Impactul similarităţii documentelor web asupra traficului

Absolvent: Diana - Florina HALIŢĂ
Coordonator ştiinţific: Lect. Dr. Darius - Vasile BUFNEA
Universitatea "Babeş-Bolyai" Cluj-Napoca

1 Iulie 2014

CUPRINS

DE CE?

De ce să alegem un CMS?

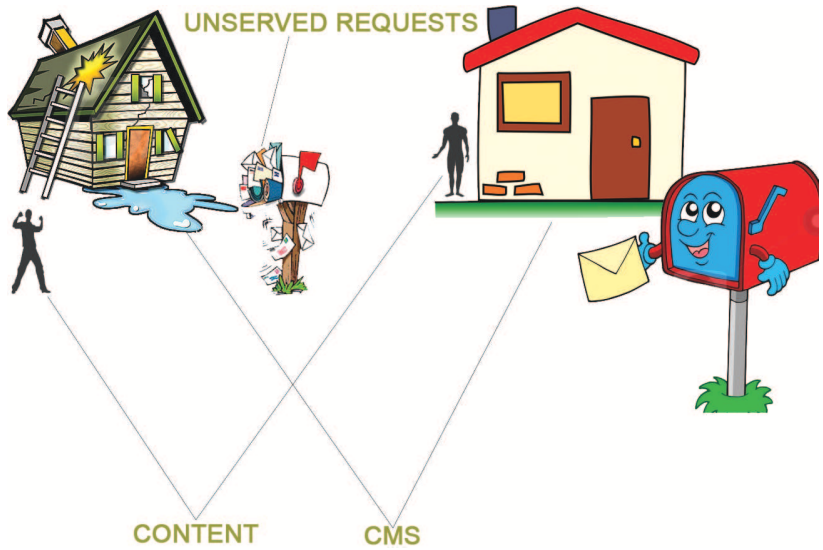
- Întreţinerea facilă - web;
- Independenţa conţinutului de prezentare;
- Update-uri periodice de securitate;
- Roluri multiple pentru utilizatori;

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

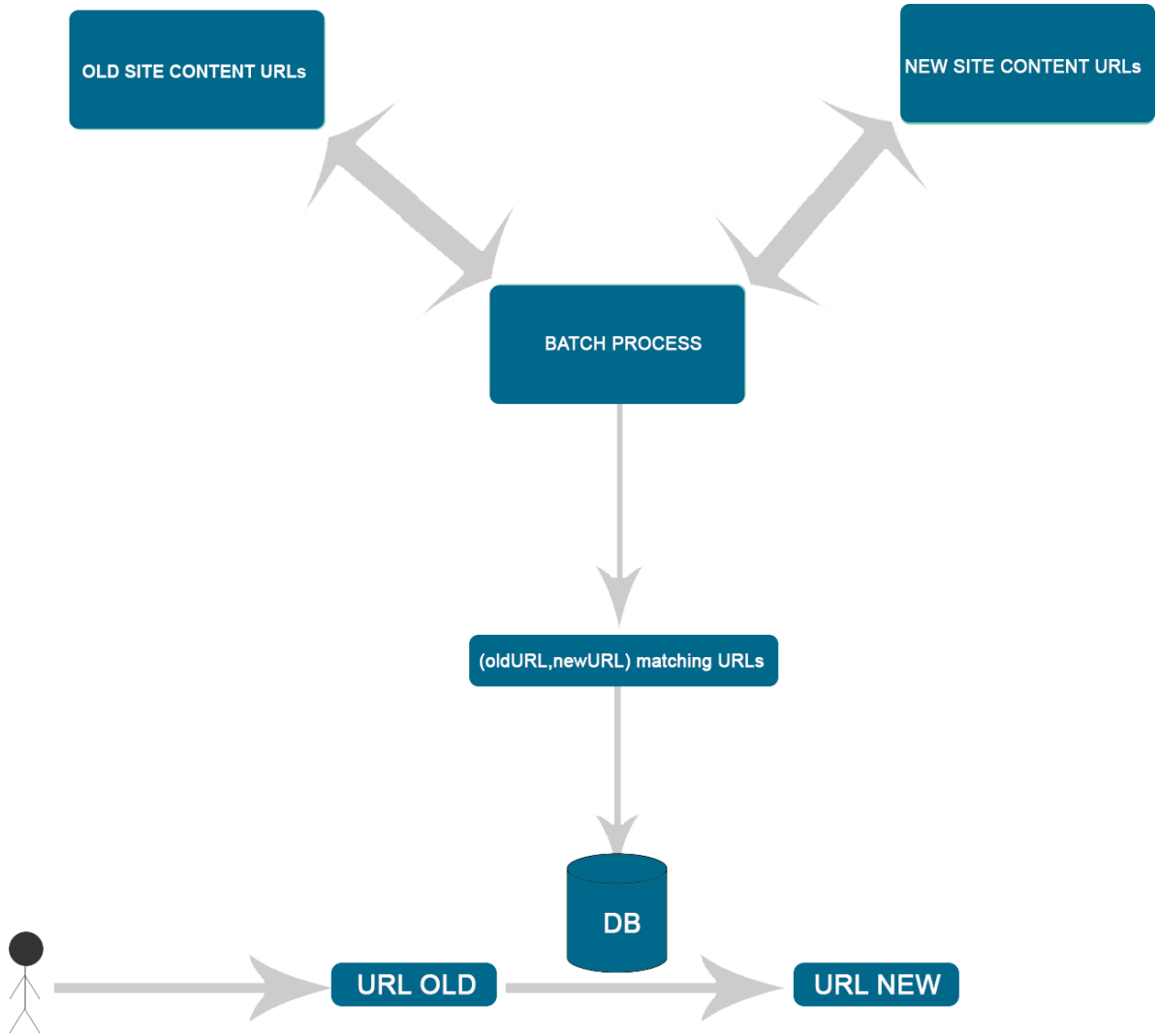


Figure : Batch processing

REZULTATE

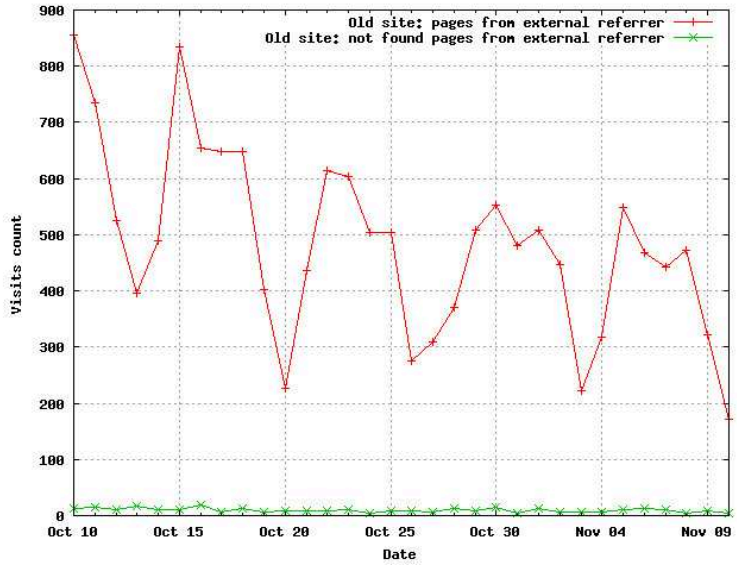


Figure : Vechiul site: numărul de pagini accesate de către un referrer extern şi numărul de pagini care au generat eroarea 404 şi au fost accesate de către un referrer extern

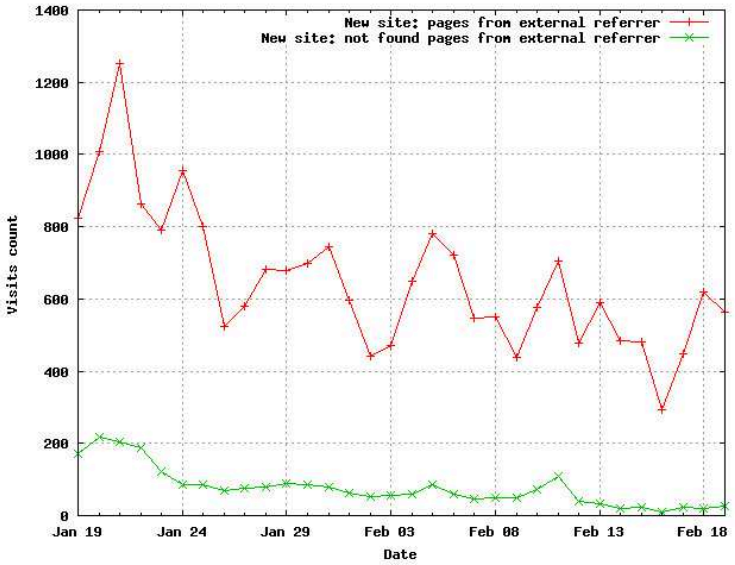


Figure : Noul site: numărul de pagini accesate de către un referrer extern şi numărul de pagini care au generat eroarea 404 şi au fost accesate de către un referrer extern

REZULTATE

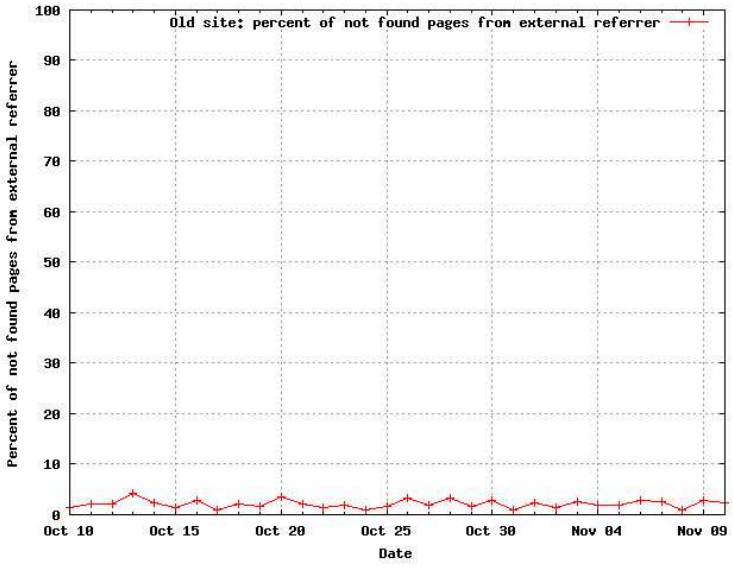


Figure : Vechiul site: procentul paginilor ce au generat eroare 404 şi au fost accesate de către un referrer extern

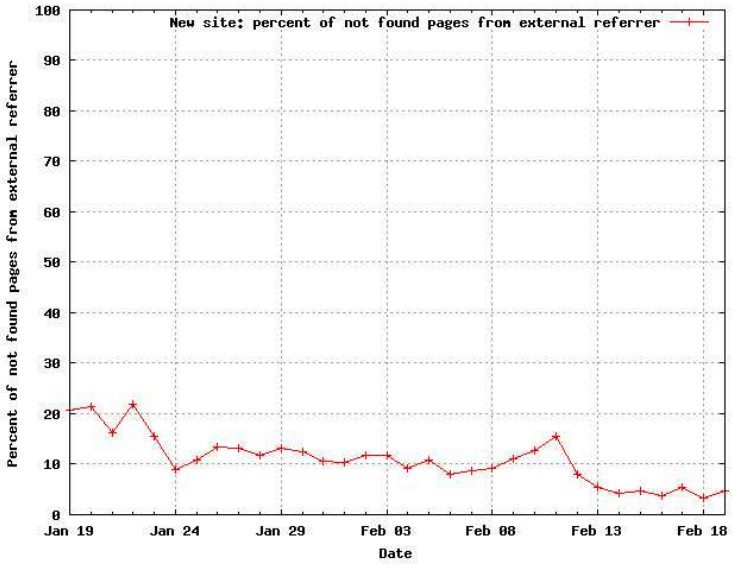


Figure : Noul site: procentul paginilor ce au generat eroare 404 şi au fost accesate de către un referrer extern

REZULTATE

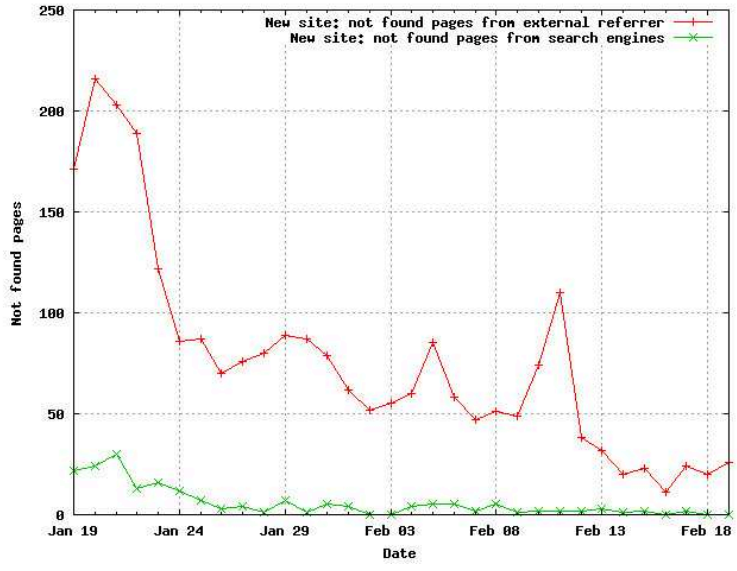


Figure : Noul site: adaptarea motoarelor de căutare la noua structură a site-ului

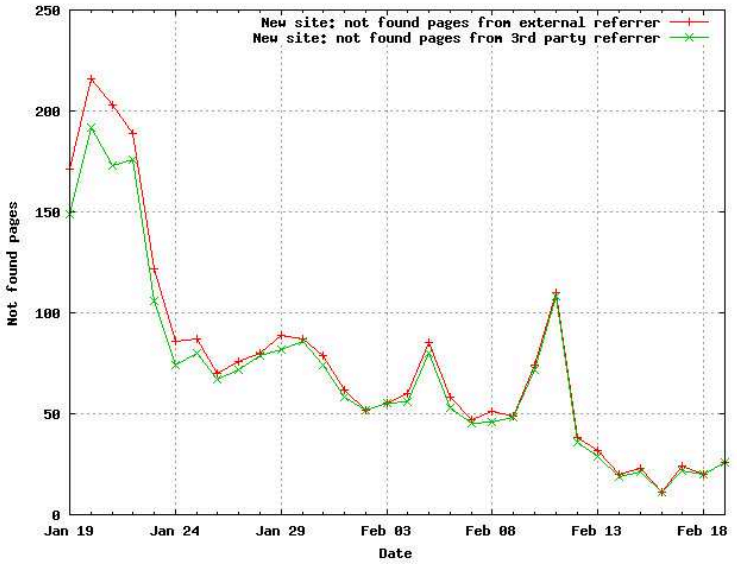


Figure : Noul site: majoritatea erorilor 404 sunt generate de către utilizatorii care vin de la referreri 3rd party

REZULTATE

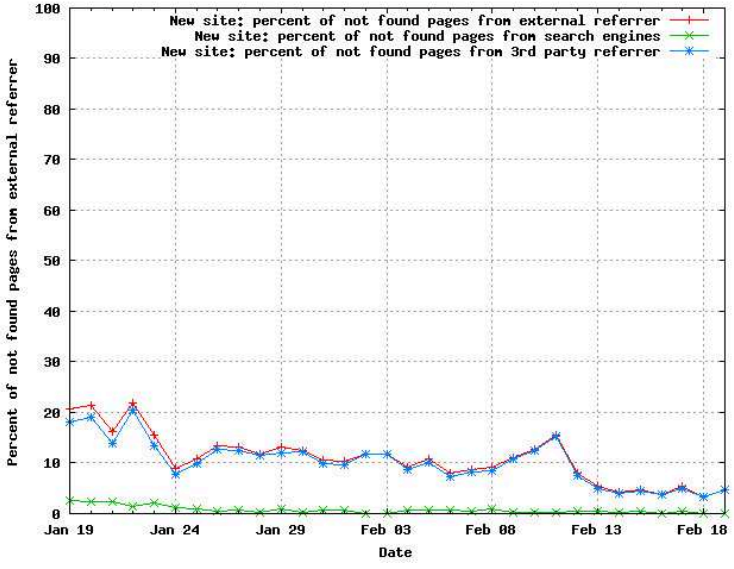


Figure : Noul site: procentul paginilor ce generează eroarea 404 şi procentul paginilor ce generează eroarea 404 şi vin de la motoarele de căutare sau de la referreri 3rd party

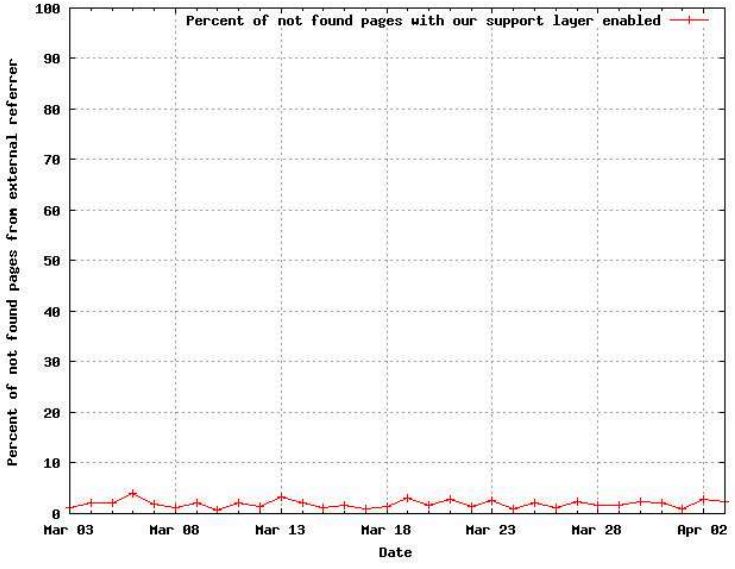


Figure : Procentul paginilor ce generează eroarea 404 având un referrer extern, odată ce a fost activat layer-ul suport propus

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

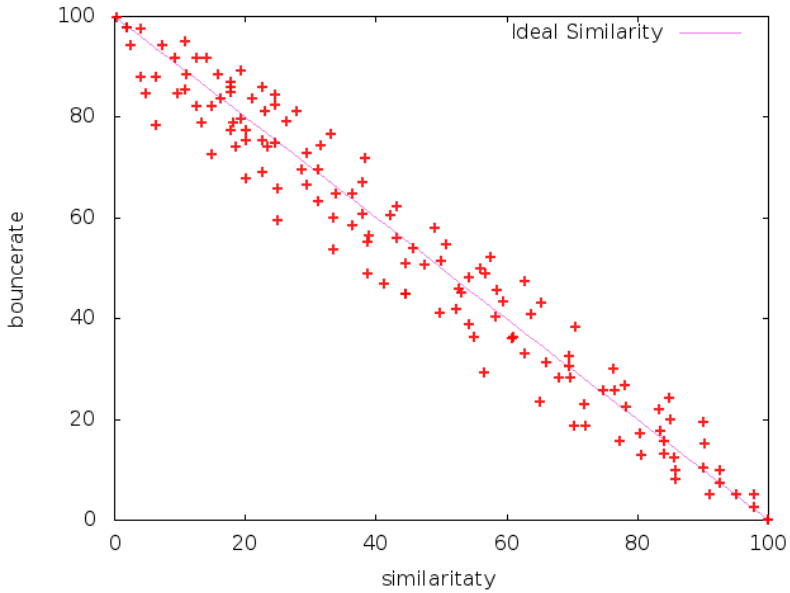


Figure : Similaritate ideală

- 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

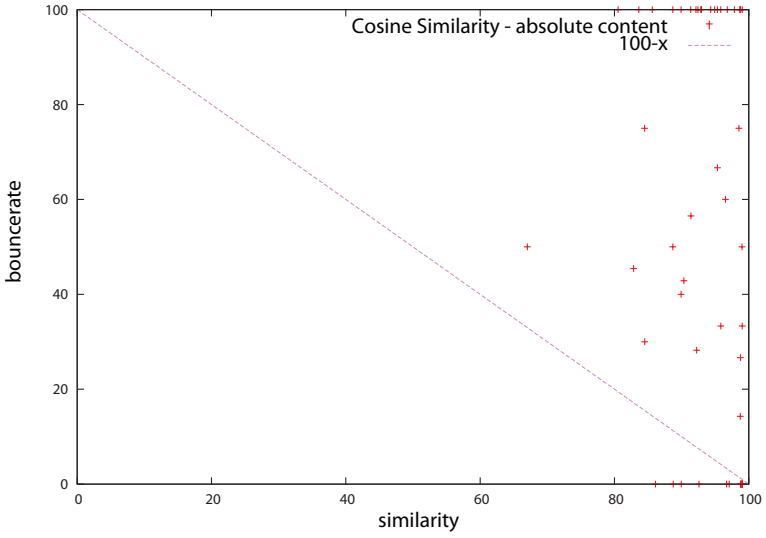


Figure : Similariatea Cosinus

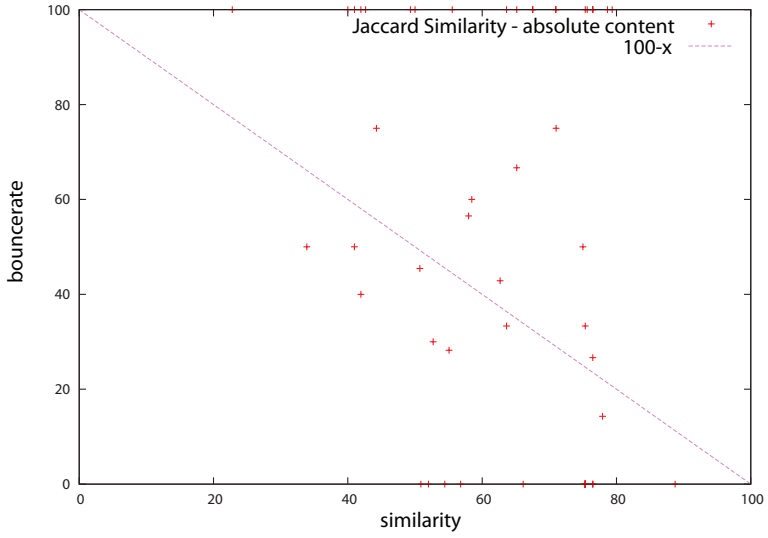


Figure : Similaritatea Jaccard

REZULTATE

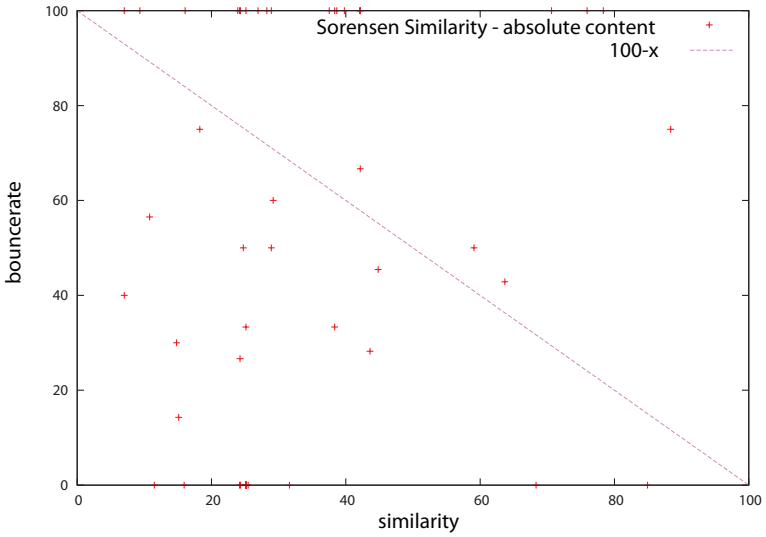


Figure : Similariatea Sorensen

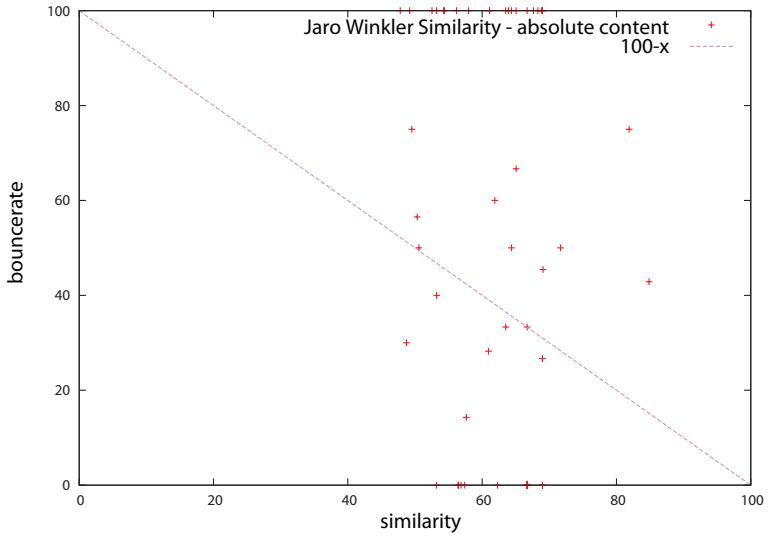


Figure : Similaritatea Jaro Winkler

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 $y = x - 100$

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

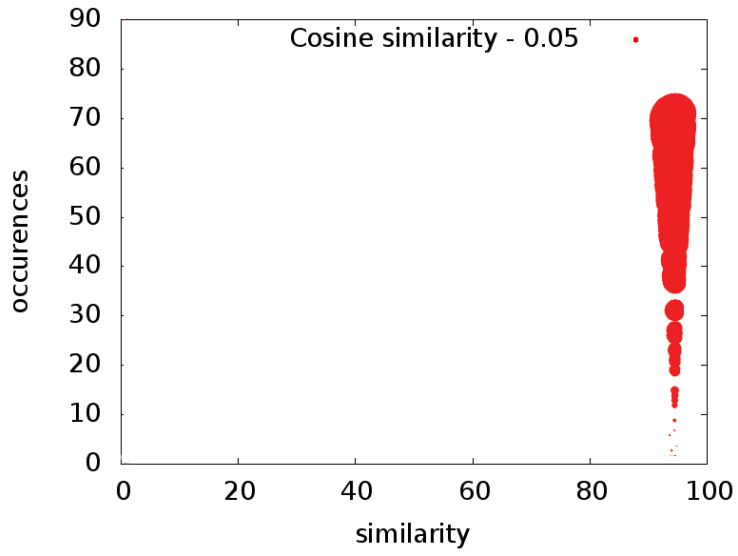


Figure : Scraper site:
Similaritatea Cosinus - $\epsilon = 0.05$

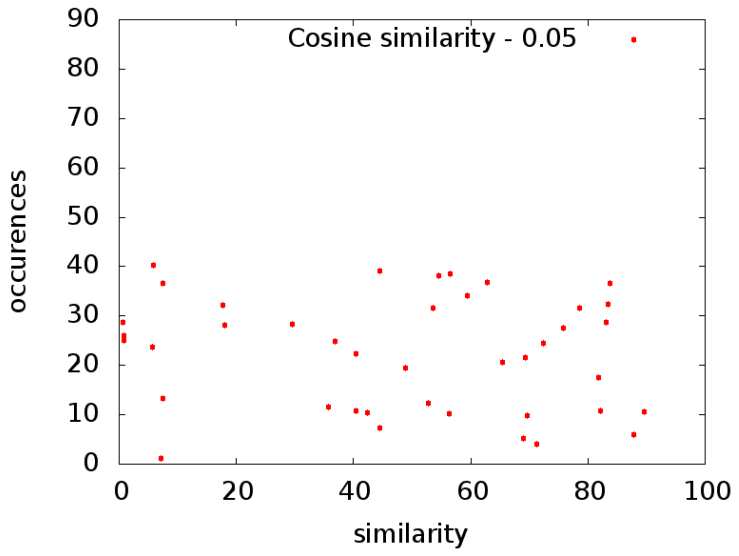


Figure : Non-Scraper site:
Similaritatea Cosinus - $\epsilon = 0.05$

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.

Mulţumesc!

Q & A