Relevanţa conţinutului web şi a comportamentului utilizatorilor în analiza traficului

Student doctorand: Diana - Florina HALIŢĂ
Coordonator ştiinţific: Prof. Dr. Florian Mircea BOIAN
Universitatea "Babeş-Bolyai" Cluj-Napoca
Proiect de cercetare

10 Iunie 2015

Cuprins

Migrarea transparentă a unui site web între două sisteme de management de conţinut

De ce?

De ce să alegem un CMS?

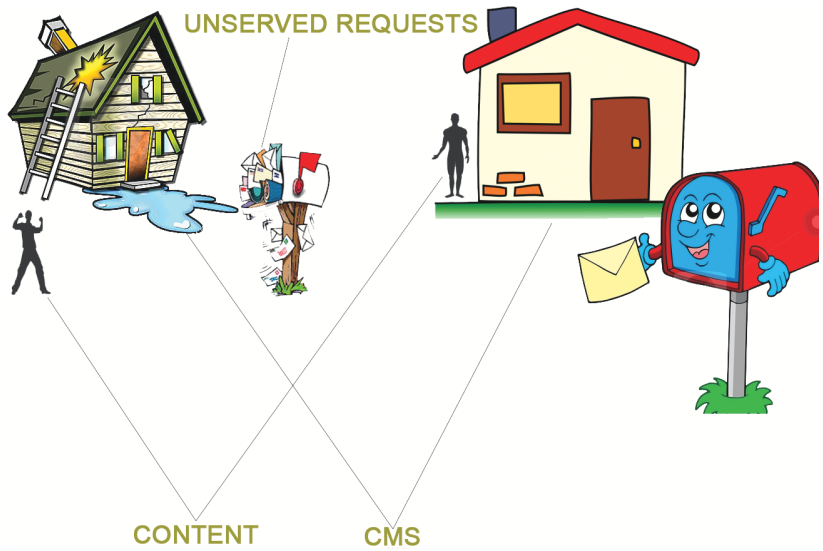
- Întreţinerea facilă - web;
- Independenţa conţinutului de prezentare;
- Update-uri periodice de securitate;
- Roluri multiple pentru utilizatori;

De ce este necesară migrarea?

- CMS-ul vechi este considerat învechit;
- CMS open source;
- Scalabilitatea CMS-ului nou;
- Fructificarea noilor tehnologii web: CSS3, Ajax, HTML5.

Provocări

- expunerea conţinutului site-ului web la un URL nou
- creşterea numărului de vizitatori care sunt induşi în eroare
- pierderea diverselor beneficii câstigate în timp



Soluţii

Soluţii posibile:

- migrarea automată sau manuală;
- estimarea timpului necesar migrării;
- evaluarea procesului de migrare.

Soluţii propuse de alţii:

- plugin-uri care ţin cont de comportamentul utilizatorilor în timp.

Soluţia propusă:

- potrivirea perechilor de URL-uri;
- AVANTAJ:
  algoritmul se utilizează înaintea lansării site-ului.


REZULTATE

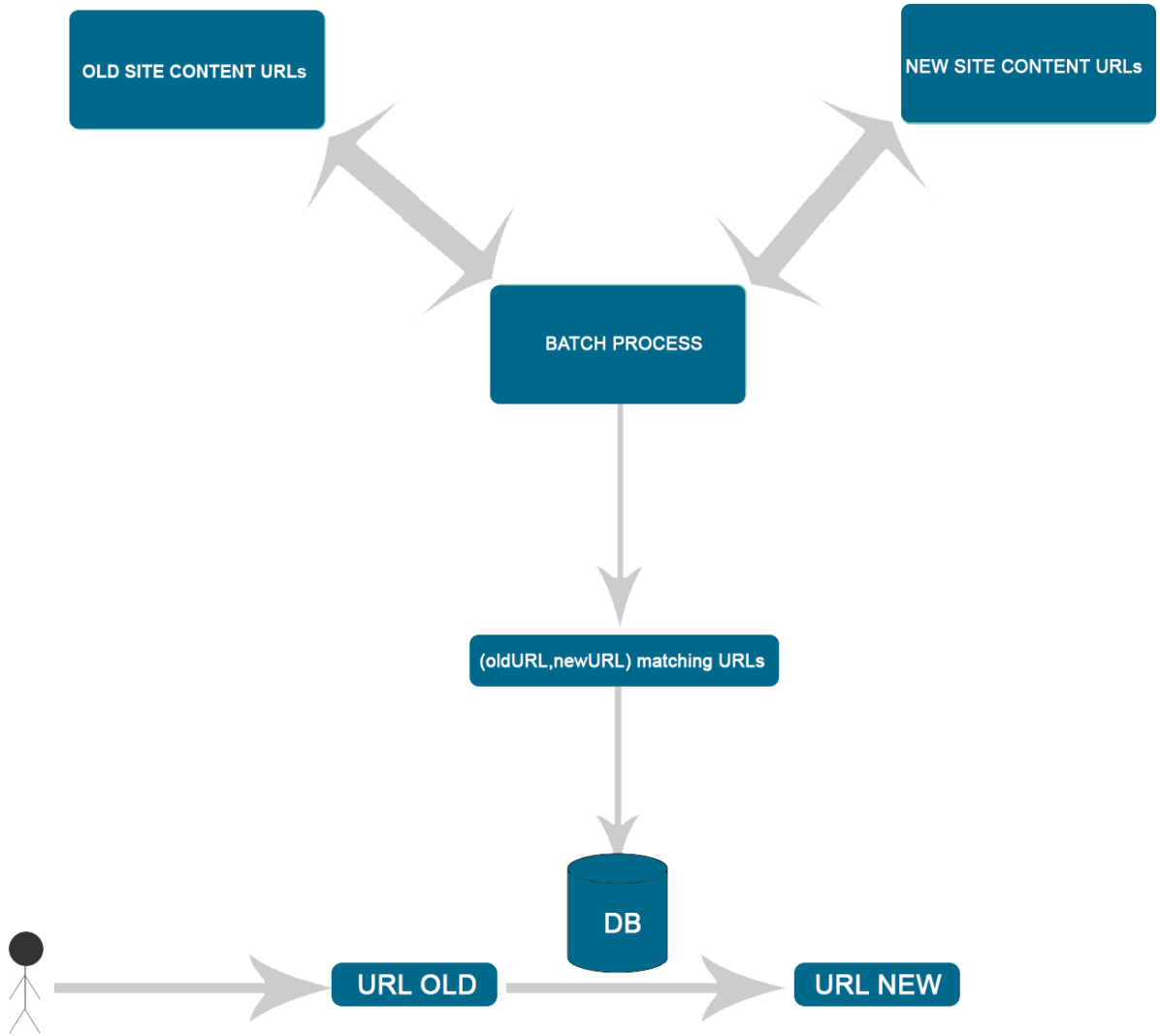


Figure: Batch processing

REZULTATE

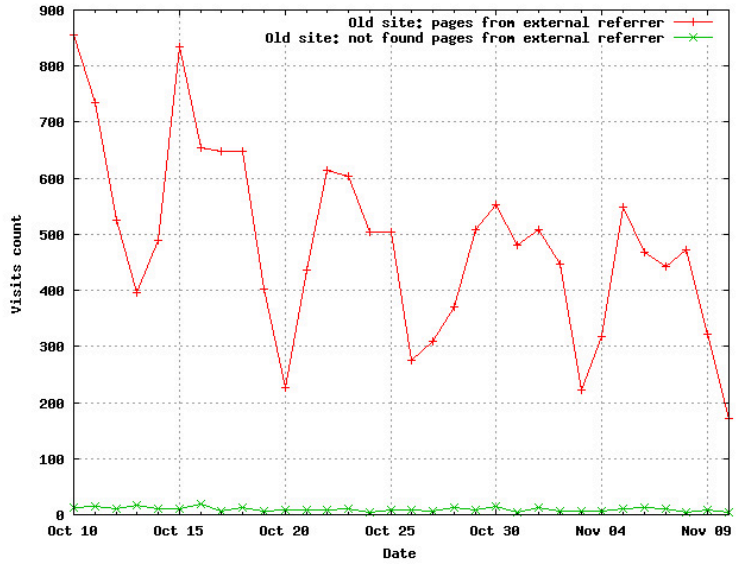


Figure: Vechiul site: numărul de pagini accesate de către un referrer extern şi numărul de pagini care au generat eroarea 404 şi au fost accesate de către un referrer extern

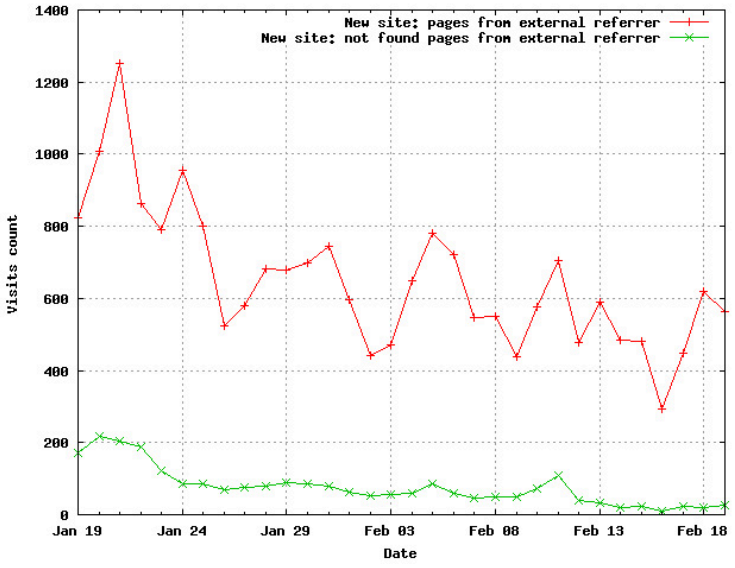


Figure: Noul site: numărul de pagini accesate de către un referrer extern şi numărul de pagini care au generat eroarea 404 şi au fost accesate de către un referrer extern

REZULTATE

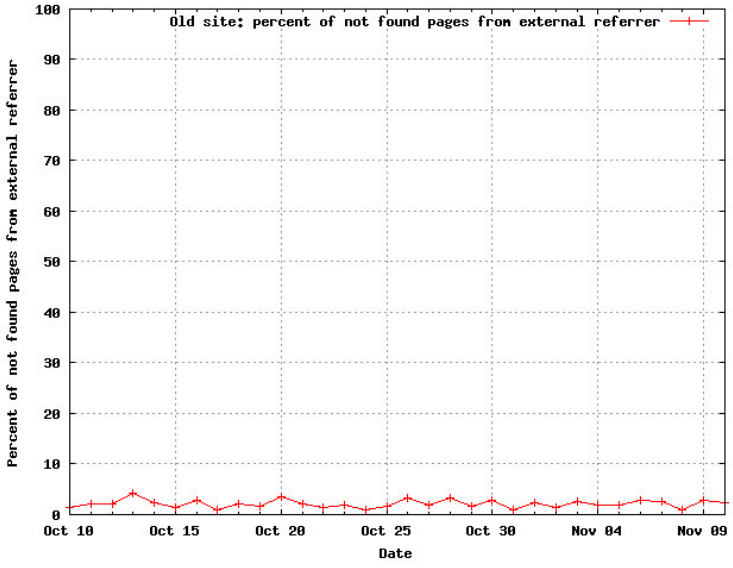


Figure: Vechiul site: procentul paginilor ce au generat eroare 404 şi au fost accesate de către un referrer extern

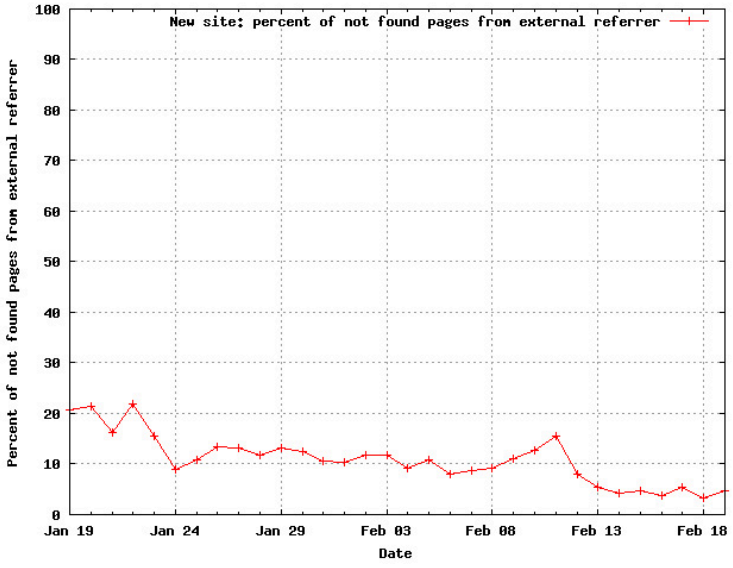


Figure: Noul site: procentul paginilor ce au generat eroare 404 şi au fost accesate de către un referrer extern

REZULTATE

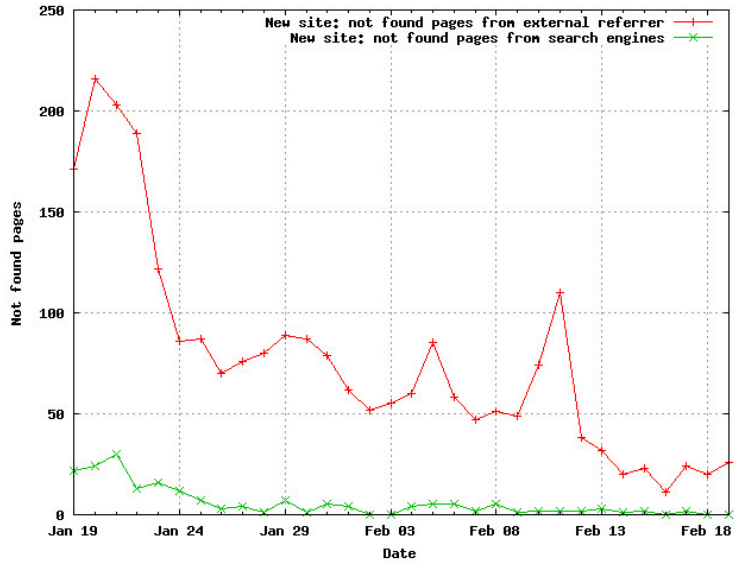


Figure: Noul site: adaptarea motoarelor de căutare la noua structură a site-ului

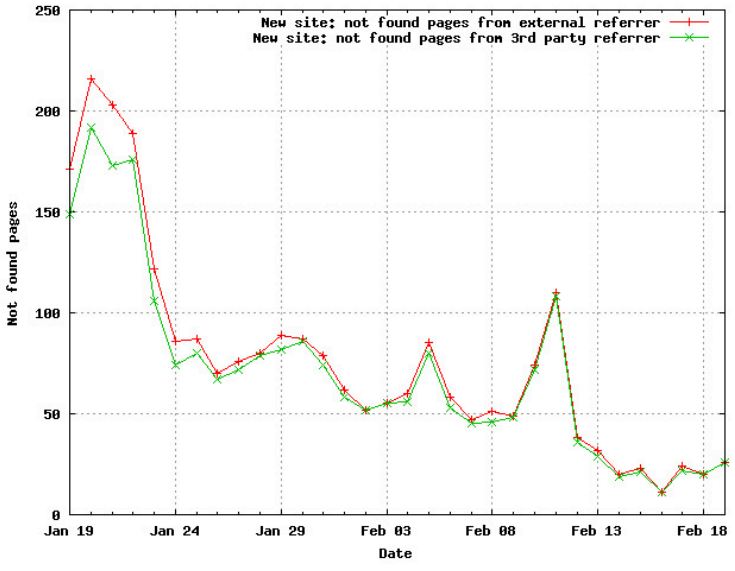


Figure: Noul site: majoritatea erorilor 404 sunt generate de către utilizatorii care vin de la referreri 3rd party

REZULTATE

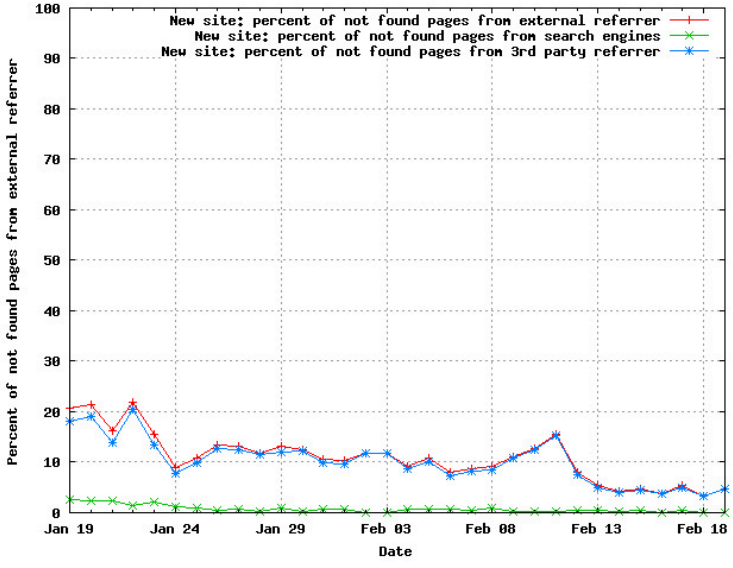


Figure: Noul site: procentul paginilor ce generează eroarea 404 şi procentul paginilor ce generează eroarea 404 şi vin de la motoarele de căutare sau de la referreri 3rd party

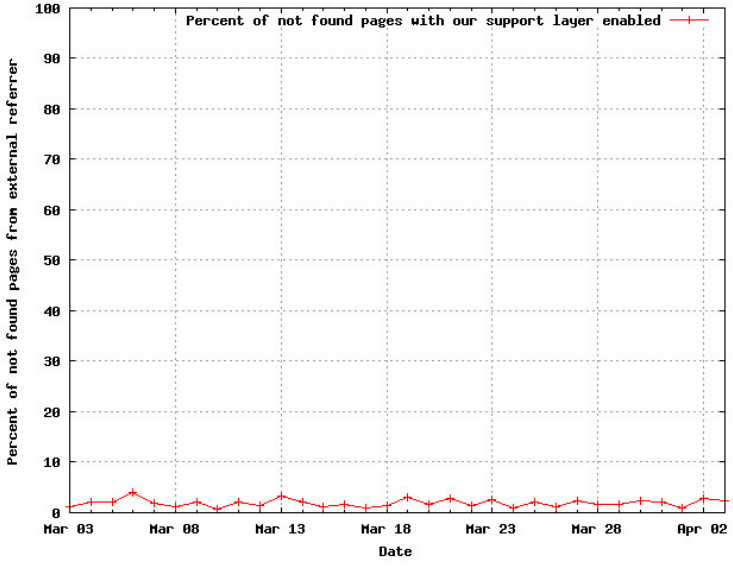


Figure: Procentul paginilor ce generează eroarea 404 având un referrer extern, odată ce a fost activat layer-ul suport propus

Analizarea legăturii dintre similaritatea documentelor web şi a ratei de respingere generate de link-urile dintre aceste documente

De ce?

Link-urile sunt folosite abuziv:

Scop:

- creşterea page rank-ului domeniului destinaţie;

Rezultat:

- utilizatorului nu îi este prezentată o informaţie de care să fie interesat;

Locaţie:

- sunt localizate fie sitewide, fie în cadrul conţinutului absolut.


Soluţii

Tehnici de detectare a link-urilor abuzive:

- analiza conţinutului
- analiza link-urilor
- analiza comportamentului utilizatorilor
- metode de clasificare automată, supervizată sau nesupervizată

Acurateţe:

- 80% din paginile care contin link-uri abuzive sunt detectate

Recomandare:

- combinarea tehnicilor $\Rightarrow$ creşte procentul de pagini abuzive detectate

REZULTATE

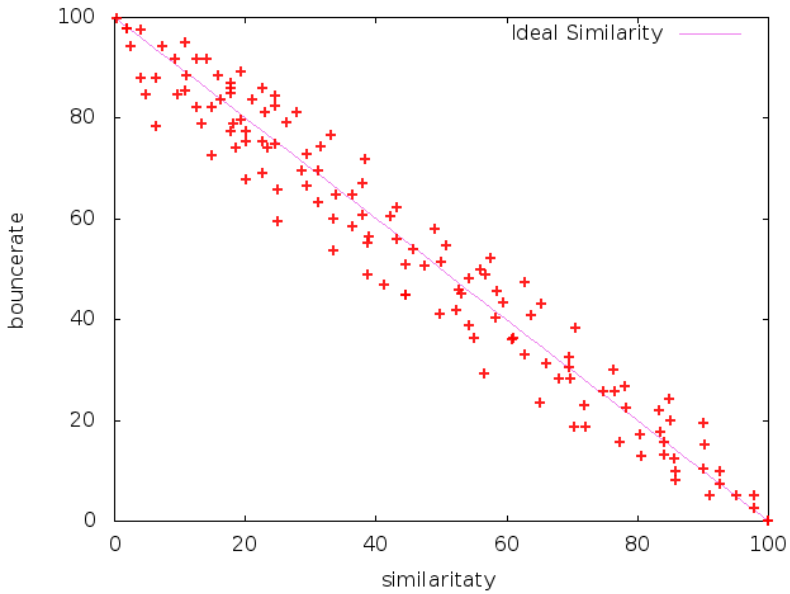


Figure: Similaritate ideală

- similaritate mare $\Rightarrow$ rata de respingere mică
- similaritate mică $\Rightarrow$ rata de respingere mare
- cea mai bună funcţie de similaritate:
  condiţii: suma

$$\sum \frac{|x_i + y_i - 100|}{\sqrt{2}}$$

este minimă

REZULTATE

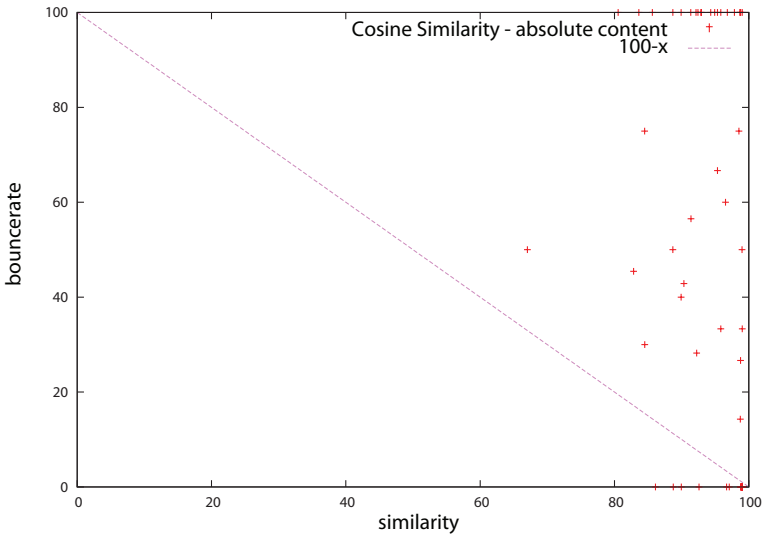


Figure: Similariatea Cosinus

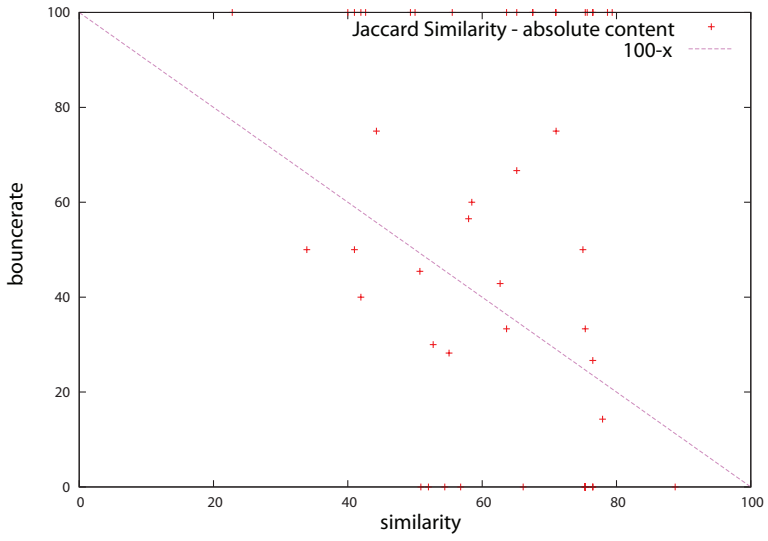


Figure: Similaritatea Jaccard

REZULTATE

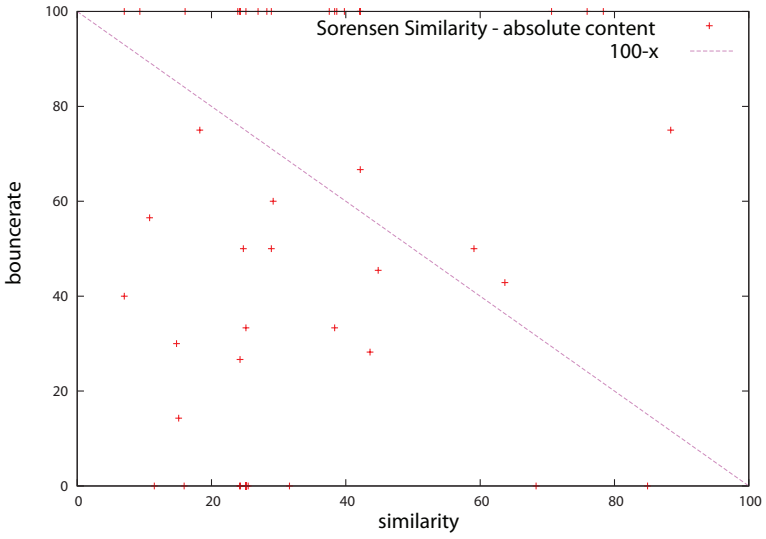


Figure: Similariatea Sorensen

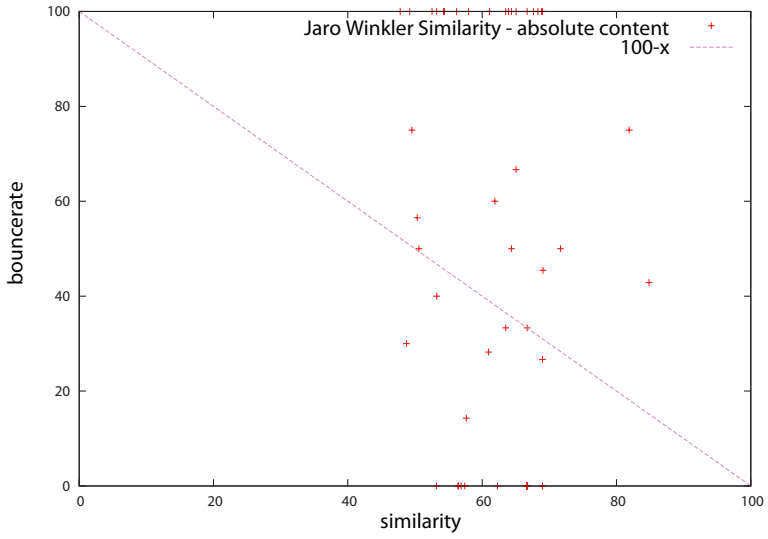


Figure: Similaritatea Jaro Winkler

REZULTATE

Funcţie de similaritate	Metodă	Valoarea sumei
Cosinus	conţinut absolut	1804.476000579978
Jaccard	**conţinut absolut**	**1414.03085464388**
Sorensen	conţinut absolut	1769.3699189543242
Jaro-Winkler	conţinut absolut	1528.5359097516346

Table: Suma distanţelor de la toate punctele de pe grafic la dreapta de ecuaţie $y = x - 100$


Măsurarea şi vizualizarea nivelului de scrapping al unui site web

De ce?

Probleme:

Prezenţa în SERP a unor site-uri care:

- direcţionează greşit utilizatorii;
- direcţionează utilizatorii spre un scraper site.

Consecinţe:

- scăderea performanţei motoarelor de căutare;
- nemulţumirea utilizatorilor referitoare la informaţia găsită

SOLUŢII

Identificarea scraper site-urilor:

- algoritmi automaţi;
- pe baza feedback-ului utilizatorilor - programe de învăţare automată;
- pe baza similarităţii dintre conţinutul aflat la pagina care face parte dintr-un scraper site şi pagina sursă de la care a fost preluat conţinutul.

REZULTATE

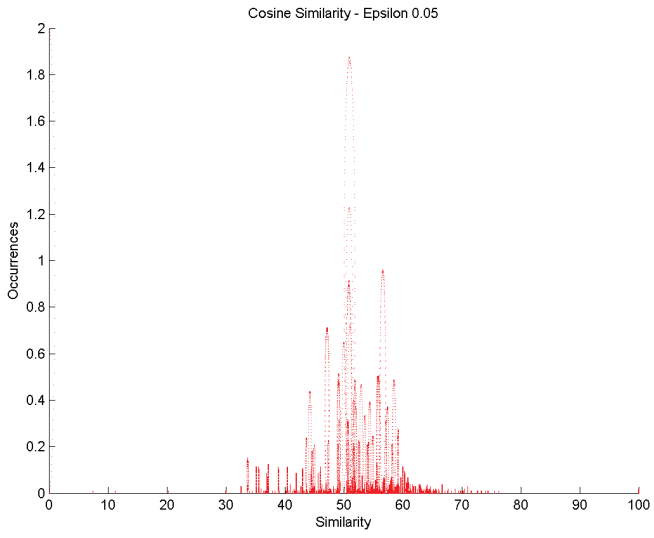


Figure: Non-Scraper site: Similaritatea Cosinus - $\epsilon = 0.05$

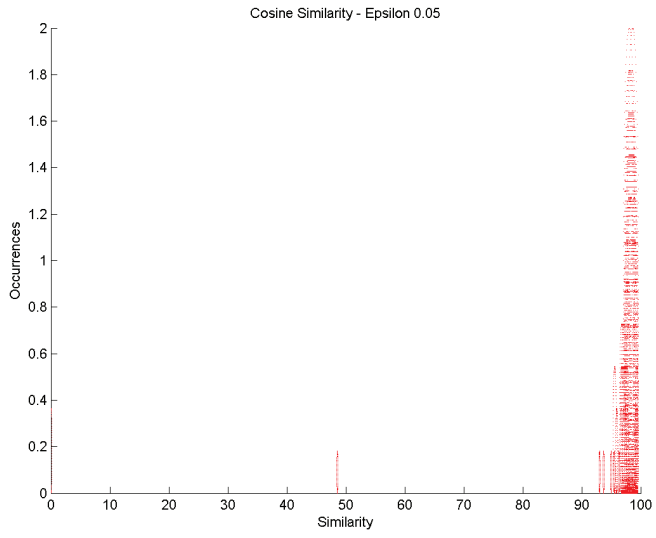


Figure: Scraper site: Similaritatea Cosinus - $\epsilon = 0.05$

Concluzii

Migrarea unui site web

- ponderarea proprietăţilor conţinutului prezentat la un anumit URL;
- redirecţionarea vizitatorului spre o pagină similară cu cea a referrer-ului.

Similaritate şi bouncerate

- analiza similarităţii conţinuturilor $\Rightarrow$ (referrer, $\forall$ link intern);
- funcţii de similaritate semantică;
- ponderarea funcţiilor de similaritate;
- ponderarea unor proprietăţi specifice ale conţinutului.

Analiza Formală Conceptuală

FCA

Definiţie

*Un **context formal** este un triplet (G, M, I), unde:*

- *G reprezintă mulţimea obiectelor;*
- *M reprezintă mulţimea atributelor;*
- *$I \subset G \times M$ reprezintă o relaţie binară între mulţimea obiectelor şi mulţimea atributelor,numită relaţie de incidenţa*

Observaţie

***gIm** se citeşte **obiectul g are atributul m**.*

FCA

Definiţie

Operatorul de derivare *in FCA este o conexiune Galois între mulţimile G şi M:*

- $A' = \{m \in M | \forall g \in A, gIm\}, A \subseteq G;$
- $B' = \{g \in G | \forall m \in B, gIm\}, B \subseteq M.$

3FCA

Definiţie

*Un **context formal triadic** este un cvadruplu* (K_1, K_2, K_3, Y)*, unde:*

- K_1 *reprezintă mulţimea obiectelor;*
- K_2 *reprezintă mulţimea atributelor;*
- K_3 *reprezintă mulţimea condiţiilor;*
- Y *este o relaţie ternară,* $Y \subseteq K_1 \times K_2 \times K_3$.

Observaţie

***(g,m,b)** se citeşte **obiectul g are atributul m sub condiţia b**.*

3FCA

Definiţie

*Un **triconcept** al unui tricontext (K_1, K_2, K_3, Y) este un triplet maximal (A_1, A_2, A_3) cu $A_i \subseteq K_i$.*

Observaţie

Pentru un triconcept (A_1, A_2, A_3):

- *A_1 se numeşte **extent**-ul triconcept-ului;*
- *A_2 se numeşte **intent**-ul triconcept-ului;*
- *A_3 se numeşte **modus**-ul triconcept-ului.*


Web Usage Mining

Web Usage Mining

- tehnicile din Data Mining precum şi cele din Statistică au fost utilizate pentru a extrage informaţii utile din log-urile web;
- tipologia site-urilor:
  - site-uri de e-commerce - **SCOP**: - vânzarea de produse;
  - site-uri de e-learning - **SCOP**: - oferirea de informaţii.
- o vizită pe un site educaţional nu se aplica euristicilor utilizate de majoritatea instrumentelor de analiză.

Rezultate

- **platforma de e-learning**: PULSE;
- **perioada de colectare a datelor**: Februarie - Iulie 2013;
- **ce investigăm**: tipare de comportament al utilizatorilor care folosesc PULSE;
- **date folosite**: clasele URL-urilor accesate, clase de referreri, timestamp-urile accesărilor username-ul studenţilor;
- ELBA, TOSCANA & Toscana2TRIAS.


CLASELE URL-URILOR ACCESATE (AF_CLASS)

- URL-urile accesate au fost împărţite în **9 clase**.

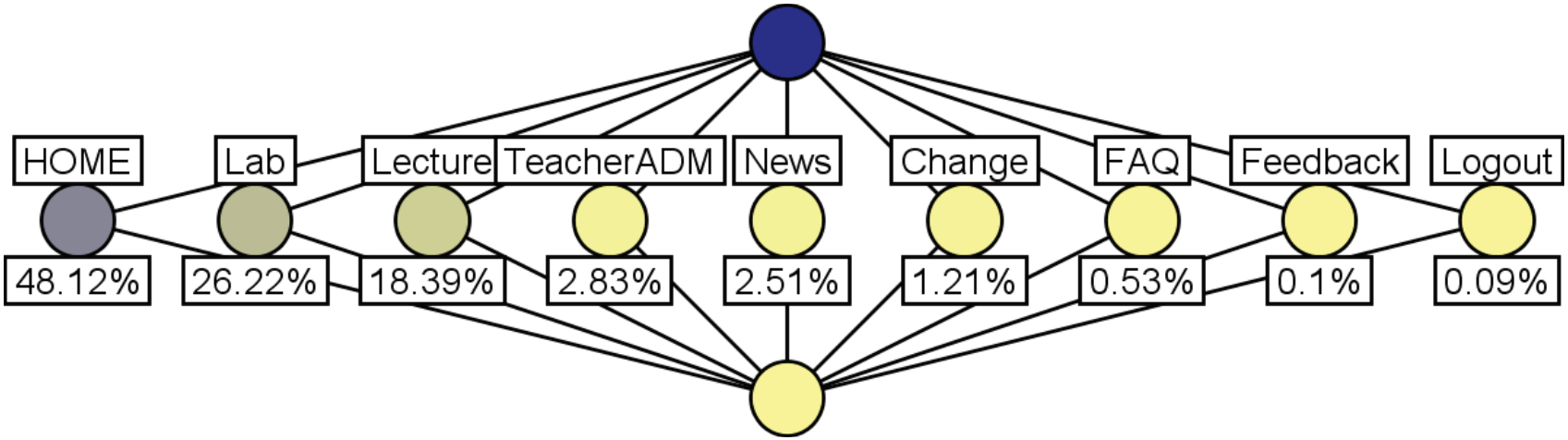


Figure: Clasele URL-urilor accesate

CLASE DE REFERRERI (R_CLASS)

- URL-urile referrer-ilor reprezintă paginile dinspre care utilizatorii vizitează o pagină dintr-un site;
- referrer-ii externi platformei PULSE nu au fost consideraţi în acest studiu.

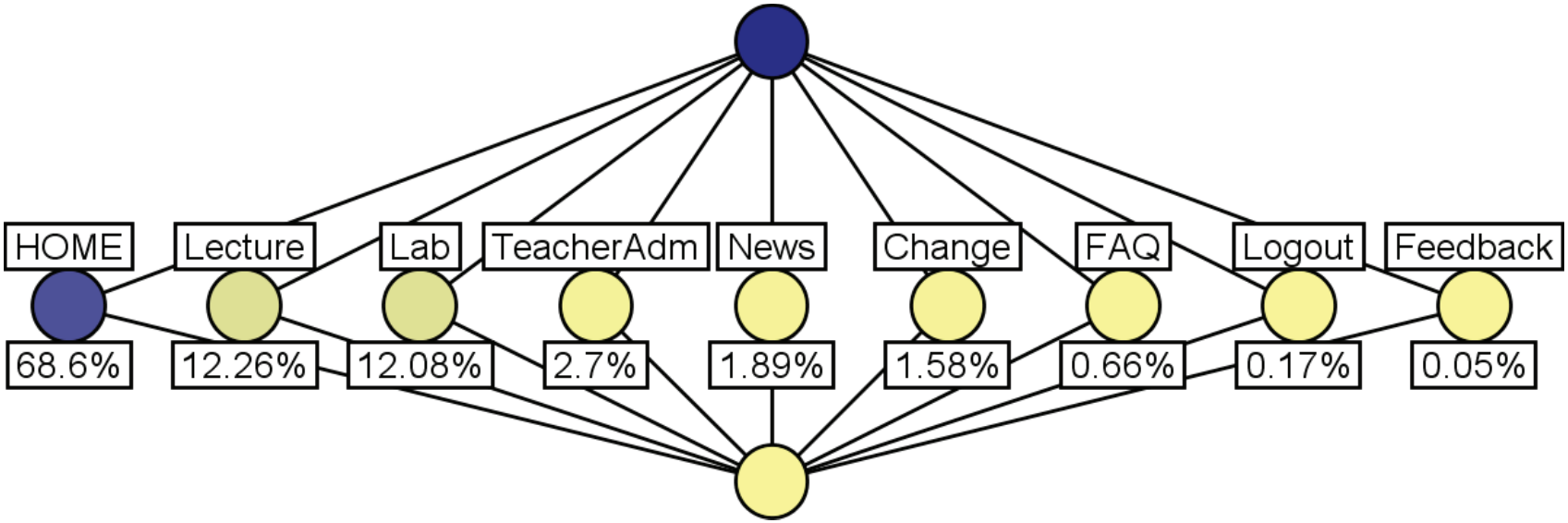


Figure: Clase de referreri

Web Usage Mining folosind 3FCA

- Toscana2Trias permite selectarea datelor triadice pornind de la un set de scale preprocesate în ToscanaJ;
- Date de intrare pentru Trias:
  - atribute: perechea (R_CLASS - AF_CLASS);
  - conditii: timestamp;
  - obiecte: studenti.

Interpretarea rezultatelor obţinutelor în 3FCA folosind CIRCOS

Interpretarea rezultatelor obţinutelor în 3FCA folosind CIRCOS

Observaţie

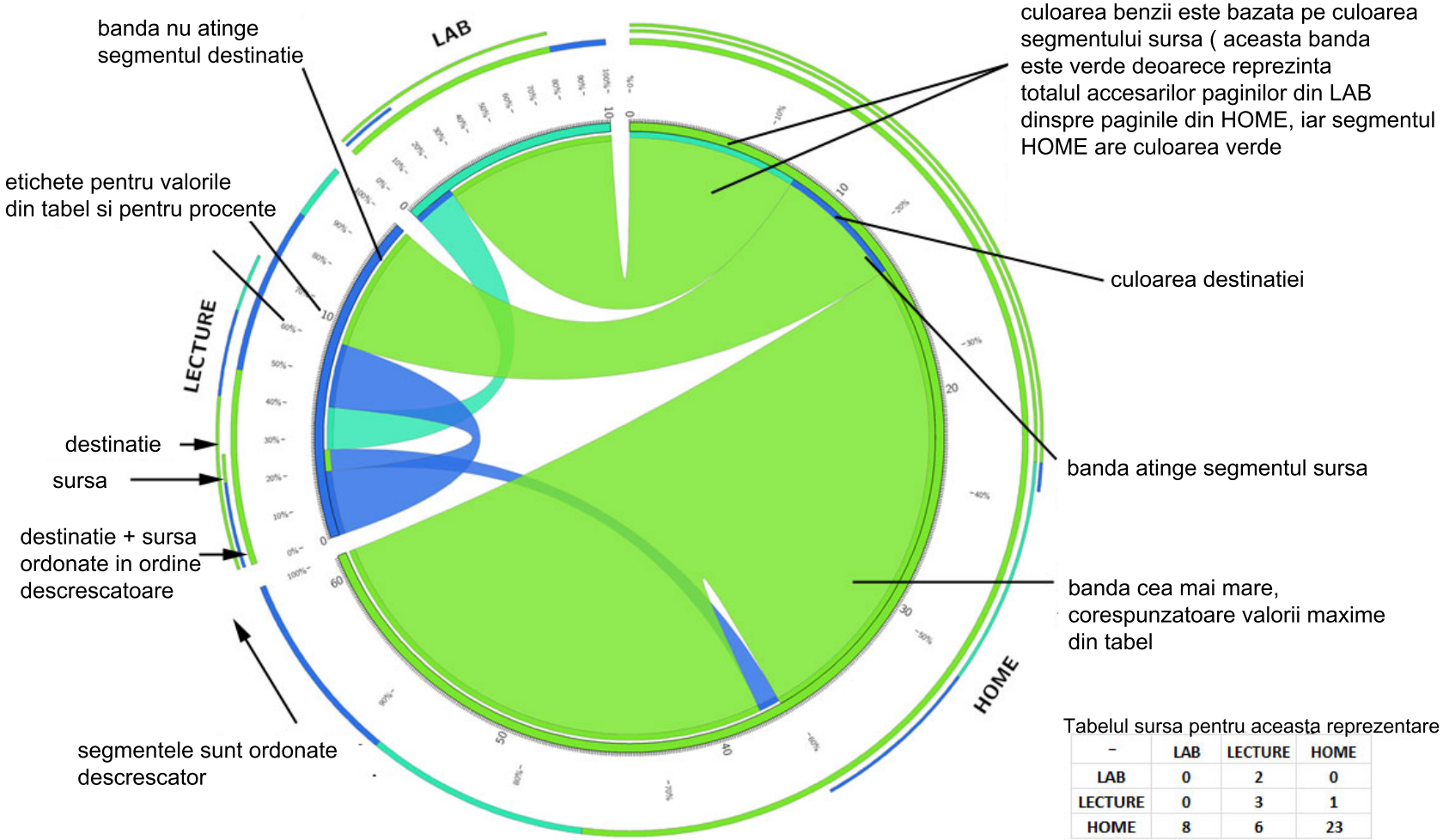
- *XML-ul generat de TRIAS conţine toate triconceptele derivate din tricontext;*
- *fiecare triconcept e definit de extent, intent şi modus;*
- *datele de intrare pentru CIRCOS trebuie reprezentate într-un tabel bidimensional, $R \times C$.*

INTERPRETAREA REZULTATELOR OBŢINUTELOR ÎN 3FCA FOLOSIND CIRCOS

Observaţie

- *(G,M,B,Y) $\Leftrightarrow$ (G,(B,M),I), $(g,(b,m)) \in I \Leftrightarrow (g,m,b) \in Y$;*
- *$\forall$ (b,m) $\Rightarrow$ numărul de elemente din extent este (b,m)';*
- *C = mulţimea care defineşte indicatorii de coloană, obţinută prin proiectarea relaţiei Y pe M;*
- *R = mulţimea care defineşte indicatorii de linie, obţinută prin proiectarea relaţiei Y pe B;*
- *valorile numerice ale tabelului se calculeaza astfel: $\forall(c,r) \in C \times R$, numărul de elemente din extent (c,r)' este calculat direct din XML-ul rezultat din Trias.*


INTERPRETAREA REZULTATELOR OBŢINUTELOR ÎN 3FCA FOLOSIND CIRCOS



Tabelul sursa pentru aceasta reprezentare

-	LAB	LECTURE	HOME
LAB	0	2	0
LECTURE	0	3	1
HOME	8	6	23

Figure: Grupa: ”ar” - săptămână

Rezultate

Test specification

- **durata**: un semestru;
- **date**: grupe de studentţi;
- **triconcepte**:
  - **obiecte**: R_class;
  - **atribute**: AF_class;
  - **condiţii**: timestamps;
- **comportament**:
  - relaxat;
  - intens;
  - normal.

Rezultatele obţinute sunt publicate:
`http://www.cs.ubbcluj.ro/~fca/tests-2013/`

REZULTATE

- comportamentul relaxat:
  - apare în special în timpul vacanţelor;
  - **observaţie**: număr redus de URL-uri accesate;
  - **tipar**: HOME $\Rightarrow$ LAB $\Rightarrow$ LECTURE;
  - **tipar**: HOME $\Rightarrow$ LECTURE $\Rightarrow$ HOME;

REZULTATE

- comportamentul normal:
  - apare în timpul semestrului, când nu sunt examene sau vacanţe;
  - **observaţie**: aproape toate clasele de URL-uri accesate sunt vizitate;
  - paginile din clasa LAB sunt cele mai vizitate.

REZULTATE

- comportamentul intens:
  - apare în special în perioadele cu examene;
  - **observaţie**: număr mare de accesăr;
  - paginile din clasa LECTURE sunt cele mai vizitate;
  - HOME reprezintă un liant între restul facilităţilor oferite de PULSE;


REZULTATE - GRUPA: "AR"

Figure: SO: comportament relaxat

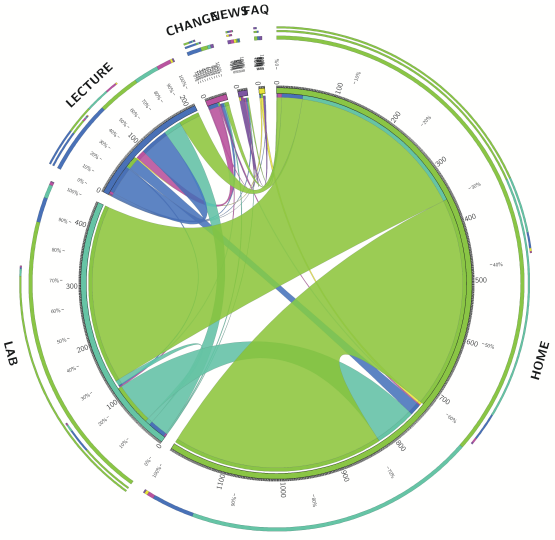


Figure: SO: comportament normal

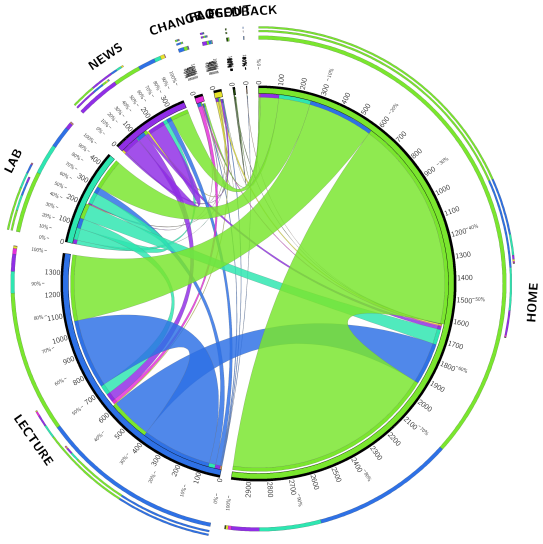


Figure: SO: comportament intens

Rezultate - grupa: "ri"

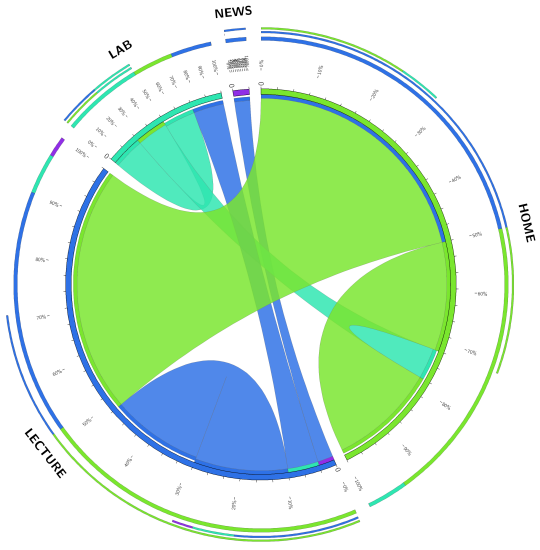


Figure: SO: comportament relaxat

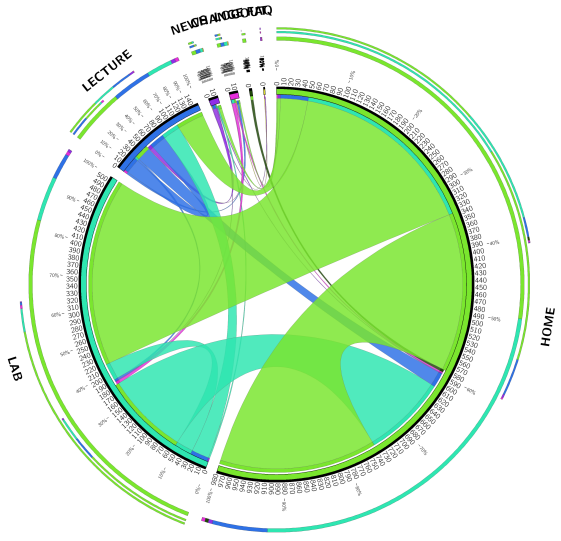


Figure: SO: comportament normal

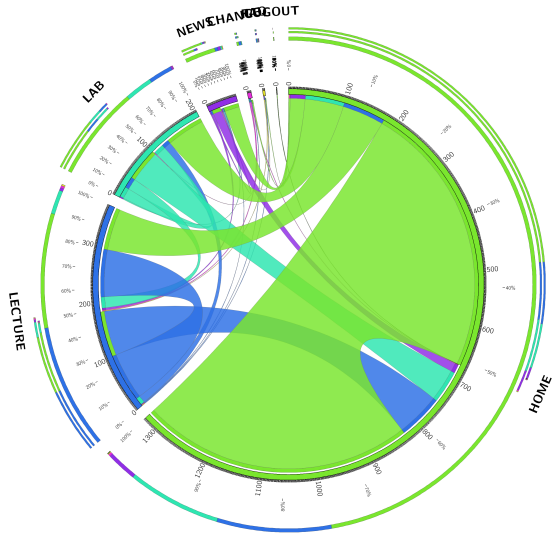


Figure: SO: comportament intens

Rezultate - grupa: "IE"

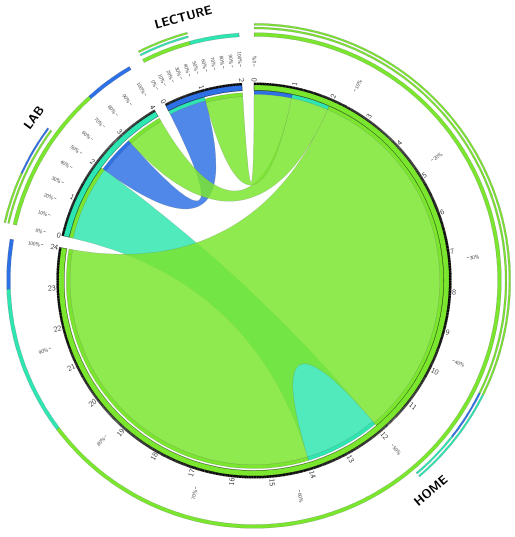


Figure: WDO: comportament relaxat

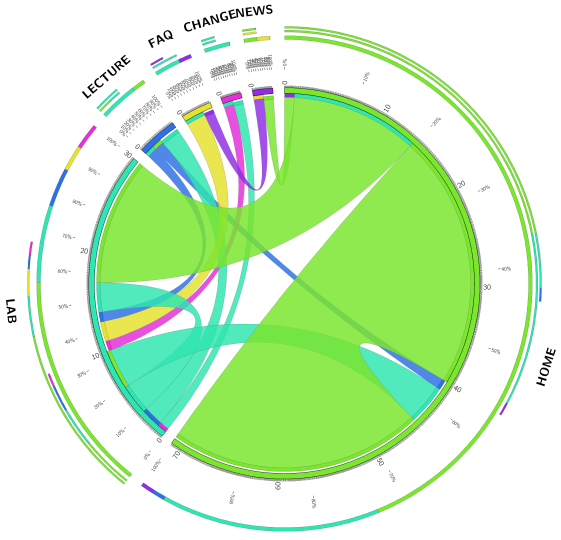


Figure: WDO: comportament normal

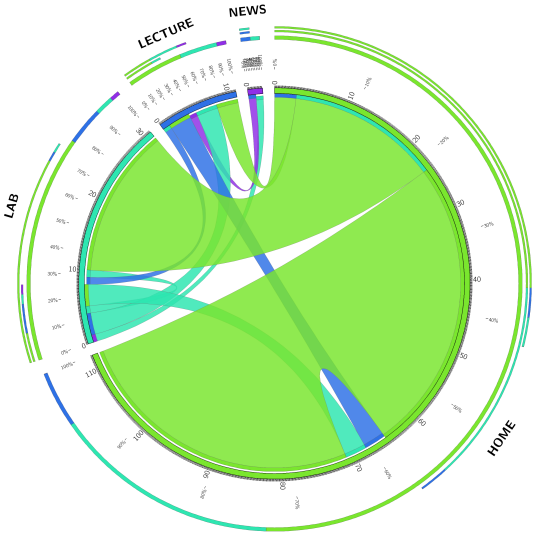


Figure: WDO: comportament intens

REZULTATE - GRUPA: "EI"

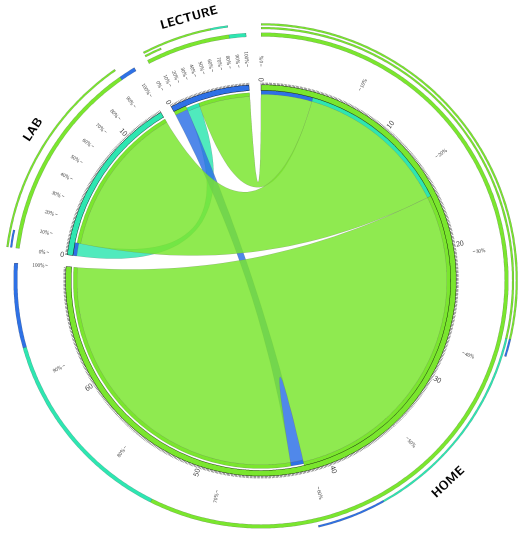


Figure: WDO: comportament relaxat

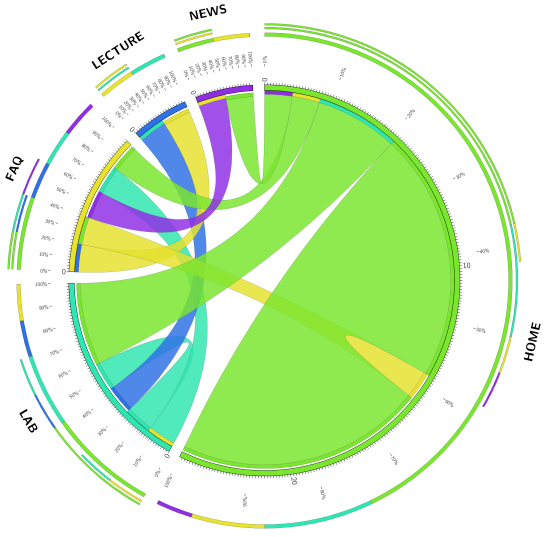


Figure: WDO: comportament normal

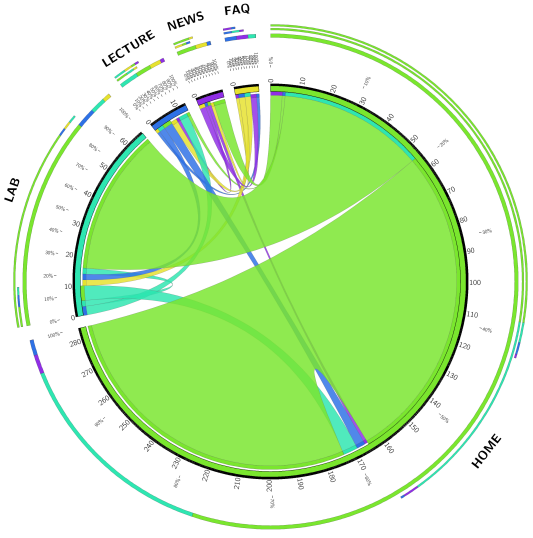


Figure: WDO: comportament intens

Concluzii

- ajustarea parametrilor în fişierele de configurare ale CIRCOS, precum şi afişarea datelor într-un alt format decât cel circular.
- construirea unui algoritm bazat pe probabilităţi condiţionale pentru a şterge conceptele cele mai puţin importante.
- utilizarea TCA pentru a evidenţia din punct de vedere temporal comportamentul utilizatorilor, atât individual cât şi pe grupuri.

Mulţumesc!

Q & A