(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) **RO 134591 A2**
(51) **Int.Cl.**
G06Q 30/02 (2012.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2019 00190**

(22) Data de depozit: **26/03/2019**

(41) Data publicării cererii:
27/11/2020 BOPI nr. **11/2020**

(71) Solicitant:
• **ZA CLOUD S.R.L.**, *STR. GOVORA, NR.16A, AP.16, CLUJ-NAPOCA, CJ, RO*

(72) Inventatori:
• **ŞTEŢCO EMIL IOAN**, *STR.GOVORA, NR.16A, AP.16, CLUJ-NAPOCA, CJ, RO;*
• **BARA GEORGE ANTONIU**, *STR.HAŢEG, NR.28, BL.K2, AP.24, CLUJ- NAPOCA, CJ, RO;*
• **SUCIU MIHAI ALEXANDRU**, *ALEEA PADIN, NR.20, AP.61, CLUJ-NAPOCA, CJ, RO*

(54) # METODĂ DE PUBLICITATE ONLINE DE TIP SEMANTIC

(57) **Rezumat:**

Invenţia se referă la o metodă de publicitate online. Metoda conform invenţiei furnizează reclame pe baza contextului utilizatorului care este dedus prin analiza semnatică a textului relevant din pagina web vizualizată de utilizator folosind metode de procesare a limbajului natural. Metadatele pe baza cărora se construiesc regulile de livrare sunt extrase din textul relevant folosind taxonomii standard, iar extragerea entităţilor şi a sentimentului dedus, adică negativ, neutru sau pozitiv, se face cu algoritmi de învăţare profundă.

Revendicări: 5
Figuri: 2

Fig. 1

RO 134591 A2

Metodă de publicitate online de tip semantică

Invenţia se referă la o metodă de publicitate online de tip semantic care se bazează pe algoritmi de inteligenţă artificială şi de teoria jocurilor (Game Theory). Reclamele sunt servite utilizatorului pe baza contextului acestuia. Contextul este dedus pe baza datelor extrase din pagina online vizualizată de utilizator.

Internetul oferă un mediu versatil pentru publicitate. Un avantaj pentru publicitatea online faţă de mediile tradiţionale (ex. afişe pe marginea drumului, ziar, etc.) este personalizarea conţinutului livrat. Conţinutul adaptat utilizatorului este mult mai eficient pentru furnizorul de publicitate deoarece reclamele furnizate sunt mai de interes pentru utilizator. Probabilitatea ca un utilizator care beneficiază de conţinut personalizat să ignore reclame afişate în strânsă legătură cu pagina online vizualizată este mai mică faţă de un utilizator care primeşte conţinut standard, fără corelarea conţinutului vizualizat cu reclamele afişate. Astfel, în primul caz reclamele furnizate vor genera venituri mai mari pentru furnizorul de publicitate, iar în al doilea caz reclamele furnizate vor genera mai multe costuri pentru advertiser, costuri generate de vizualizări (impressions) fără acţiune directă asupra reclamelor. De asemenea reclamele personalizate pot reduce costurile pentru furnizor deoarece acestea sunt distribuite la un număr mai mic de utilizatori cu probabilitate foarte mare de a fi interesaţi de reclamele afişate care sunt legate exact de conţinutul paginii online vizualizată în acel moment.

Abordarea semantică se bazează pe analiza automată a conţinutului pus la dispoziţie de către publisher (entitatea care integrează anunţurile publicitare în conţinutul online publicat) unde va fi afişat anunţul digital şi selectarea unui anunţ cât mai relevant pentru pagina/aplicaţia respectivă. Acest algoritm este unul bazat pe analiză predictivă şi implică folosirea de Procesare a Limbajului Natural (NLP), Învăţare Automată (Machine Learning) şi a Teoriei Jocurilor (Game Theory). Teoria Jocurilor este o ramură a matematicii aplicate care abordează problema comportamentului optim în jocurile cu două sau mai multe persoane, într-un cadru descris de un ansamblu de reguli precise care stabilesc posibilităţile de acţiune ale fiecărui jucător, precum şi modul cum li se acordă acestora, în final, anumite valori.

Scopul abordării semantice este de a corela intenţia utilizatorului, care se poate determina din corelarea conţinutului pe care acesta îl vizualizează într-un anumit moment şi / sau istoricul preferinţelor sale cu un material de marketing relevant pentru intenţie, livrat într-un mod personalizat şi cu efect emoţional. Rezultatul abordării semantice poate fi cuantificat prin numărul de numărul de click-uri pe reclamă, numărul de vizualizări sau alte acţiuni ale utilizatorului pe pagină.

Prezenta invenţie are ca scop furnizarea de reclame publicitare online utilizatorilor pe baza conţinutului digital explorat de aceştia şi protejarea mărcilor afişate pe baza sentimentului conţinutului paginilor vizualizate.

În scopul furnizării reclamelor într-un mod personalizat se cunoaşte de asemenea patentul US 0059713 A1/2012 *Matching advertisers and users based on their respective intents*, ce prezintă o metodă de selecţie a scopului unui advertiser pe baza informaţiilor primite de la utilizator. Dezavantajul metodei

7

constă în atributele statice pe baza cărora se deduce intenţia utilizatorului şi nu pe baza conţinutului paginii vizualizate de utilizator la un moment dat.

Figura 1 prezintă platforma propusă. Textul din pagina accesată de utilizator este procesat. Pe baza meta-datelor NLP extrase se va alege reclama potrivită din inventarul de reclame corespunzătoare unei campanii de publicitate şi va fi livrată în pagina utilizatorului. Teoria jocurilor intervine în alegerea celei mai potrivite reclame, astfel încât toţi actorii implicaţi (utilizator, publisher şi advertiser) să îşi optimizeze câştigurile.

Antrenarea algoritmilor de procesare de limbaj natural (NLP) s-a făcut în trei cicluri succesive, pentru a obţine o acurateţe de cel puţin 80%. Pentru fiecare ciclu de antrenare / re-antrenare s-a folosit un set de texte relevante statistic extrase dintr-un corpus de date folosind metoda de eşantionare statistica simplă pentru a minimiza cantitatea de date necesară unui ciclu de antrenare / re-antrenare.

Pentru a putea antrena un algoritm de extragere entităţi dintr-un document din limba engleză, s-a utilizat un corpus de date ce constă din 50.000 de articole extrase din feed-uri RSS ale principalelor agenţii de presă din SUA şi Marea Britanie (Fox News, CNN, Reusters, New York Times, Washington Post, BBC, The Sun etc.). Articolele vin într-un format HTML, din care - pentru corpus – este necesar doar textul principal al articolului, lăsând la o parte conţinutul irrelevant, cum ar fi reclamele sau textele laterale. Pentru a crea un corpus adecvat pentru antrenare algoritmi NLP, entităţile trebuie identificate şi adnotate în textul articolelor. Entităţile din articol sunt identificate şi marcate corespunzător. 70% din articole sunt folosite pentru antrenarea modelului iar restul sunt folosite pentru testarea acestuia. Textul este adnotat ca in următorul exemplu:

> According to the former <START:NATIONALITY> Egyptian <END> <START:TITLE> Antiquities Minister <END> <START:PERSON> Zahi Hawassa <END> , who participated in the excavations <START:TEMPORAL_TIME> 30 years <END> ago.

Pentru antrenarea modelului în limba româna s-au urmat aceeaşi paşi ca în varianta pentru limba engleză, corpusul pentru limba română este obţinut din site-uri şi agenţii de ştiri din România (adevarul.ro, hotnews.ro, stirileprotv.ro, zf.ro, mediafax.ro, agerpres.ro, etc.). S-a folosit un corpus de 20.000 de articole.

Paşii urmaţi în procesul de antrenare / re-antrenare modelului sunt:

1. Pregătirea corpusului folosit pentru antrenare şi testare.
2. Extragerea textul principal al articolului din sursa HTML a articolelor.
3. Identificarea entităţilor pentru articolele din corpus.
4. Adnotarea entităţilor.
5. Antrenarea modelului.
6. Verificarea modelului.

Modelele obţinute pentru limbile engleză şi română au fost încărcate de un micro-serviciu Java Spring care rulează sub Java Virtual Machine (JVM). Serviciul numit "nlp-service" poate fi accesat de serviciul API Gateway, care este - la rândul lui - un alt micro-serviciu, numit "api-gateway-service". API Gateway are un endpoint /rest/v1/entities. End-pointul accepta un JSON în formatul:

DocumentRequest {

content (string, optional): conţinutul în format text

contentUri (string, optional): URI către conţinut (content şi contentUri se exclud reciproc), în acest caz textul articolului este obţinut prin descărcarea articolului, după care se extrage textul din codul HTML

language (string, optional): limbajul textului în format ISO 639-3

}

Rezultatul procesării este întors în format JSON şi conţine entităţile extrase, se întoarce o listă cu entităţi, fiecare element din listă având un câmp ce descrie tipul, numărul apariţiilor şi entitate extrasă. Un posibil rezultat:


```
{
    "entities": [{
            "type": "PERSON",
            "count": 13,
            "mention": "Trump"
        }, {
            "type": "TITLE",
            "count": 10,
            "mention": "President"
        }, {
            "type": "LOCATION",
            "count": 10,
            "mention": "US"
        }, {
            "type": "ORGANIZATION",
            "count": 3,
            "mention": "Democratic"
        }, {
            "type": "NATIONALITY",
            "count": 2,
```

```
            "mention": "Americans"
        }, {
            "type": "TEMPORAL:DATE",
            "count": 1,
            "mention": "January"
        }, {
            "type": "TEMPORAL_TIME",
            "count": 1,
            "mention": "two minutes"
        }, {
            "type": "IDENTIFIER_EMAIL",
            "count": 1,
            "mention": "donald.j.trump@gov.usa.com"
        }
    ]
}
```


Setul de meta-date ce trebuie extrase pentru personalizarea semantică a reclamelor livrate conţine, pe lângă entităţile extrase, şi clasificarea conţinutului după taxonomii standard (IAB – The Interactive Advertising Bureau şi, respectiv, IPTC - News Categories Taxonomy for the Media), împreună cu sentimentul articolului (negativ, neutru, pozitiv).

Pentru antrenarea algoritmilor de clasificare IAB, IPTC şi sentiment s-au folosit algoritmi cu învăţarea profundă (Deep Learning – CNN). În învăţarea profundă (Deep Learning), o reţea neuronală convoluţională (CNN sau ConvNet) este o clasă de reţele neuronale profunde, aplicată cel mai frecvent la analizarea imaginilor vizuale. Cercetările noastre au arătat că reţelele neuronale convoluţionale se pretează foarte bine şi domeniului NLP. Corpusul de date folosit pentru antrenarea reţelelor CNN a fost clasificat automat, cu intervenţie umană pentru calibrare, iar eşantioanele folosite pentru ciclurile de antrenare / re-antrenare au fost extrase folosind aceeaşi metodă de eşantionare simplă aleatorie.

După extragerea meta-datelor NLP din pagina vizualizată de utilizator se pot livra reclamele personalizate. Soluţia permite selecţia reclamelor pe baza unor reguli predefinite. Entităţile extrase din pagina vizitată sunt potrivite cu regulile de livrare pe baza entităţilor (cu tip), clasificare a conţinutului (taxonomiile IAB şi IPTC) şi analiza sentimentului (neutru, negativ sau pozitiv).

Regulile de livrare create pot asigura siguranţa unui client de publicitate (a mărcii). De exemplu dacă sentimentul extras din text este pozitiv sau neutru faţă de marca Adidas, se vor livra reclame despre marca Adidas. Prin impunerea unui sentiment neutru sau pozitiv se asigură ca marca pentru care se face

publicitate nu este asociată cu un articol posibil negativ. Dacă sentimentul extras din pagină este negativ faţă de marca Adidas, platforma livrează un anunţ publicitar pentru altă marcă de pantofi sport.

Regulile de livrare ce folosesc clasificarea conţinutului furnizează reclame pe baza categoriei asociate cu conţinutul paginii vizitate. Dacă pentru o pagină se obţine clasificarea "Travel" (IAB) or "holiday or vacation" (IPTC), entitatea extrasă este „Zurich" de tipul LOCATION, iar articolul citit este neutru sau pozitiv faţă de o călătorie în Zurich, se vor livra reclame legate de călătorie sau vacanţe în Zurich. Se poate ţine cont şi de locaţie în alcătuirea regulilor de livrare.

Toate regulile de livrare semantice (entităţi, sentiment, clasificari) se pot combina între ele pentru a obţine reguli mai complexe pentru definirea exactă a contextului în care reclamele asociate unei campanii de publicitate să apară pentru maximizarea ROI (Return On Investment).

Figura 2 prezintă cum funcţionează serviciul de publicitate online de tip semantic bazat pe inteligentă artificială.

Pe baza metadatelor extrase algoritmul de alegere a reclamelor încearcă o potrivire între cuvintele cheie asociate reclamelor şi aceste metadate. Printr-un mecanism de licitaţie în timp real se determină preţul pentru reclamă şi se asigură echitabilitatea între actorii implicaţi.

Revendicări

1. Metodă de furnizare a reclamelor pe baza contextului utilizatorului, **caracterizată prin aceea că** contextul utilizatorului este dedus prin analiza semantica a textului relevant din pagina web vizualizată de utilizator folosind metode de procesare a limbajului natural cu Inteligenţă Artificială. Algoritmi inspiraţi din Teoria Jocurilor intervin în alegerea celei mai potrivite reclame, astfel încât toţi actorii implicaţi (utilizator, publisher şi advertiser) să îşi optimizeze câştigurile.

2. Metodă ce permite extragerea textului relevant din sursa brută, neprelucrată HTML a articolelor online, **caracterizată prin aceea că** textul relevant este extras din sursa brută folosind algoritmi euristici, independenţi de structura HTML a articolelor procesate şi a limbii folosite în conţinutul articolelor. Articolele vizualizate de utilizator sunt preluate în format brut HTML de pe site-ul public al Publisherului, din care, pentru analiza semantică pentru furnizarea reclamelor, este necesar doar textul relevant articolului lăsând la o parte părţile irelevante din pagina brută HTML, cum ar fi reclamele, textele laterale sau alte referinţe.

3. Metodă de antrenare / re-antrenare algoritmi de Inteligenţă Artificială cu învăţarea profundă (Deep Learning – CNN), **caracterizată prin aceea că** antrenarea se face în cicluri succesive, pentru a obţine o acurateţe de cel puţin 80%. Pentru fiecare ciclu de antrenare / re-antrenare se folosesc un set de texte relevante statistic extrase dintr-un corpus de date folosind metoda de eşantionare statistica simplă pentru a minimiza cantitatea de date necesară unui ciclu de antrenare / re-antrenare. Folosind un set iniţial de date, relevant statistic, este antrenat un model. Setul iniţial este obţinut folosind o metodă de eşantionare aleatoare, proporţia acestuia din corpusul datelor fiind 11%. Mai apoi, 30% din date sunt folosite pentru testare, fază în care se decide care dintre algoritmi (extragere entităţi, clasificare, sentiment) vor fi reantrenaţi în ciclurile următoare. Se repetă ciclul de antrenare/testare de un număr prestabilit (>=3) până se obţine o acurateţe satisfăcătoare (> 80%).

4. Metodă de extragere a meta-datelor NLP multi-limbă din pagina online, **caracterizată prin aceea că** textul preluat din pagina online vizualizată de utilizator este clasificat automat folosind taxonomii standard (IAB şi IPTC), iar extragerea entităţilor şi a sentimentului (pozitiv, negativ si neutru) din text se face cu algoritmi cu învăţarea profundă (Deep Learning – CNN).

5. Metodă de alegere a reclamei relevante în funcţie de meta-datele NLP extrase din pagina online vizualizată de utilizator, **caracterizată prin aceea că**, pentru fiecare reclama din inventarul de reclame asociate unei campanii, se definesc reguli de stricte de livrare a reclamelor. În plus, regulile de livrare pot fi combinate cu alte atribute statice legate de comportamentul utilizatorului, cum ar fi: ţara de origine a utilizatorului, limba acestuia, setări temporare. Algoritmi inspiraţi din Teoria Jocurilor intervin în alegerea celei mai potrivite reclame, astfel încât toţi actorii implicaţi (utilizator, publisher şi advertiser) să îşi optimizeze câştigurile. Printr-un model de licitaţie în timp real se determină preţul şi asigură echitabilitatea între jucători. Pentru a găsi oferta optimă, problema este formulată ca o problemă de optimizare.

Desene

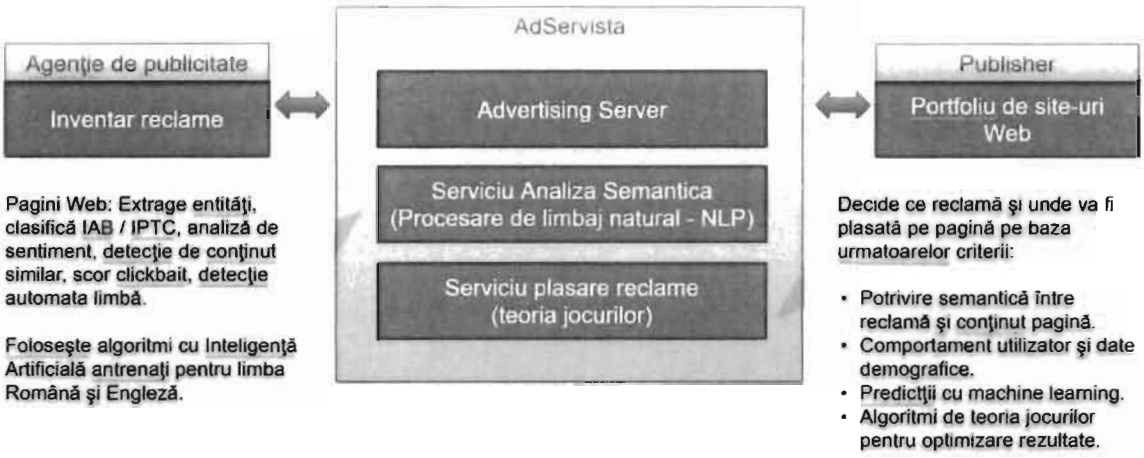

Figura 1

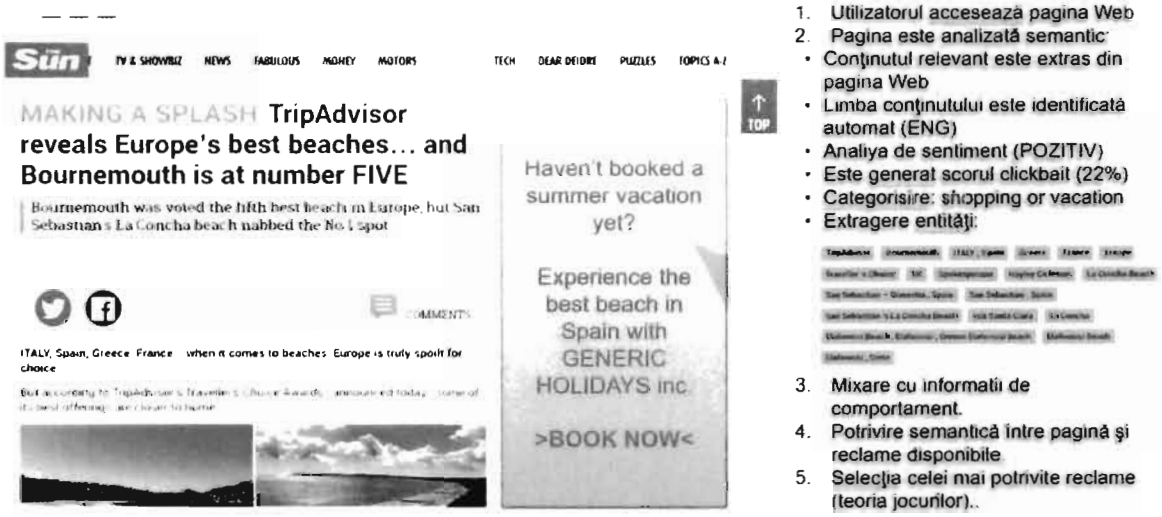

Figura 2