MINISTERUL EDUCAȚIEI, CULTURII ȘI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică și Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susținere
Șef departament:
Ion Fiodorov, conferențiar universitar, doctor în informatică
_____
"____" _____2021

Instrument de analiză a datelor textuale utilizînd tehnici Data Mining

Teză de master

Student: **Scutelnic Bogdan, TIA-191M**

Conducător: **Rogovschi Nicoleta, conf. univ.dr.**

Chișinău, 2021

ADNOTARE
la teză pentru obținerea titlului de master cu tema
La teză de master "INSTRUMENT DE ANALIZĂ A DATELOR TEXTULAE UTILIZÎND TEHNICI DATA MINING"

Scutelnic Bogdan

Specializarea Tehnologia Informației pentru Afaceri

Actualitatea temei. Data Mining este un proces de a găsi modele potențial utile din seturi de date imense. Este o abilitate multidisciplinară care folosește învățarea automată, statistici și AI pentru a extrage informații pentru a evalua probabilitatea evenimentelor viitoare. Statisticile derivate din Data Mining sunt utilizate pentru marketing, detectarea fraudei, descoperirea științifică etc.

O dată cu dezvoltarea rapidă a tehnologiilor informaţionale a apărut necesitatea dezvoltării sistemelor informaţionale, pentru a programa sisteme soft care ar putea stoca cantităţi enorme de informaţii şi ar putea reda prin comenzi simple şi uşor de asimilat date exacte despre conţinutul bazei de date, precum şi prelucrarea unor altor tipuri de date nefiind obligatoriu să prelucreze date sub formă de tabele dar pot fi date grafice sau date de avertizare sau de alt tip din domeniul informaticii, care sigur este domeniul cel mai important la ziua de azi precum şi cel mai cercetat care datorită posibilităţi vaste pe care le poate îndeplini un soft combinat cu hard. Astfel viitorul zilei de mâne al oricărui domeniu nu poate fi conceput fără folosirea unui soft al domeniului respectiv.

Structura lucrării: lucrarea este alcătuită din introducere, trei capitole, concluzii și bibliografie.

În capitolul unu sunt analizate principalele noțiuni despre data mining, text minig, depozite de date, precum și noțiunea de Big Data.

În capitolul doi sunt prezentate tehnologiile și metodele pentru explorarea datelor prin tehnici specifice proceselor data mining.

În capitolul trei este prezentată aplicația care extrage date utilizând tehnicile data minig.

Cuvinte-cheie: Data mining, text mining, depozit de date, descoperire de resurse, statistica datelor.

Scopul lucrării: Elaborarea unui instrument care va permite analiza datelor text stocate în baza prelucrării statistice și a tehnicilor Data Mining.

Tehnologii utilizate: data mining, statistica datelor, limbajul R+.

În lucrare sunt studiate metode de prelucrare a textului şi crearea modulelor de determinare a modelelor de detectare a conţinutului. Este un lucru enorm în cadrul unui sistem de căutare centralizat sau pentru un sistem inteligent. A fost creat un instrument de anilza a datelor textuale cu ajutorul tehnicilor Data Mining si desigur intrumentul a fost verificat, fiind prezentate rezultate unei anilize.

ANNOTATION

to the Master's thesis with the theme

For my master degree thesis with the topic " TEXTUAL DATA ANALYSIS TOOL USING DATA MINING TECHNIQUES"

Scutelnic Bogdan

Specialization Business Information Technology

The actuality of the subject. Data Mining is a process of finding potentially useful models from huge datasets. It is a multidisciplinary skill that uses machine learning, statistics and AI to extract information to assess the likelihood of future events. Statistics derived from Data Mining are used for marketing, fraud detection, scientific discovery, etc.

With the rapid development of information technologies came the need to develop information systems to program software systems that could store huge amounts of information and could render through simple and easy to assimilate accurate data about the contents of the database, as well as processing other types of data are not required to process data in the form of tables but can be graphical data or warning or other data in the field of informatics, which is certainly the most important field today and the most researched due to the possibilities vast that a software combined with hardware can accomplish. Thus, the future of tomorrow of any field cannot be conceived without the use of software of that field.

Structure of the thesis: the paper consists of introduction, three chapters, conclusions and bibliography.

Chapter one analyzes the main notions about data mining, text minig, data warehouses, as well as the notion of Big Data.

Chapter two presents the technologies and methods for data exploration through techniques specific to data mining processes.

Chapter three presents the application that extracts data using minig data techniques.

Keywords: Data mining, text mining, data storage, resource discovery, data statistics.

Purpose of the paper: Development of a tool that will allow the analysis of text data stored based on statistical processing and Data Mining techniques.

Technologies used: data mining, data statistics, R + language.

The paper studies methods of word processing and the creation of modules for determining content detection models. It's a huge thing in a centralized search system or for a smart system. A text data analysis tool was created using Data Mining techniques and of course the tool was verified, presenting the results of an analysis.

CUPRINS

INTRODUCERE

Domeniul gestiunii informaţiilor abordează problema organizării, stocării şi regăsirii în timp util a datelor de care dispunem despre un anumit subiect. Neîndoielnic că fiecare dintre noi se confruntă continuu cu această problemă. Agenda personală, cartea de telefoane, sunt mijloace simple de organizare, păstrare şi regăsire a datelor de care avem nevoie la un moment dat.

"Avem foarte multe date colectate, ce să facem cu ele acum?" Această problemă a devenit una obişnuită în multe organizaţii mari. Informaţia digitală este ieftin de obţinut si relativ ieftin de stocat. Dar care este scopul stocării unei cantităţi de date atât de mari? În afara argumentului legat de modalităţile convenabile de stocare a datelor în format electronic, mai există şi altul: firmele colectează date deoarece managerii "simt" că aceste date reprezintă un activ valoros, care ar putea fi folosit la un moment dat. În institutele de cercetare, datele se referă la observaţii asupra fenomenelor aflate sub studiu, culese cu atenţie. În organizaţiile economice, datele şi informaţiile se referă la pieţe, concurenţi, furnizori, distribuitori, clienţi, precum si la informaţii interne legate de procesele de producţie, organizarea muncii, depozitare etc.

Pentru multe firme din Occident, procesul de explorare şi analiză a datelor nu este unul nou. Dar programele de *data mining* permit realizarea acestor analize mult mai rapid şi, potenţial, cu o eficienţă sporită. În multe cazuri, tehnicile de explorare a datelor permit ca datele colectate pentru un anumit scop să poată fi folosite în multe alte scopuri.

Ca urmare a perfecţionării modalităţilor de colectare şi stocare a datelor, foarte multe companii mari s-au trezit deodată că stau pe un munte de date cărora ar trebui să le găsească o utilizare. De exemplu, companiile furnizoare de cărţi de credite înregistrează permanent datele referitoare la tranzacţiile comerciale, supermarketurile înregistrau date privind cumpărăturile, utilizarea cupoanelor de reduceri ş.a.m.d. Apariţia necesităţii de a valorifica toate aceste date era doar o chestiune de timp.

Chiar dacă proiectanţii interfeţelor şi a conţinutului web-ului includ comportamentul utilizator şi câteva deprinderi în unele motoare de căutare, utilizatorii pot avea dificultăţii în efectuarea de căutări pe web deoarece aceştia nu sunt capabili să formuleze interogările corecte care să reducă numărul de rezultate obţinute şi să crească calitatea acestora. De obicei interfaţa utilizator este dificil de utilizat, funcţiile de organizare ierarhică sunt aplicate static şi informaţiile de pe web sunt foarte rar structurate şi organizate pentru o recunoaştere rapidă.

Pentru a rezolva problemele prezentate mai sus acest domeniu de cercetare a crescut continuu în direcţii cum ar fi algoritmii, strategiile şi arhitecturile. Sute de idei au fost testate, unele fiind implementate iar altele existând doar ca şi prototipuri. Unele dintre aceste idei sunt: organizarea documentelor în categorii predefinite, îmbunătăţirea modului de prezentare a rezultatelor, monitorizarea unei pagini specifice, ajutarea utilizatorului in formularea interogării corecte şi dezvoltarea de programe pentru filtrarea rezultatelor căutării.

O dată cu dezvoltarea rapidă a tehnologiilor informaţionale a apărut necesitatea dezvoltării sistemelor informaţionale, pentru a programa sisteme soft care ar putea stoca cantităţi enorme de informaţii şi ar putea reda prin comenzi simple şi uşor de asimilat date exacte despre conţinutul bazei de date, precum şi prelucrarea unor altor tipuri de date nefiind obligatoriu să prelucreze date sub formă de tabele dar pot fi date grafice sau date de avertizare sau de alt tip din domeniul informaticii, care sigur este domeniul cel mai important la ziua de azi precum şi cel mai cercetat care datorită posibilităţi vaste pe care le poate îndeplini un soft combinat cu hard. Astfel viitorul zilei de mâne al oricărui domeniu nu poate fi conceput fără folosirea unui soft al domeniului respectiv.

REFERINȚE BIBILIOGRAFICE

[1] Andone I. Tugui, A., s.a. Dezvoltarea sistemelor inteligente în economie, Ed. Economică, Bucureşti, 2011

[2] Cipolla, E. T. Data Mining: Techniques to Gain Insight Into Your Data, Enterprise Systems Journal, Decembrie 2005;

[3]

[4] Guandong Xu, Zanchun Yhang, Lin Li, Web Mining and Social Networking Techniques and Applications, USA:Springer, 2011

[5] J. Palau, M. Montaner, B. Lopez and J.L. de la Rosa. Collaboration analysis n the recommender system using social networks. In CIA, pages 137-151,2004

[6] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. Of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '98), pages 668-677, 1998

[7] Zaiane O., Han J.: WebML: Querying the World Wide Web for resources and knowledge. In: Workshop on Web Information and Data Management WIDM98, Bethesda, 1998, 9-12.

[8] Yang, Q., Zhang, H.H. and Li, I.T, 2001. Mining Web logs for prediction models in www caching and prefetching, Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 26-29, pp. 473-478.

[9] Yan, T.W., Jacobsen, M. Garcia-Molina, H. and Dayal, U., 1996, Knowledge Discovery from users web page navigation, Seventh International Workshop on Research Issues in Data Engineering, Birmingham, England, aprilie 7-8, pp 20-29.

[10] Berry, M., Linoff, G.: Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons, Chichester (1997)

[11] Dunham, M.H.: Data Mining : Introductory and Advanced Topics. Prentice Hall, Pearson Education Inc. (2003)

[12] Prinzie, A. , Van den Poel, D.: Investigating Purchasing Patterns for Financial Services using Markov, MTD and MTDg Models. In: Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 03/213 (2013)

[13] Agrawal, R., Srikant, R.: Mining sequential patterns, *International Conference on Data Engineering(ICDE'95)*, Taipei, Taiwan, pp. 3-14, martie 1995

[14] Inmon, W.H. Buiding the Data Warehouse, John Wiley & Sons, New York, 2005

[15] Turban, E., Jay Aronson J. Decision Support Systems and Intelligent Systems, Ed. Prentince Hall, New Jersey, USA, 2011