



IKER
GAZTE
NAZIOARTEKO
IKERKETA EUSKARAZ

II. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2017ko maiatzaren 10, 11 eta 12
Iruñea, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

UPV/EHUko eZerbitzu baten
modelatzea ikasketa
automatikoaren bidez

*Ainhoa Yera, Iñigo Perona,
Olatz Arbelaiz eta Javier Muguerza*

111-118 or.
<https://dx.doi.org/10.26876/ikergazte.ii.03.16>

ANTOLATZAILEA:



ELKARLANEAN:



LAGUNTZAILEAK:



UPV/EHUko eZerbitzu baten modelatzea ikasketa automatikoaren bidez

Yera, Ainhoa; Perona, Iñigo; Arbelaitz, Olatz; Muguerza, Javier.

*Euskal Herriko Unibertsitatea (UPV/EHU). Konputagailuen Arkitektura eta Teknologia saila. Aldapa.
ainhoa.yera@ehu.eus*

Laburpena

Lan honetan UPV/EHUko matrikulazioaren inguruko eZerbitzuko nabigazioa aztertu dugu web-meatzaritzako tekniken bidez, nabigazio-saioen sailkapena era automatikoan egiten saiatu gara. Emaizten arabera, baieztatu dugu maila handi batean definitutako arrakasta / porrota nabigazio-portaerak detektatzeko gai garela. Esaterako, arrakasta gisa etiketatu diren klusterretan dauden saioren % 90etik gora arrakasta motakoak dira eta porroten kasuan % 90 inguru. Gainera, gainbegiraturako ikasketa bidez, bi saio-motak era automatikoan desberdintzeko gai gara % 96ko asmatze-tasa batekin. Beraz, lan hau etorkizunean eZerbitzu hau hobetu ahal izateko oinarri ona dela uste dugu.

Hitz gakoak: eGobernua, eZerbitzuak, web erabilpenaren meatzaritza, nabigazio-ereduak.

Abstract

In this work we have analyzed the enrollment eService navigation of the UPV/EHU and using data mining techniques we have attempted to automatically perform navigation sessions classification. The results show that we are able to detect the defined success and failure navigation behaviours. For example, more than 90 % of the sessions of the clusters labelled as success are of success type and in the failure case, around 90%. Besides, using supervised learning we are able to automatically distinguish the two nabigation types with an accuracy rate of 96 %. Thus, we think that this research is a suitable basis to improve the eService analyzed in a near future.

Keywords: eGovernment, eServices, web usage mining, navigation models.

1. Sarrera eta motibazioa / momentuko egoera

eGobernua lehen mailako joera bilakatu da informazioaren iraultzan eta munduko ia edozein gobernu horren parte izan da (Taylor *et al.* (2007)). eGobernua informazio-teknologia (IT) eta informazio-sistemen berrikuntzaren arlo garrantzitsua da. Horregatik, gobernuak eta instituzioak orokorrean, gobernu prozesuetan eGobernua txertatzen ari dira (Alshehri eta Drew (2010)). eGobernuaren garapena zeregin inportantea da. Ez soilik instituzioek informazioa eta zerbitzuak emateko modua nahiz publikoarekin tratatzeko era azkar aldatzea eragiten duelako, baizik eta haien estrategien parte integrala bilakatzen ari delako (Zhang *et al.* (2014)).

2011ko irailean, Espainia arlo honetan herrialderik aurreratuenen artean zegoen eta bosgarren postuan Europa mailan online motako zerbitzu publikoen erabilgarritasun eta sofistikazio mailan (Spanish Institute of Public Administration (2011)). Espainian, eGobernuaren aurrerapenerako bi faktore nagusi egon dira: batetik, bertako zerbitzu erabiltzaile potentzialek erakutsitako ezagutza eta konpromisorako aurretiko jarrera ona eta bestetik, azken urteetan arlo publikoak egindako lege-arloko planifikazioa eta ahalegin handiak (Gonzalez *et al.* (2007)). 2007ko ekainaren 1ean onartutako Hiritarren Zerbitzu Publikoen Sarbide Elektronikoaren Legeak eGobernuaren inplementaziorako konpromisoa babesten zuen. Lege honen helburu nagusia da "Hiritarrek zerbitzu publiko guztietara sarbidea izatea eta izapide administratibo guztiak internet bidez, edozein tokitatik eta edozein ordutan egiteko ahalmena izatea". Autonomia-erkidegoetako agintarien helburu nagusia lege hori aplikatu ahal izateko, eGobernuko zerbitzuak (eZerbitzuak) hiritar guztientzat eskuragarri egotea izan da (Cegarra-Navarro *et al.* (2012)).

Legezko-ekintzak W3C-k edo Web Content Accessibility Guidelines (WCAG) 2.0 (W3C (2012)) gidak proposatutako irisgarritasun estandarretan oinarritzen dira. Estandar hauek nabigazioa pertsona ororentzat irisgarria izatea dute helburu.

Bestalde, datu-meatzaritza eta modelatze-teknikak erabiltzaileen nabigazio-ereduak lortzeko eta eGobernuaren garapenerako lagungarriak izan daitezke. Izan ere, alde batetik, jokabide interaktiboa iragartzea ahalbidetzen dute eta bestetik, esteka baten edukia, webgune baten informazio-arkitektura eta web orrialde baten diseinuaren egokitasuna ebaluatzeko ere baliagarriak dira (Vigo eta Harper (2013)). Web erabilpenaren meatzaritza (Fujimoto *et al.* (2011)) behatutako ekintzetatik ezagutza erauzteko erabil daiteke, hala nola, erabiltzaileen profilak lortzeko (Schiaffino eta Amandi (2009)). Resource Description Framework (RDF) (Lassila eta Swick (2002)) edo RDF Schema (RDFS) (Brickley eta Guha (2002)) bezalako Web Semantikoko teknologiek aldiz, aukera eskaintzen digute baliabideen domeinua, baliabideak, erabiltzaileak edo estekak modelatu eta adierazteko. Ontologiak gaur egun oso erabiliak dira erabiltzaile ereduak sortzeko, kontzeptualizazio semantiko komuna eskaintzen dutelako (Dolog eta Nejdl (2007)).

Gure ustez, erabiltzaileen profilak sortzeko aipatu metodoak erabiltzeak eGobernuari bi ekarpen nagusi egiten dizkio (Gugliotta *et al.* (2005)): lehenengoa, eGobernuen zerbitzu desberdinak erabilerrazago bilakatzea erabiltzaile mota anitzentzat eta bigarrena, zerbitzu hauen integrazioa hobetzea.

Artikulu hau gure ikerketa taldea eGobernuen inguruan egiten ari den lanaren parte da. Zentzu honetan, lan hau Euskal Herriko Unibertsitatearekin (UPV/EHU) egindako elkarlanaren emaitza da. Gure helburua, haiek eAdministrazio gisa eskaintzen duten zerbitzua analizatzea izan da, zehazki, matrikulazioari dagokion atala (www.ehu.eus/web/sarrera-acceso).

Zerbitzuak egokitzeko eta irisgarriak egiteko lehenengo urratsa, haien egitura eta uneko erabilera aztertzea da. Beraz, lehenik eta behin, unibertsitatearen webgunearen egitura eta edukia aztertu ditugu, matrikulazioari dagokion zonaldea detektatzeko eta URLen sailkapena egiteko. Bigarrenik, zerbitzarietan metatutako log fitxategiak erabiliz, zonalde horren erabilera aztertu dugu. Erabiltzaile-saioak bi modutan adierazi ditugu: erabiltzaileek saioetan atzitutako URLen sekuentzia moduan eta saioen informazioan oinarritutako atributu mekanikoen bidez. Bi datu-base (DB) hauetan, ikasketa automatikoko teknikak erabili ditugu eta aurretik definitutako bi erabiltzaile mota detektatzeko gai izan gara: saioaren amaieran matrikulazioari buruzko informazio asko duten URLetan interesa adierazi dutenak (arrakasta) eta informazio gutxiko URLetan amaitu duten erabiltzaileak (porrota). Honek bide emango die webgunea kudeatzen dutenei etorkizunean zerbitzua erabiltzaile mota desberdinei egokitzeko eta beraz, eZerbitzu erakargarriago bat eskaintzeko.

2. Ikerketaren helburuak

Ikerketa honek UPV/EHU-k eAdministrazio gisa eskaintzen duen matrikulazio eZerbitzuan erabiltzaile-profilak detektatzea eta haien ezaugarriak ezagutzea du helburu. Horretarako, bertako nabigazioaren log fitxategien informazioa erabili dugu eta web-meatzaritzako tekniken bidez nabigazio-profilak aztertu ditugu. Gure etorkizuneko asmoa da profil hauetan oinarrituz eZerbitzu honen erabiltzaile berriak automatikoki sailkatzeko gai den eta haiei egokitzen zaien sistema sortzea.

Gure esperimentuek aditzera eman dute aztertutako eZerbitzuan erabiltzaile-profil desberdinak era automatikoan detektatu daitezkeela. Zehazki, erabilitako ikasketa automatikoko algoritmo ezagunen emaitzak aintzat hartuz, baieztatu dezakegu eZerbitzu honen erabiltzaile berrien nabigazio-patroiak aurreikusteko bidean gaudela. Horregatik, uste dugu web-meatzaritzako teknikak lagungarriak izango direla zerbitzu honetan erabilgarritasuna, eraginkortasuna edota erabiltzailearen asebetetzea hobetzeko.

eZerbitzuetan, nabigazioaren amaierak erabiltzailearen sailkapena ahalbidetu ohi du. Esaterako, tramite bat burutzeko arazoak izan dituzten erabiltzaileak ez dira eZerbitzuaren amaieraraino iristen. Horregatik, matrikulazio zonaldea eZerbitzu baten moduan modelatu ahal izateko gure sailkapena ere erabiltzaileak saioa amaitu duen moduan oinarritu da. Horren arabera bi saio mota bereizi ditugu, arrakasta / porrota. Hurrengo lana erabiltzaile-saioak

automatikoki sailkatzeko saioak bi modutan adieraztea izan da: lehena, nabigazio-zonaldean (URL sekuentzia) oinarritzen da eta bigarrena, nabigazio-moduan (atributu mekanikoak). Bi adierazpide hauen bidez, portaera desberdinen arabera sailkatu ahal izango ditugu erabiltzaile-saioak.

3. Ikerketaren muina

3.1 Esperimentuaren garapena

Ikerketa honetan, UPV/EHUko webgunean matrikulazioaren inguruko informazioa ematen duen zonaldeko nabigazioa aztertu dugu eZerbitzu ikuspuntutik. Horretarako lehenengo urratsa, webgune osoaren nabigazioari lotutako 49 eguneko log fitxategiaren aurreprozesamendua izan da (2016/02/23-tik 2016/04/12-ra). Osotara, hiru URLtik gorako luzera duten 25.467 saio lortu ditugu matrikulazio zonaldean (www.ehu.eus/web/sarrera-acceso). Luzera minimoa hiru klik izatea erabaki dugu, nabigazioa esanguratsua izan dadin.

Bigarrenik, matrikulazio zonaldea eZerbitzu modura modelatzeko erabiltzaile-saioen sailkapen bat definitu dugu. Horretarako, saioetan ikusitako URLen mota aztertu dugu, bi irizpide aintzat hartuz: URLaren edukia (testua ala estekak nagusi diren) eta URLaren zonaldea (matrikulaziokoa edo beste zonaldekkoa den). URLen edukiari dagokionez, testu-fitxategi formatua duten URLak (.pdf / .doc / .docx) edukizko gisa sailkatu ditugu. Beste kasuetan, URLa edukizkoa ala bideratzailea (estekak nagusi dituen) den jakiteko honako heuristikoa (1) definitu dugu:

$$\frac{\text{Hitz erreal kop.}}{\text{Esteka kop.}} = \frac{\text{Hitz kop.} - \text{Esteketako hitz kop.}}{\text{Esteka kop.}} \leq 10 \rightarrow \text{URL bideratzailea} \quad (1)$$

$$> 10 \rightarrow \text{edukizko URL} - a$$

eZerbitzu erreal batean erabiltzaileek helburu edo zeregin zehatzak izan ohi dituzte (tramite bat egin ...) eta erabiltzaile-saioen sailkapena agerikoa da (zeregin hori burutu duen edo ez ...). Gure kasuan ordea, erabiltzaileek ez dute zeregin zehatzik eta beraz, saioen sailkapenerako bestelako irizpideak erabili behar izan ditugu. Gure sailkapena erabiltzaileen nabigazioaren bukaerako orrian oinarritzen da; erabiltzaileek atzitutako azken URLaren motan hain zuzen. Hortaz, erabiltzaileak atzitutako azken URLa bere nabigazioaren lorpenarekin lotu dugu.

Matrikulatzeko helburua duten erabiltzaileengan pentsatuz, matrikulazioari buruzko informazio asko duten URLetan (edukizkoak) amaitzen diren erabiltzaile-saioak arrakastarekin lotu ditugu. Bestalde, informazio gutxiko web orrialdeetan (bideratzaileak) amaitu diren erabiltzaile-saioak porrotarekin lotu ditugu. Hots, arrakasta motako saioak azken URLa edukizkoa (1) eta matrikulazio zonaldekkoa dutenak dira. Porrota motako saioetan aldiz, azken URLa bideratzailea (1) eta edozein zonaldekkoa da. Lehen hurbilpen honetan, edukizkoa eta matrikulazio zonaldekkoa ez den URL batean amaitutako saioak analisitik kanpo utzi ditugu. Izan ere, mota horretako saioak arrakastarekin lotura izan dezakete, baina matrikulaziokoa ez den testuinguruan. Esaterako, beste zonalde bateko informazioa eskuratzea helburu zutenak. 1. Taulan definitutako saioen sailkapena eta saio-mota bakoitzeko kasu kopuruak azaltzen dira. Osotara arrakasta motako 10.734 saio lortu ditugu (% 42,1) eta porrota motako 14.733 saio (% 57,9).

1. taula. EHUko matrikulazio zonaldeko erabiltzaile-saioen sailkapena

Saio mota	Saio kopurua	Saioaren azken URLa	
		Zonaldea	Mota
Arrakasta	10.734 (% 42,1)	Matrikulazioa	Edukizkoa
Porrota	14.733 (% 57,9)	Matrikulazioa	Bideratzailea
		Ez Matrikulazioa	Bideratzailea

Hautatutako saioekin bi datu-base (DB) sortu ditugu: lehenengoan, erabiltzaile-saioak adierazteko erabiltzaileek atzitutako URLen segidak (sekuentziak) erabiltzen dira zuzenean. Bigarrenean,

erabiltzaile-saioak adierazteko, log fitxategietako informazioan oinarrituz, eskuz kalkulaturako atributu mekanikoez osaturako bektorea erabiltzen da. Bigarren DBan, saio bakoitza adierazteko erabilitako atributuak kalkulatzeko denboran, URLen sailkapenean (edukia / zonaldea) edota klik kopuruetan oinarritzen gara. 2.taulan ageri da atributu hauen zerrenda eta deskribapena. Beraz, lehenengo DBak adierazten du erabiltzaileak nondik nora ibili diren eta bigarrenak aldiz, nola egin duten nabigazio hori.

2. taula. Erabiltzaile-saioak adierazteko atributuak

Atributua	Azalpena
Klik kop.	Klik kopurua (saioaren luzera).
% Bid. kop.	Bideratzaile motako URL kopurua / saioaren luzera
% Edu. kop.	Eduki motako URL kopurua / saioaren luzera
% M. kop.	Matrikulazio zonaldeko URL kopurua / saioaren luzera
% Ez-M. kop.	Matrikulazioa ez den zonaldeko URL kopurua / saioaren luzera
% Ind. kop.	Matrikulazioa zonaldeko hasierako orrialdea motako URL kopurua / saioaren luzera
Ref-B. kop	Erreferentzia bilatzailea duten URL kopurua
Sek-T	Saioaren iraupena (s)
Klik-T_mean	Klik baten iraupena batazbeste (s)
Bid.-T_mean	Bideratzaile motako URLa duen klik baten iraupena batazbeste (s)
Edu.-T_mean	Eduki motako URLa duen klik baten iraupena batazbeste (s)
M.-T_mean	Matrikulazio zonaldeko URLa duen klik baten iraupena batazbeste (s)
Ez-M.-T_mean	Matrikulazioa ez den zonaldeko URLa duen klik baten iraupena batazbeste (s)
Ind-T.-mean	Matrikulazio zonaldeko hasierako orrialdea motako URLa duen klik baten iraupena batazbeste (s)
Edu.-Bid. kop.	Eduki - bideratzaile motako URLen arteko trantsizio kopurua
Bid.-Edu. kop	Bideratzaile - eduki motako URLen arteko trantsizio kopurua
M.-Ez-M. kop.	Matrikulazio – matrikulazioa ez den zonaldeko URLen arteko trantsizio kopurua
Ez-M.-M. kop.	Matrikulazioa ez den – matrikulazioa den zonaldeko URLen arteko trantsizio kopuru

Gure ikerketaren helburua ez da soilik saio-mota bakoitza automatikoki detektatzea, hots, bi klaseko (K) sailkatzailea sortzea arrakasta eta porrota desberdintzen dituen. Gure erronka nagusia, eZerbitzu hau hobetzea da eta prozesu horretan erabilgarria izango zaigun ahal bezainbeste informazio eskuratu nahi izan dugu. Esaterako, saio-mota bakoitzaren barruan dauden portaera desberdinak. Helburu horrekin, datu-base bakoitzari ikasketa automatikoko algoritmoak aplikatu dizkiogu. Zehazki, URL sekuentziez osaturako DBako PAM (k-medoids) (Kaufman eta Rousseeuw (1990)) algoritmoa erabili dugu. Izan ere, PAM sekuentziak multzotan biltzea ahalbidetzen duen clustering algoritmoa da. Bestetik, saioen atributu mekanikoez osaturako DBako, atributuak normalizatu ostean (banaketa normala), K-means (Lloyd (1982)) algoritmo ezaguna erabili dugu. Oro har, clustering-ean desiragarria da kluster kopuru txikia aukeratzea, mota bereko ahal bezainbeste kasu biltzen dituzten klusterak lortzeko. Gure kasuan, frogatutako K desberdinek (50, 75, 100 ...) ez dutenez klusterren egituraren eragin gehiegirik izan, frogatutako baliorik txikiena aukeratu dugu bi algoritmoetan (K=50).

Erabiltzaileen portaera desberdinak aztertzeko, arrakasta edo porrotaren %74ko nagusitasuna duten klusterretan zentratu gara. Bi dira arrazoi nagusiak: batetik, kluster horietan dagoen saio kopuru totala

adierazgarria delako (DB osoaren % 42a da) eta bestetik, saio-mota bakoitzaren ordezkartza egokia lortzen delako (% 12 - %14 arrakasta eta % 29 porrota).

Gehigarri moduan, nabigazio-zonaldea eta nabigazio-modua ikuspuntuaren arteko erlazioa aztertzeke, bi bideetatik lortutako partizioak konparatu ditugu Jaccard indizea (Jaccard (1908)) erabiliz. Azkenik, gainbegiratutako ikasketako J-48 sailkapen-zuhaitza erabili dugu atributu mekanikoen gure sailkapena iragartzeko duten ahalmena ikertzeke. Erabilitako J-48 Weka (Hall et al. (2009)) programa barruko C4.5 (Quinlan (2014)) algoritmo ezagunaren inplementazioa da. Hurrengo azpi atalean, deskribatutako dira lortutako emaitzak.

3.2 URL segidarekin sortutako DBko clustering-aren emaitzak

Lehen aipatu moduan, URLez osatutako sekuentzien datu-basean PAM (k-medoids) (Kaufman eta Rousseeuw (1990)) algoritmoa aplikatuz 50 kluster sortu dira. Kluster hauen azterketa egin da haietan metatutako sekuentzia motaren arabera, arrakasta / porrota. Emaitzak aztertuz, hauek lirateke datu aipagarrienak:

- Arrakasta nagusitzen den zortzi kluster lortu dira, arrakasta-saioen proportzioa % 74 baino handiagoa dutenak.
- DB osoko saioren % 12 erabiliz (arrakastari lotutako zortzi klusterrak) arrakasta motako saioren %24a biltzeko gai gara.
- Porrota nagusi duten 17 kluster lortu dira, porrota-saioen proportzioa % 74 baino handiagoa dutenak.
- Saio guztien % 29a hartuz (porrotari lotutako 17 klusterrak) porrota motako saioren % 45a detektatzeko gai gara.

3. taulan, arrakastari lotutako zortzi kluster eta porrota nagusi den 17 klusterren ezaugarriak adierazi ditugu.

3. taula. PAM (K=50) URLen DBaren gainean aplikatu osteko emaitzak

Atributua	Arrakasta-saio kop. > % 74 duten klusterrak	Porrota-saio kop. > % 74 duten klusterrak
Kluster kop.	8	17
Arrakasta-saio kop.	2.551 (% 81,9)	842 (% 11,2)
Porrota-saio kop.	564 (% 18,1)	6.669 (% 88,8)
Saio kop. klusterretan	3.115	7.511
Saio kop. - DB (%)	% 12,2	% 29,5

3. taula aztertuz, esan genezake kluster kopuruaren erdian (25) saio-mota bat edo bestearen nagusitasuna % 74tik gorakoa dela. DB osoan, porrota motako 14.733 erabiltzaile-saio ditugu eta horrela izendatutako klusterretan, porrota motako 6.669 saio daude. Hortaz, DB osoan porrota motako kasuak % 58 izan arren, kluster horietan erortzen direnak probabilitate handiagoa dute porrota izateko, % 89. Era berean, DB osoan arrakasta motako erabiltzaile-saio kopurua % 42 izanik, hautatutako klusterretan % 82ra igotzen da kopuru hau. Honengatik guztiagatik ondoriozta dezakegu nabigazioan zeharkatzen diren zonek (URLak), definitutako arrakasta / porrota sailkapenarekin zerikusia dutela eta clustering-a egitura hori automatikoki lortzeko gai dela. Hortaz, esan genezake UPV/EHUko matrikulazio eZerbitzuko erabiltzaile berrien nabigazioa era automatikoan sailkatzeko aukerak daudela nabigazio berrien nondik norakoei (URLak) erreparatuz gero.

3.3 Nabigazioaren atributu mekanikoen sortutako DBko clustering-aren emaitzak

Era berean, K-means (Lloyd (1982)) algoritmoa atributu mekanikoen (2. taula) osatutako saioren datu-basean aplikatu ostean 50 kluster sortu dira. Hemen, arrakasta motako saioren kopurua % 74tik gora duten sei kluster lortu dira, kluster hauetan dagoen saio kopurua guztira DB osoaren % 14 izanik.

Testuinguru honetan, saio guztien % 14a hartuz, arrakasta motako saioen % 30 biltzeko gai gara Bestalde, porrota-saio kopurua % 74tik gora duten 21 kluster lortu dira, kluster hauetan dagoen erabiltzaile-saio kopurua DB honen % 29 delarik. Horrekin batera, saio guztien % 29a hartuz porrota motako saioen % 46 detektatzeko gai gara. 4. taulan deskribatzen dira emaitza hauek.

4. taula. K-NN (K=50) atributuen DBaren gainean aplikatu osteko emaitzak

Atributua	Arrakasta-saio kop. > % 74 duten klusterrak	Porrota-saio kop. > % 74 duten klusterrak
Kluster kop.	6	21
Arrakasta-saio kop.	3.212 (% 87,7)	544 (% 7,4)
Porrota-saio kop.	451 (% 12,3)	6.842 (% 92,6)
Saio kop.	3.663	7.386
Saio kop. - DB (%)	% 14,4	% 29

Aurreko DBan gertatu bezala, 4. taula aztertuz baieztatu daiteke klusterren erdietan baino gehiagotan guk definitutako bi saio moten nagusitasuna % 74tik gorakoa dela. Kasu honetan, porrota gisa hautatutako klusterretan, porrota motako 6.842 saio daude. Aurreko kasuan bezalaxe, nahiz eta DBaren % 58 porrota izan, aipatu klusterretan eroritako saioek probabilitate handiagoa dute porrota izateko, % 93 hain zuzen ere. Halaber, DB osoan arrakasta motako saio kopurua % 42 izanik, aukeratutako klusterretan nabarmenki igotzen da kopuru hori, % 88ra hain zuzen. Beraz, uste dugu UPV/EHUko webgunearen matrikulazio zonaldeko nabigazioa atributu mekanikoez osatutako saioen bidez adierazi daitezkeela. Hau horrela izanik, eZerbitzu honen erabiltzaile-saio berriak arrakasta / porrota moduan automatikoki sailkatzeko erabiltzaile-saio horien atributu mekanikoak erabil litezke.

3.4 Sortutako bi partizioen arteko konparaketa

Nabigazio-sekuentziekin sortutako DBan erabilitako PAM algoritmoaren emaitzak eta nabigazioaren atributu mekanikoen sortutako datu-basean erabilitako K-means algoritmoaren emaitzak alderatzeko, multzoak konparatzeko oso ezaguna den Jaccard indizea (Jaccard (1908)) erabili dugu. Bi partizioen konparaketan, indizeak 0.04 balorea eman du eta balore txikia izanik, bi emaitzak oso desberdinak direla adierazi du. Honen aurrean, esan genezake UPV/EHUko matrikulazio eZerbitzuko nabigazioan, zonaldea (URLa) eta atributu mekanikoen deskribatzen duten nabigatzeko modua independenteak direla. Beraz, eZerbitzu honen erabiltzaile berrien nabigazioa arrakasta / porrota bezala sailkatzeko aipatu bi ezaugarriak (zonaldea / mekanikoak) erabilgarriak lirateke eta aztertzeke legoke bi sistemak osagarriak ote diren.

3.5 Atributu mekanikoen oinarritutako gainbegiraturako ikasketaren emaitzak

Atributu mekanikoen DBan erabilitako gainbegiraturako ikasketako J-48 algoritmoak sortutako sailkapen-zuhaitzak asmatze-tasa handia du: % 95,6koa. Zuhaitzaren arabera, definitutako saio-motak diskriminatzeke garaian honako atributuak lirateke garrantzitsuenak:

- Bideratzaile motako URLa duen klik baten iraupena batazbeste (Bid.-T_mean).
- Eduki motako URLa duen klik baten iraupena batazbeste (Edu.-T_mean).
- Bideratzaile motako URL kopurua saioaren luzerarekiko (% Bid. kop.) .
- Matrikulaziokoa den - matrikulaziokoa ez den zonaldeko URLen arteko trantsizio kopurua (M.-Ez-M. kop.).

Beraz, esan dezakegu atributu mekanikoak baliagarriak direla saioak guk definitutako moduan era automatikoki sailkatzeko, lehen adierazitakoak garrantzi gehiago dutelarik.

4. Ondorioak

Lan honetan UPV/EHUko matrikulazio-eZerbitzua web-meatzaritzako tekniken bidez modelatu da. Lehenik eta behin, eZerbitzu honen nabigazioen amaierari erreparatuz (azken klikaren zonaldea eta eduki mota) bi saio-mota definitu ditugu: arrakasta eta porrota. Arrakasta nabigazio-saioak zehazki, UPV/EHUko matrikulazio-zonaldekoa den eta testua nagusi duen URL batean amaitutakoak lirateke. Bestalde, porrota motako saioak matrikulazioa nahiz beste edozein zonaldekoa den eta bideratzailea den UPV/EHUko URL batean bukatutakoak dira (1. taula). Definizio hauekin eta nabigazio-zonaldea edo nabigatzeko modua aintzat hartuz bi DB sortu ditugu eta hauen gainean PAM eta K-means clustering algoritmoak erabili ditugu. Horren helburua, definitutako nabigazio-motak (arrakasta / porrota) era automatikoki biltzen diren jakitea eta eZerbitzu hau etorkizunean hobetzeko baliagarria zaigun informazioa eskuratzea da.

Bi prozesuetan antzeman dugu sortutako klusterren erdietan gutxienez, definitutako nabigazio-mota bat edo bestearen nagusitasuna % 74tik gorakoa dela. Horrez gain, PAM eta K-means algoritmoak gai dira DB osoko saioen % 29a erabiliz, porrota-saioen % 45 - % 46a detektatzeko. Era berean, bi algoritmoek saioen % 12 - % 14a aztertuz, arrakasta-saioen % 24 - % 30 detektatzeko ahalmena dute. Horregatik, esan dezakegu erabilitako bi adierazpideen partizioek eZerbitzu honetan definitutako bi nabigazio-motak automatikoki detektatzeko bide ematen dutela.

Gehigarri bezala, lortutako bi partizioak konparatu ditugu Jaccard indize ezaguna (Jaccard (1908)) erabiliz. Indize honek 0.04 balioa eman du aipatu bi partizioen konparaketan eta beraz, ondorioztatu dezakegu nabigazio-zonaldea eta modua ez daudela erlazionatuta. Horrek esan nahi du, gerora eZerbitzu honen etorkizuneko erabiltzaileen nabigazioa arrakasta edo porrota gisa sailkatzeko aztertutako bi bideak lagungarriak izango direla. Etorkizunean, nabigazio-zonaldean oinarritutako ikuspuntua eta nabigazio-moduan oinarritutakoa osagarriak ote diren aztertu beharko dugu.

Bukatzeko, atributu mekanikoekin gainbegiraturako ikasketako J-48 algoritmoa erabili dugu, % 95,6ko asmatze-tasa duen sailkapen zuhaitza lortuz. Zuhaitz honen egitura aztertuz, nabigazio-portaerak diskriminatzeko garaian garrantzitsuen diren atributuak ezagutu ahal izan ditugu. Beraz, bide honetatik ere nabigazio-ereduak lortzen ari gara.

Lan honi esker UPV/EHUko matrikulazio-eZerbitzua modelatzeko lehen pausua eman dugu eta emaitzak ikusita baieztatu dezakegu web-meatzaritzako teknikak prozesu horretan lagungarriak direla.

Etorkizuneko lerro moduan, lehenik eta behin, DBa adierazteko bi hurbilpenak osagarriak ote diren aztertu beharko dugu. Hala balitz, sistemen doitasuna hobetu nahiko genuke. Sortutako ereduak sakonkiago aztertu nahi ditugu, etorkizuneko erabiltzaileen nahiei aurre hartzeko, eta erabileran eragin negatiboa duten elementuak identifikatu eta hobetzeko.

Bukatzeko, eZerbitzu honen azterketaren ondorioak erabili nahi genituzke bestelako eZerbitzuak hobetzeko gidalerro moduan eta oro har, eGobernua hobetzeko baliagarriak diren ezagutzak plazaratzeko.

6. Erreferentziak

- Alshehri, M. eta Drew, S. (2010), Challenges of eGovernment Services Adoption in Saudi Arabia from an e-Ready Citizen Perspective, *World Academy of Science, Engineering & Technology*, 4 (6), 1086-1092.
- Brickley, D. eta Guha, R. V. (2002), Resource Description Framework (RDF) Schema Specification 1.0, <http://www.w3.org/TR/rdf-schema>.
- Cegarra-Navarro, J., Córdoba-Pachón, J. R. eta Moreno-Cegarra, J. L. (2012), eGovernment and citizen's engagement with local affairs through e-websites: The case of Spanish municipalities, *Int. Jour. of Inf. Management*, 32, 469-478.

- Dolog, P. eta Nejdl, W. (2007): Semantic web technologies for the adaptive web, in *The adaptive web*, P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.), *Lecture Notes in Computer Science*, Springer-Verlag, Berlin, 4321, 697-719.
- Fujimoto, H., Etoh, M., Kinno, A. eta Akinaga, Y. (2011): Web user profiling on proxy logs and its evaluation in personalization, in Du, X., Fan, W., Wang, J., Peng, Z. eta Sharaf M. A. (Eds.), *Procs. of the 13th Asia-Pacific web conference on Web technologies and applications (APWeb'11)*, Springer-Verlag, Berlin, 107-118.
- Gonzalez, R., Gasco, J. eta Llopis, J. (2007), eGovernment success: Some principles from a Spanish case study, *Industrial Manag. & Data Systems*, 107(6), 845-861.
- Gugliotta, A., Cabral, L., Domingue, J., Roberto, V., Rowlatt, M., Council, E. C. eta Davies, R.A. (2005), Semantic web service-based architecture for the interoperability of e-Government services, *WISM 2005*, 21.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. eta Witten, I. H. (2009). The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter*, 11(1), 10-18.
- Jaccard, P. (1908), Nouvelles recherches sur la distribution florale, *Bulletin de la Société Vaudoise de Sciences Naturelles*, 44, 223-370.
- Kaufman, L. eta Rousseeuw, P.J. (1990), Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience.
- Lassila, O. eta Swick, R. (2002), W3C resource description framework (rdf) model and syntax specification, <http://www.w3.org/TR/REC-rdfsyntax/>.
- Lloyd, S. (1982), Least squares quantization in PCM, *IEEE Transactions on Information Theory*, 28 (2), 129-137.
- Quinlan, J.R. (2014): *C4. 5: programs for machine learning*. Elsevier.
- Schiaffino, S. eta Amandi, A. (2009): Intelligent user profiling, in M. Bramer (Ed), LNAI 5640 chapter Intelligent user profiling, *Artificial Intelligence: An International Perspective*, Springer-Verlag, 193-216.
- SIPA (Spanish Institute of Public Administration) (2011): eGovernment in Spain, *Report to the European Commission*, www.epractice.eu/files/eGovernmentSpain.pdf.
- Taylor, J., Miriam, L. eta Organ, J. (2007), Information-intensive government and the layering and sorting of citizenship, *Public Money and Management*, 27(2), 161-164.
- Vigo, M. eta Harper, S. (2013): Challenging information foraging theory: screen reader users are not always driven by information scent, In *Procs. of the 24th ACM Conference on Hypertext and Social Media*, ACM, NY, USA, 60-68.
- W3C. (2012): WAI, Web Content Accessibility Guidelines (WCAG), <http://www.w3.org/WAI/intro/wcag.php>.
- Zhang, H., Xiaolin, X. eta Jianying, X. (2014), Diffusion of eGovernment: A literature review and directions for future directions, *Government Information Quarterly*, 31(4), 631-636.

7. Eskerrak eta oharrak

Lan honek honako unitateen finantzaketa jaso du: Batetik, Euskal Herriko Unibertsitateko UPV/EHU finantzaketa (PIF15/143 beka). Bestetik, Espainiako gobernuko ekonomia eta lehiakortasun ministerioak eta Europako eskualdeen garapenerako funtsak (ERFD) finantzatutako eGovernAbility proiektuarena (erreferentzia TIN2014-52665-C2-1-R). Azkenik, eusko jaurlaritzako hezkuntza saileko unibertsitateak eta ikerketa atalak sustengatzen duen ADIAN ikerketa taldearena (erreferentzia IT980-16).