruzko informazioa metatzen den. Hau da, erabiltzaileak egiten duen klik bakoitzeko informazioa metatzen da bertan. 2. taulan ikus liteke *log* fitxategi hauetako baten adibidea non erabiltzailearen IP helbidea, atzipeneko data eta ordua, atzitutako orria eta kode bat (eskaera zuzen ala oker burutu den adierazten duena) ageri diren.

**2. taula.** *Log* fitxategiaren adibide bat.

```
66.249.71.104 [09/Jan/2012:04:04:01 +0100] "GET /index.php?option=com_content&view=ar" 500
213.190.9.153 [09/Jan/2012:04:04:02 +0100] "GET /templates/pie_abajo_v2/error500.php HTTP/1.1" 403
194.30.43.2 [09/Jan/2012:10:35:59 +0100] "GET /templates/pie_abajo_v2/images/bienvenida_es.jpg" 200
82.130.196.242[09/Jan/2012:10:20:39 +0100] "GET /plugins/content/loaded/41_Get_Flash_Player.jpg" 200
158.227.96.228 [09/Jan/2012:10:35:59 +0100] "GET /templates/pie_abajo_v2/video/bienvenida_es.avi" 200
66.249.71.104 [09/Jan/2012:04:57:01 +0100] "GET /index.php?view=details&id=S01DAG&lang=en" 200
157.55.16.87 - - [09/Jan/2012:04:58:18 +0100] "GET /index.php?&view=details&el_mcal_month=12" 200
```

*Log* fitxategiak aurreprozesatzea da lehen urratsa. Eskaera okerrak eta erabiltzaile klikei ez dagozkienak ezabatzea izango da lehenengo egin beharreko lana eta ondoren, erabiltzaileak eta erabiltzaile-saioak identifikatu beharko dira [9,10]. Erabiltzaile-saioa erabiltzaile batek, IP helbide jakin baten bidez identifikatuko duguna, webgunean zehar nabigatzen duenean atzitzen dituen orriez dago osatuta. Hau da URL helbide sekuentzia gisa adierazten dugu. Lortutako datu-basea beraz, erabiltzaile-saio multzoa izango da. Hurrengo urratsa izango da ikasketa automatikoko teknikak erabiltzea erabiltzaileen nabigazio profilak topatzeko. Lan hau bi urratsetan banatu dugu. Lehenengoan PAM (*Partitioning Around Medoids*) [5] algoritmoa eta *edit distance* [6] izeneko distantzia erabiliz antzeko nabigazio ezaugarriak dituzten erabiltzaile-saioak batu ditugu multzotan. Bigarren urratsean aldiz, multzo bakoitzerako profil bana eraiki dugu; multzo bakoitzean ohikoen diren URL sekuentziez osatutakoa hain zuzen ere. Horretarako SPADE (*Sequential PAttern Discovery using Equivalence classes*) [7] ohiko diren sekuentziak bilatzen dituen ikasketa automatikoko algoritmoa erabili dugu eta hari esker ondoko irteera lortu da: *cluster* edo talde bakoitzeko, taldea osatzen duten saioetan bisitatua izateko probabilitate handia duen URL helbide multzoa. Hauek izango dira, erabiltzaileen nabigazio profilak eta prozesuari profilatzea esango zaio. 8 Irudiko goiko parteak laburtzen du prozesu hau guztia.

Erabiltzaile profil hauek lagungarri izango dira erabiltzaile berrien nabigazioa errazteko. 6. irudiko beheko partean ageri den moduan, erabiltzaile berriak webgunean nabigatzen hasi ahala, esplotazioa esango diogun atalean, haien antzekoenak diren multzoak detekta daitezke. Gure sistemaren kasuan *K-Nearest Neighbour* (K-NN) [11] gainbegiratutako ikasketako



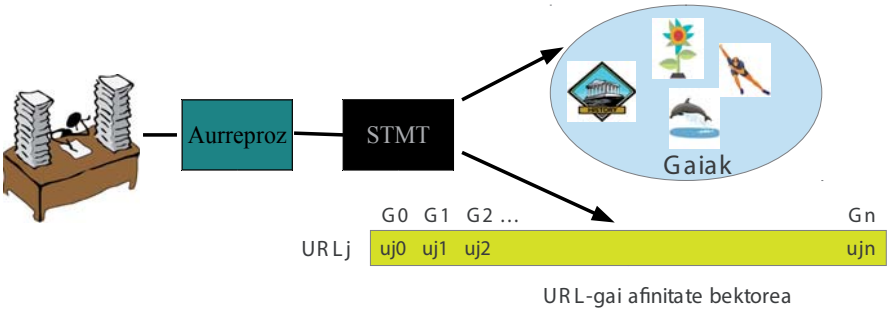

**6. irudia.** Nabigazio profilak sortu eta webgunea egokitzeko erabiltzearen prozesua.

teknika eta *edit distance* erabili ditugu erabiltzaile berrien nabigazioaren sekuentzia eta multzoen arteko distantzia kalkulatzeko.

Antzekoena den taldea topatu ostean, haren profila erabili ahal izango dugu erabiltzaile berriari nabigatu ahala interesgarri gerta dakizkiokeen loturak proposatzeko edo eskema egokitu eta haiek nabarmentzeko.

### 3.3. **Interes profilak**

Interes profilak eraiki ahal izateko erabiltzaileak webgunean nondik nora dabiltzan jakiteaz gain, beharrezkoa da web orri edo URL bakoitza gai batekin lotzea. Webguneak modu egokian diseinatuta daudenean, web orriak modu berezian metatu ohi dira eta lantzen dituzten gaien inguruko informazioa etiketen bidez atxikia izaten dute; Bidasoa Turismoko webgunean eta webgune gehienetan aldiz, datu-basea fitxategi multzo soila besterik ez da. Beraz, erabiltzaileen interes profilak lortu ahal izateko beharrezkoa izango da lehenago bertan pilatzen diren dokumentuen gaikako egitura lortzea. Azkenaldian eredu estatistiko andana garatu da dokumentu multzo handien gaikako egitura lortzeko [8]. Lan honetan erabili duguna *Latent Dirichlet Allocation* (LDA) [12] oinarritzen den eta *topic modeling* izenez ezagutzen den eredua izan da. *Stanford Topic Modeling Toolbox* (STMT) [13] izeneko tresna erabili dugu hain zuzen ere Bidasoa Turismoko webguneko dokumentuetan ezkutatuta dauden gaien egiturari buruzko informazioa lortzeko.



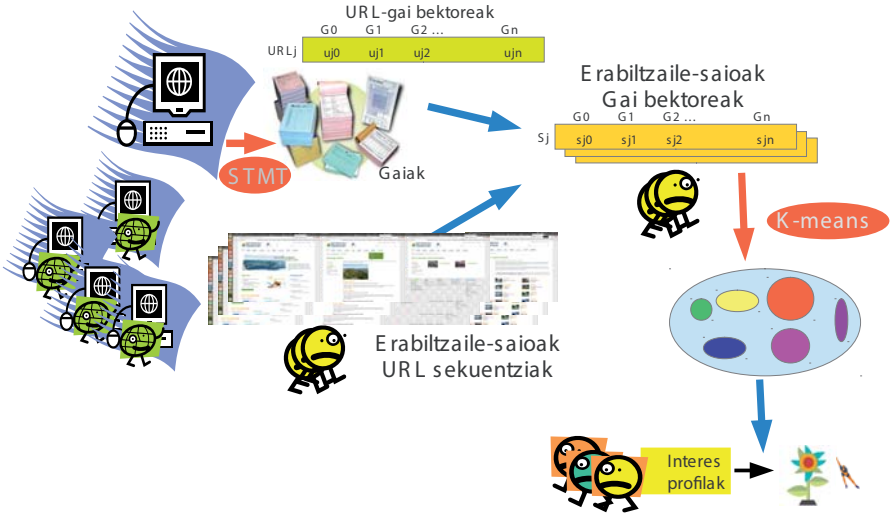

**7. irudia.** Webgune baten gaikako egitura lortzea STMT tresna erabilz.

STMT tresnak sarrera gisa webguneko URL bakoitzeko fitxategiaren testu edukia hartu eta irteera gisa lortutakoa zera da: gai zerrenda bat, hitz gakoen bidez adierazia eta URL helbide bakoitzak gai horietako bakoitzarekin duen afinitatea (URL-gai afinitate-bektorea). 3. taulan ageri da prozesu honen eskema.

**3. taula.** Bidasoa Turismoko webgunearen gaikako egitura.

| Etiketak | Hitz gakoen zerrenda |
|---|---|
| itsasoa eta kirola (Sea & Sport, SS) | sea, surf, boat, sport, marina, kayak, canoe... |
| ostatua kanpina (Accomodation Camping, AC) | accomodation, pilgrim, camp, hostel, youth, guesthouse, holiday... |
| ostatua hotela (Accomodation Hotel, AH) | room, countryside, obispo, goikoerrota, jauregui, martindozenea ... |
| ekitaldiak (Events, Ev) | events, festival, music, exhibition, theater, film, history, popular... |
| tradizioa (Tradition, Tr) | tradition, marcial, typical, basque, celebration, cuisine, guadalupe... |
| kultura (Culture, Cl) | culture, organization, visitor, amaia, artist, exhibition... |
| natura (Nature, Na) | bay, mountain, river, rout, path, beach, region, nature... |
| gastronomia(Cuisine, Cu) | cuisine, restaurant, zuloaga, tradition, gourmet, ciderhouse, bar, eat... |
| kirolak (Sport, Sp) | sport, pelota, tennis, ride, golf, fronton... |
| monumentu historikoak (Historical Monuments, HM) | roman, century, building, church, chapel, castle, museum... |

Bidasoa Turismoko webgunearen kasuan, 10 gai nagusi identifikatu dira eta emaitzak ulerterrazagoak egiteko, gaiak eskuz izendatu ditugu izenak erabakitzeko gai bakoitzari esleitutako hitz-gakoen esanahia oinarri gisa hartuz. Sistemak itzulitako gaikako egitura 8. irudian ageri da. Hitz




**8. irudia.** Interes profilak lortzeko prozesu osoaren eskema.

gakoak ingelesez ageri dira tresna estandarrak erabili ahal izateko webgunearen ingelesezko bertsioarekin lan egin dugulako baina lan hau berdin egin daiteke beste edozein hizkuntzatan. Lortutako gaikako egitura Bidasoa Turismoko arduradunekin eztabaidatu dugu eta haien buruan zegoen egiturarekin bat datorrela adierazi digute. Nola nahi ere erabiltzaileen interes profilak lortzerakoan, ostatuari dagozkionak, AC eta AH, bazterrean utzi ditugu, interesa baino gehiago beharra adierazten dutelakoan.

Informazio berri hau izanik, URL-gai afinitate bektorea erabil dezakegu URL bakoitza gai batekin erlazionatzeko, eta ondorioz, lehen URL sekuentzia moduan adierazita genituen erabiltzaile saioak, sekuentziako URL guztiei dagozkien URL-gai afinitate bektoreak konbinatuz, gaiekin duten afinitatearen arabera adieraz daitezke orain, saio-gai afinitate bektore gisa hain zuzen ere. Hau da, datu-base berri bat izango dugu, non saio bakoitza bektore baten bidez adieraziko den, gai bakoitzarekin duen afinitatea adierazten duena.

Datu-base berri honetatik, ikasketa aumatikoko algoritmoen bidez lortu ahal izango ditugu erabiltzaileen interes profilak. Horretarako, *K-means* [14] taldekatze algoritmoa eta *Hellinger* distantzia [15], testu konparaketan erabili ohi den distantzia bat, erabiliz antzeko interesak dituzten erabiltzaile saioak multzo berean batuko dira. Eta ondoren multzo bakoitzean bildutako URL-gai afinitate bektoreak aztertuz bertako erabiltzaileen interes profila lortuko dugu. Lehenengo hurbilketa honetan gai bakarra hautatu dugu multzoko; afinitate handiena duena hain zuzen ere. Interes profilak

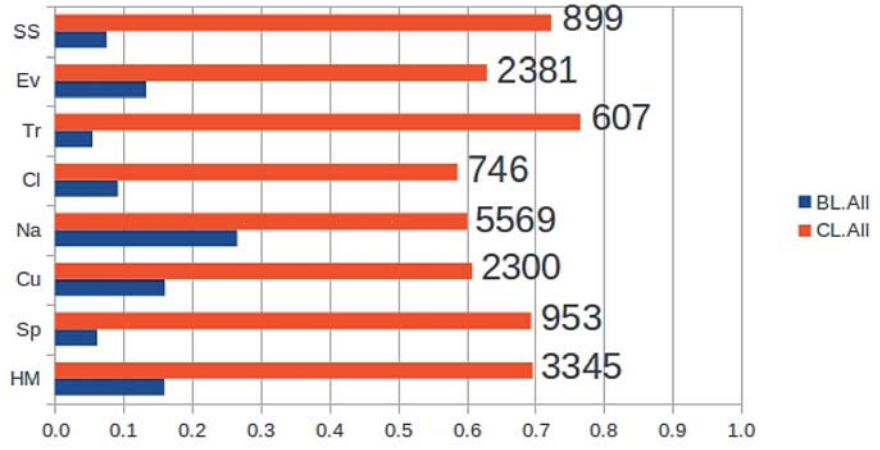
lortzeko prozesu osoa 8. irudian laburbiltzen da. Bertan ikus daiteke, abiapuntua web orrien edukia eta web zerbitzarietan metatutako *log*-ak izanik, datu-meatzaritza prozesu bati esker erabiltzaileen interes profilak lor daitezkeela.

## 3.4. **Interes profilek emandako ezagutza**

### 3.4.1. *Profil globalak*

Lortutako interes profilek webgunearen administratzaileei ekar diezaiekeen onura ebaluatzeko ondoko balioak konparatu dira: datu-base osoko erabiltzaileen gaiekiko afinitateak alde batetik eta *clustering*a eta profilak lortzeko prozesuaren ondoren lortutako afinitate mailak bestetik. Lehenengo kasurako gai bakoitzaren afinitate maila datu-baseko saio guztiak kontuan izanda kalkulatu da. Bigarrenerako aldiz, gai jakin horrekin etiketatutako taldeetako erabiltzaileak soilik hartu dira kontuan. Interes profilak lortu ondoren, gai batekin etiketatutako *cluster*rentzat zera suposatu dugu: *cluster* horretan bildu diren erabiltzaile guztiek interesa dutela gai horretan eta erabiltzaile horientzat afinitate maila *cluster*rarentzat lortu den afinitate mailaren berdina dela.

9. irudian ageri dira, interes profilak topatzeko erabili ditugun 8 gaietarako, bi aukerekin lortutako afinitate mailak (datu-base osoarekin BL.ALL eta *clustering* bidez lortutako profilak lortu ondoren CL.ALL). Y ardatzean 8 gai nagusiak ageri dira eta afinitate mailak aldiz, X ardatzean. Taldekatzeari esker lortu diren profiletan oinarritzen den aukerarako, informazio gehigarri gisa interes handiena gai horretan duten saio kopurua ageri da.




**9. irudia.** Afinitate mailen konparaketa: datu-base osoa vs. lortutako profilak.

9. irudiak argi uzten du datu-base osoaren profila kontuan hartzen badugu, lortzen diren afinitate mailak oso baxuak direla (altuena % 25aren ingurukoa da) eta beraz, ez ginatekeela gai izango esateko erabiltzaileen interesak zein gaien ingurukoak diren. C*lustering* bidez lortutako profiletatik lortutako balioak erabiltzen baditugu aldiz, afinitate mailak izugarri hazten dira. Adibidez, badago 607 saio biltzen dituen talde bat tradizioarekin (Tr) ia % 80ko afinitatea duena, edo, beste muturra aztertuz gero, 5569 saioko multzoa naturarekin (Na) duen afinitatea % 60koa dena.

Konparaketa honek nabarmen egiten du interes profilak bilatzeak onura dakarrela, hau da, haiei esker gai konkretuetan interes handia duten erabiltzaile multzoak identifikatzeko gai garela. Lorpen hau, etorkizuneko pertsonalizazio eta marketin estrategietan erabili ahal izango da.

### 3.4.2. *Jatorriaren araberako profilak*

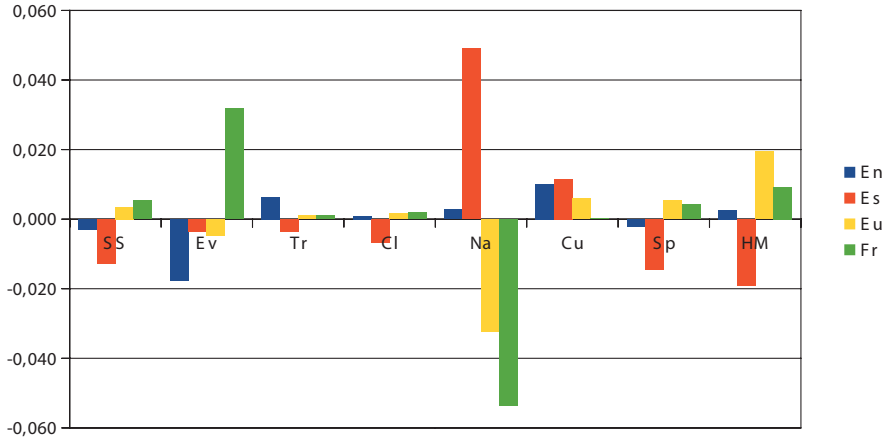
Bidasoa Turismoko webgunea 4 hizkuntza desberdinetan atzi daiteke: euskaraz, gaztelaniaz, frantsesez eta ingelesez. Atzipenerako erabilitako hizkuntza jatorriaren araberakoa izango da, denek baitugu ohitura gure ama hizkuntzan nabigatzeko. Beraz, atzipen hizkuntzaz baliatu ahal izango gara jatorri desberdinetako erabiltzaileei buruzko informazioa eskuratzeko eta ondorioz, jatorri desberdineko erabiltzaileen ezaugarrien berri izateko. Informazio hau oso baliagarria izango da turismo bulegoko arduradunentzat.

Helburu horrekin web zerbitzaritik eskuratutako *log* datu-basea erabiltzaileak webgunea atzitzeko erabilitako hizkuntzaren araberako 4 zati desberdinetan banatu da eta aurreko atalean deskribatutako modu berean, hizkuntza bakoitzeko erabiltzaileentzat interes profilak lortu dira.

Lortutako profilei esker jakin ahal izango dugu ba ote dagoen jatorri berdineko erabiltzaileen artean patroi komunik. Horrez gain, patroi hauek beste jatorri batzuetako erabiltzaileen patroiekin eta patroi globalekin konparatu eta haien arteko diferentziak topatzeko aukera ere izango dugu. Jatorri bakoitzeko erabiltzaileentzat egitura argia topatuz gero, lortutako ezagutzak bi aplikazio alor garrantzitsu edukiko lituzke: alde batetik erabiltzaileen jatorriaren araberako adaptazio desberdinak proposatu ahal izango lirateke eta, beste aldetik, zerbitzu emaileari informazio garrantzitsua emango lioke etorkizuneko marketin kanpaina edo webaren diseinu berrietan aplikatu ahal izateko.

10. irudiak erakusten ditu datu-base osoarekin lortutako profilak eta hizkuntzaka lortutako profilaren arteko konparaketa. Balioak lortzeko *clustering*a egin ondoren, gai bakoitzeko afinitate maila eta saio kopurua balio bakar batean konbinatu dira. Horretarako, datu base bakoitzean lortutako afinitate balioak eta erabiltzaile kopuruak (estaldura) balio tarte berean egon daitezen normalizatu ditugu. Aukeratutako normalizazioa haietako bakoitzean balio guztien batura 1 egiten duena izan da. Gero, *topic* bakoi-

tzerako bi balio normalizatuak, afinitatea eta estaldura, batu eta haien arteko bataz bestekoa kalkulatu dugu. Azkenik, grafikoan adierazten dena, hizkuntzaren menpeko profilentzat eta profil globalentzat lortutako balioen arteko kendura da. Horrela, hizkuntza bakoitzean lotutako profila konparatu dugu profil globalarekin. Balio positiboek hizkuntza jakin batean lortutako afinitate maila sistema globalean lortutakoa baino altuagoa dela adierazten dute eta negatiboek aldiz, afinitate maila baxuagoa.




**10. irudia.** Hizkuntza desberdinetako profilek profil globalekin duten aldea.

10. iruditik atera dezakegun lehenengo ondorioa da, interesei dagokienean, jatorri desberdineko erabiltzaileen artean desberdintasunak daudela. Irudiaren arabera, Ingelesez nabigatzen duten erabiltzaileak (En), ez daude oso urrun profil globaletik baina antza denez, haiek baino interes handiagoa adierazten dute gastronomia (Cu) eta tradizio (Tr) gaietan eta interes gutxiago aldiz, ekitaldietan (Ev). Turista espainiarrek aldiz, profil orokorretik aldenduago dirudite; natura (Na) gaian interes handiagoa dutela dirudi, gastronomia (Cu) ahaztu gabe noski, eta gainerako gai guztietan interes maila txikiagoa nabarmentzen zaie. Turista euskaldunen kasuan interesgune nagusia monumentu historikoak (HM) eta itsasoa eta kirolak (SS and Sp) gaietan ageri da. Turista euskaldunak antza denez gutxi erabiltzen dute webgunea naturarekin lotutako gauzen berri izateko. Azkenik, turista frantsesek nabarmenki ekitaldiak (Ev) dituzte gustuko. Gainerakoetan interes baxuagoa adierazten dute eta bereziki txikia naturan (Na).

Automatikoki lortutako ondorio hauetako batzuk bat etorri dira Bidasoa Turismoko arduradunen esperientziarekin eta besteak berriz, oso interesgarri iruditu zaizkie etorkizunean hartuko diren erabakietan kontuan hartzeko.

## 4. **ONDORIOAK**

Artikulu honetan azaldu denez, datu-meatzaritza eta ikasketa automatikoa erabiliz, metatutako informaziotik ezagutza lor daiteke era automatikoan. Beraz, teknika hauek erabilgarri izango dira esparru anitzetan metatutako informazio guztiari etekin handiagoa ateratzeko.

Teknika hauen ezaugarri eta aukera orokorrez hitz egiteaz gain, txosten honetan datu-meatzaritza prozesuaren adibide konkretu bat deskribatu da; Bidasoa Turismoko webguneari aplikatutako datu-meatzaritza prozesuaren deskribapena hain zuzen ere. Webgunearen edukitik eta erabiltzaileek nabigatzean utzitako arrastotik abiatuz datu-meatzaritza prozesuari esker hainbat lorpen egin dira: erabiltzaileen nabigazioa aurreikusi da, webgunearen egitura semantikoa ondorioztatu da, interes konkretuak dituzten erabiltzaile multzoak identifikatu dira eta hizkuntza desberdinetan aritzen diren erabiltzaileen interesak desberdinak direla jakin ahal izan da. Ezagutza honi esker aukera izango dugu webgunea erabiltzaile berriei egokitzeko haien nabigazioa errazagoa izan dadin, etorkizuneko diseinuetan hizkuntza kontuan hartzeko, marketin kanpainetan eraginkorragoak izateko, etab.

## 5. **BIBLIOGRAFIA**

[1] MOBASHER B. 2006. 12 *Web Usage Mining*. *Encyclopedia of Data Warehousing and Data Mining*. Springer, Berlin.

[2] SRIVASTAVA T., DESIKAN P. eta KUMAR, V. 2005. «Web Mining - Concepts, Applications and Research Directions». *Foundations and Advances in Data Mining* 275-307.

[3] ARBELAITZ O., GURRUTXAGA I., LOJO A., MUGUERZA J., PEREZ J.M. eta PERONA I. 2012. «Enhancing a Web Usage Mining based Tourism Website Adaptation with Content Information». *Proceedings of the 4th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2012) & 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR)*. 287-292.

[4] ARBELAITZ O., GURRUTXAGA I., LOJO A., MUGUERZA J., PEREZ J.M. eta PERONA I. 2013. «A Navigation-log based Web Mining Application to Profile the Interests of Users Accessing the Web of Bidasoa Turismo». *e-Review of Tourism Research (eRTR). Conference on Information and Communication Technologies in Tourism (ENTER*): Short Papers.

[5] KAUFMAN L. eta ROUSSEEUW P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.

[6] GUSFIELD D. 1997. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press.

[7] ZAKI, M.J. 2001. «Spade: An efficient algorithm for mining frequent sequences». *Machine Learning* **42,** 31-60.
[8] BLEI D.M. 2011. «Probabilistic Topic Models». *Communications of the ACM 55* , 77-84.
[9] COOLEY R., MOBASHER B. eta SRIVASTAVA J. 2009. «Data Preparation for Mining World Wide Web Browsing Patterns». *Knowledge and Information System* 5-32.
[10] HE D. eta GÖKER A. 2000. «Detecting session boundaries from Web user logs». *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*.
[11] DASARATHY S. 1991. *Nearest neighbor (NN) norms : NN pattern classification techniques*. IEEE Computer Society Press.
[12] BLEI D.M., NG A.Y. eta JORDAN M.I. 2003. «Latent Dirichlet allocation». *Journal of Machine Learning Research* **3**, 993-1022.
[13] Stanford Topic Modeling Toolbox http://nlp.stanford.edu/software/tmt/tmt-0.2/
[14] LLOYD S. P. 1982. «Least squares quantization in PCM». *IEEE Transactions on Information Theory* **28**, 129-137.
[15] DEZA E. eta DEZA M. M. 2006. *Dictionary of distances*. Elsevier, Amsterdem.