

Informatikatudomány II.
2020. október 16. 13.00-15.00
Szekciófelelős:
Forczek Krisztofer, +36 30 388 0787

Matematika- és informatikatudomány

		Informatikatudomány I.	Matematikatudomány
		Szekcióelnök: Bodroginé Dr. Zichar Marianna (DE)	Dr. Bessenyei Mihály (DE)
I. panel	10.30-10.45	Al-Kuran Hussam	Beke Ákos
	10.45-11.00	Bognár Eszter Katalin	Borsos Benjámín
	11.00-11.15	Gera Imre	Janabi, Hayder Abbas
	11.15-11.30	Ghanim Hussein Ali Ahmed	Kovács Bence Máté
	11.30-11.45	Sewunetie Walelign Tewabe	
		Informatikatudomány II.	
		Szekcióelnök: Dr. Herendi Tamás (DE)	
II. panel	13.00-13.15	Hajdu Flóra	
	13.15-13.30	Gavua Ebenezer Komla	
	13.30-13.45	Mattyasovszky-Philipp Dóra Anna	
	13.45-14.00	Padányi Viktória	
	14.00-14.15	Ősz Olivér	
	14.15-14.30	Barth Áron	

Informatikatudomány II.

2020. október 16. 13.00-15.00

Szekciófelelős:

Forczek Krisztofer, +36 30 388 0787

Fuzzy logika alkalmazása nemlineáris rendszerek érzékenységvizsgálata esetében

HAJDU Flóra, **HAJDU** Csaba

Széchenyi István Egyetem

Informatikai tudományok

hajdfl@sze.hu, hajdu.csaba@ga.sze.hu

Nemlineáris rendszerek vizsgálatának egyik izgalmas területe a paraméterek érzékenységvizsgálata. Számtalan tudományterületen alkalmazzák, mint például közgazdasági folyamatok modellezése, környezetmérnöki alkalmazások, kémiai és gyártási folyamatok modellezése, orvosi kutatások, de gépészmérnöki, mechatronikai és járműmérnöki alkalmazások esetében is elkezdett terjedni a használata. Elsősorban annak vizsgálatára alkalmazzák, hogy a paraméterek változtatása hogyan hat a rendszer viselkedésére, de paraméter identifikációs, inverz szimulációs, modell kalibrációs és modell redukciós feladatok is megoldhatók vele.

Egy rendszer érzékenységének meghatározására számtalan módszer létezik a parciális differenciális technikáktól kezdve a statisztikai módszerekig. A kutatás során elsősorban az egyparaméteres, azaz az OAT (One-at-a-Time) érzékenységvizsgálattal foglalkoztunk. A módszer előnye, hogy egyszerűen implementálható és napjaink gyors számítógépeinek segítségével nagyon sok paraméterkombináció hatékonyan, rövid idő alatt vizsgálható. A klasszikus OAT érzékenységvizsgálatot ugyanakkor csak nagyon szűk vizsgálati tartomány esetében alkalmazzák. A kutatás célja az OAT érzékenységvizsgálat kibővítése szélesebb vizsgálati tartományra, amelyhez Fuzzy logikát használtunk. A fuzzy-logika bevezetése ezen kívül a klasszikus OAT érzékenységvizsgálathoz képest egy még finomabb, szélesebb körű elemzést tett lehetővé.

Az érzékenységvizsgálat során először érzékenységi függvényeket hoztunk létre, amely egy előre kiválasztott kimeneti jel változása a paraméter változása függvényében. Az érzékenységi függvények alapján a vizsgálati tartományokat felosztottuk 4 érzékenységi fuzzy halmaz szerint, majd meghatároztuk az egyes halmazok tagsági értékét. A tagsági értékek meghatározását követően a paramétereket érzékenységi fokát elemeztük defuzzifikációs módszerek (pl. COG) használatával.

Az előadás során először ismertetjük az érzékenységvizsgálati módszereket, néhány alkalmazási területet. Ezt követően részletesen bemutatjuk az OAT érzékenységvizsgálat bővítését az általunk javasolt fuzzy logikával. Ezt követően esettanulmányokat mutatunk be elsősorban nemlineáris gépészeti rendszerek érzékenységvizsgálatáról. Az előadást további kutatási feladatokkal zárjuk.

Informatikatudomány II.

2020. október 16. 13.00-15.00

Szekciófelelős:

Forczek Krisztofer, +36 30 388 0787

Application of Kmeans and Hierarchical Agglomerative Clustering Techniques on MapReduce

GAVUA Ebenezer Komla, KECSKEMÉTI Gabor

¹József Hatvany Doctoral School of Informatics, ¹Institute of Information Technology,

¹University of Miskolc, Miskolc, Hungary, ²Department of Computer Science, Liverpool John Moores University, Liverpool, UK

gavua@iit.uni-miskolc.hu, g.kecskemeti@ljmu.ac.uk

The application of cloud computing technologies on large data sets has changed the face of data mining significantly since its inception. The development of Google File System and the Hadoop Distributed file system has enabled data storage and retrieval in clouds to be managed effectively. The Hadoop distributed framework was developed with the capabilities to handle various categories of data. However, the processing of heterogeneous data remains its major challenge. This has affected the throughput of the Hadoop MapReduce framework. The main contribution of this paper is to apply two data mining techniques on the MapReduce programming model. Most data input for Hadoop MapReduce are in their raw state. Nevertheless, already clustered data set, will quicken the rate of data processing. This is to ensure a highly clustered data set after the application of the data mining techniques. The MapReduce programming tool can then be applied to the most efficient cluster. A complex iris data set was utilized for this experiment. K-means and Hierarchical Agglomerative Clustering Techniques were applied on the dataset separately. This enabled the processing behaviour of the data set to be observed and analysed. The results generated were tested for Precision, Recall and Accuracy. These tests enabled the most appropriate model to be selected. The results proved that K-means clustering provides a high level of accuracy when employed for clustering on data sets. A combination structure between MapReduce and the two techniques was designed to highlight the experiment conducted. This design, aside highlighting the experiment, presents researchers with a structure to test other data mining techniques on MapReduce. This has set the pace for further tests to be conducted on Hadoop MapReduce to determine the processing behaviour of the generated results. Aside this, future works will directed towards to the design of a retrieval algorithm for heterogeneous data on social media. Social media is targeted because, there are several of complex data sets which will require an effective mode of retrieving units of data. These data sets are mostly a combination of audio, text, pictures and video. An efficient algorithm will ensure efficient retrieval.

Keywords: Data Mining, MapReduce, Cloud Computing

Informatikatudomány II.

2020. október 16. 13.00-15.00

Szekciófelelős:

Forczek Krisztofer, +36 30 388 0787

Cognitive Information Systems - Elements, Human-Computer Interaction & Infocommunication

MATTYASOVSZKY-PHILIPP Dóra Anna

Eötvös Loránd Tudományegyetem

Informatika

philippd@caesar.elte.hu

The fast changes in information technology and business needs led to the evolution and development of Cognitive Information Systems (CIS), due to enterprises were not able to answer fast changes related challenges easily in time, that highly impacted their long-term success. The ability to answer those challenges required that the enterprise has the capability with the specific skills available in the needed time. Those are rarely available as one single person or team, and very often they are not affordable or not efficient; therefore, one of the possible solutions is to leverage the capability of a CIS. Via Human-Computer Interaction (HCI) through Cognitive Infocommunication (CI). Wang defined a model to describe the context of systems within Cognitive Informatics. The Information - Matter - Energy (IME) Model, building a bridge between the natural world (NW) and the abstract world (AW) connected to each other, where matter (M) and energy (E). Extending Wang's model with Infocommunication, Infocommunication become as a key element in HCI. This approach via Cognitive Resonance (CR) ensures CIS capability to support the enterprise's needs, meanwhile it improves the knowledge level in both silicon and carbon agent side. There have been few pieces of research on the general model for the analysis and design of CIS. This publication attempts to create a design scheme for incorporating the various model for CISs and Understanding-based management systems (UBMSS). The components were not described as modeling element and not described as an enabling tool to create a consistent and integrated system. The most significant components for modeling are: semi-structured documents, business processes, constituents of knowledge management, the enterprise and, the information architecture including self-directing software components - Artificial Intelligence (AI) – that yield function. For CIS modeling the above-mentioned elements were combined into a unified framework that follows the object-oriented paradigm and architecture approaches.

Informatikatudomány II.
2020. október 16. 13.00-15.00

Szekciófelelős:

Forczek Krisztofer, +36 30 388 0787

Álvéletlenszám-generátorok metaanalízise*

PADÁNYI Viktória, **DR. HERENDI** Tamás

Debreceni Egyetem Informatikai Kar, Számítógéptudományi Tanszék
Matematika és informatika

padanyi.viktoria@inf.unideb.hu, herendi.tamas@inf.unideb.hu

Gyakran használnak álvéletlenszám-generátorokat a különböző elméleti és gyakorlati problémák megoldása során. Az eltérő alkalmazások más-más tulajdonságokat várnak el a generátoroktól. Legfontosabb ilyen általános tulajdonságok például a generátor periódushossza (ciklushossza), használhatósága, számítási sebessége, illetve van-e elérhető vagy letölthető implementáció az adott generátorhoz. A felhasználhatóság szempontjából érdekes még az adott generátorok minősége, amit statisztikai tesztekkel mérhetünk. Az előbb felsorolt jellemzők alapján tudjuk kiválasztani az alkalmazás szempontjából a legmegfelelőbb álvéletlenszám-generátort. A munkánk során felkutattuk a használatban lévő álvéletlenszám-generátorokat, tanulmányoztuk, hogy milyen fontos tulajdonságokkal rendelkeznek, illetve melyek az egyes alkalmazásokhoz köthető előnyös tulajdonságaik. Igyekeztünk áttekinteni a klasszikus (korai) generátoroktól kezdve a modern speciális alkalmazással rendelkező generátorokig többféle sorozatot is. Ezek a következők:

1. Négyzetközép módszerek: Neumann 1951-ben vezette be a négyzetközép módszert. Az osztályba tartozó módszerek azért érdekesek, mert ezek voltak az első gyakorlati véletlenszám-generátorok.
2. Kongruenciákon alapuló véletlenszám-generátorok: Valószínűleg ez a csoport a véletlenszám-generátorok legnagyobb osztálya. A hatványmaradékot alkalmazzák, egy szám egymás utáni hatványainak maradéka modulo m .
3. Keveréses módszerek: A seed-nek az elemeit kevergető módszer.
4. Hibrid módszerek: Más fajta módszerekből egy összetett algoritmus konstruálása, szimpla algoritmusok jó tulajdonságainak ötvözése.
5. Más, egyéb módszerek: Azon módszerek csoportja, amelyek az előző kategóriákba nem sorolhatók, nem kategorizálhatók ilyen egyértelműen, egyedi ötletek alapján működő algoritmusok.

A generátorokat tulajdonságuk és a használatban lévő próbák és vizsgálatok alapján rendeztük. A vizsgálatra használt tesztek:

1. Egyszerűbb vizsgálatok: A csoportban lévő egyszerűbb vizsgálatokat gyakran alkalmazzák más tesztek részeként.
2. D. E. Knuth által javasolt próbák: Empirikus és elméleti teszteket tartalmaz.
3. DIEHARD tesztcsoomag: A próbacsoomag 12 tesztje különböző statisztikai tulajdonságokat mérnek, amelyeket George Marsaglia gyűjtött össze.
4. NIST tesztcsoomag: Ez a próbacsoomag az amerikai NIST (magyarul: Nemzeti Szabványügyi Hivatal) javaslata, amely 15 tesztből áll. A csomag célja, hogy eszközként szolgáltasson véletlen adatsorok teszteléséhez.

A végeredmény egy áttekintő tanulmány, amelyben számszerűsítettük a tapasztaltakat.

MSC: 11Z05, 11K45, 65C10, 91A60

Informatikatudomány II.

2020. október 16. 13.00-15.00

Szekciófelelős:

Forczek Krisztofer, +36 30 388 0787

kulcsszavak: álvéletlenszám-generátorok, statisztikai tesztek, egyenletes eloszlású szekvenciák, álvéletlenszám-generátorok összehasonlítása

* A kutatást az „Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein” (EFOP-3.6.3-VEKOP-16-2017-00002) című projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Informatikatudomány II.

2020. október 16. 13.00-15.00

Szekciófelelős:

Forczek Krisztofer, +36 30 388 0787

Combinatorial method for scheduling batch processes with water integration

ŐSZ Olivér; HEGYHÁTI, Máté

Széchenyi István University

Optimization

osz.oliver@sze.hu

The chemical industry uses a lot of freshwater during manufacturing. Most of the water is used for cleaning the production equipment between processing different products. This is especially true for batch process plants, where material flow is not continuous, but transfer occurs at discrete time points. Batch plants offer more flexibility over their continuous counterparts, because it allows equipment units to be used to perform different tasks during the production process. While this flexibility is good for handling the parallel manufacturing of lots of different products with low order amounts, batch plants require more complex methods for scheduling.

As environmental awareness is rising and regulations are getting stricter in these days, chemical manufacturers are looking for optimization methods to decrease their wastewater generation. One such approach is water reuse, where the water output of a task can be used as input for another task. For each task, the required amount and the maximum allowed contamination of the water input is given. Similarly, the concentration of contaminants in the water outputs is given as well. The output concentration is considered to be a constant parameter, independent of the actual contamination of the input water. The water requirement of a task can be satisfied completely by reused water, or mixed with freshwater to lower contamination level or to increase total mass. Water can be transferred directly between equipment or stored in storage tanks before it is reused.

The S-graph framework was developed in 2002 for batch process scheduling. Since then, it has been extended to various scheduling problems that arise in industrial systems. The method uses a directed acyclic graph to model precedence relations between the events associated with production tasks. Such graphs are used to model partial schedules present at intermediate steps of a branch-and-bound search procedure. This algorithm is guaranteed to find the optimal schedule in finite steps. The S-graph approach was originally used for makespan minimization, but it was also extended to cost minimization in the past. In the current work, a new extension to the framework is presented, for integrating water reuse assignments into the search algorithm, in order to minimize wastewater generation through scheduling. The correctness and performance of the proposed approach has been tested on literature examples.

Informatikatudomány II.

2020. október 16. 13.00-15.00

Szekciófelelős:

Forczek Krisztofer, +36 30 388 0787

TNMplot: egy webes alkalmazás a normál, tumoros és metasztatikus adatok transzkriptom szintű összehasonlítására

BARTH Áron^{1,2,3}, Balázs GYÓRFFY^{1,2,3}

¹*Semmelweis Egyetem, Bioinformatika Tanszék, Budapest*

²*TTK Lendület Onkológiai Biomarker Kutatócsoport, Enzimológia Intézet, Budapest*

³*Semmelweis Egyetem, II. sz. Gyermekgyógyászati Klinika, Budapest*

Kulcsszavak: tumor, normál, metasztázis, génexpresszió,

Bevezetés: Az elmúlt két évtizedben hatalmas mennyiségű, online is elérhető adatvált elérhetővé a rosszindulatú tumorok kutatásához. Ezek részben RNS szekvenálási adatokat (TCGA, TARGET, GTEx), részben gén chip alapú vizsgálatokat (NCBI-GEO) tartalmaznak. Jelenleg nincs olyan felhasználóbarát, könnyen elérhető platform, amely lehetővé teszi ezen adatbázisokon belül és ezek között a normál, tumoros és áttétes génexpressziós adatok nagy mintaszámon alapuló összehasonlítását.

Célkitűzés: Egy olyan webes felület létrehozása, amelynek felhasználásával összemérhetőek a normál, tumoros és áttétes génexpressziós adatok.

Módszerek: Munkánk során két génexpresszió mérésére alkalmas eszköz, a génchip, illetve az RNS szekvenálás adataira támaszkodtunk. A génchip adatokat az NCBI-GEO adatbázisban elérhető adatok alapján dolgoztuk fel, összesen 3180 vizsgálatot felhasználva, melyekből manuálisan választottuk ki a megfelelőnek ítélt mintákat, amelyeket a MAS5 algoritmus segítségével normalizáltunk. Az RNS szekvenálási adatokat a TCGA, a TARGET valamint a GTEx adatbázisaiból töltöttük le. A három adatbázisból összesen 23 418 minta szekvenálási adatát tudtuk felhasználni, amelyeket a DESeq2 algoritmus segítségével normalizáltunk. Az adatok kiértékeléséhez Mann-Whitney U vagy Kruskal-Wallis tesztet alkalmaztunk.

Eredmények: A 3180 génchip alapú vizsgálatból összesen 33 520 minta került bevonásra, melyből 453 áttétes, 29 376 tumoros és 3691 normál típusú, összesen 38 féle szövet típusból. A TCGA adataiból 11010 mintát használtunk fel (394 áttétes, 9886 tumoros, 730 normál) amely 33 szövet típust reprezentál, a TARGET –ből 1193 minta került felhasználásra (1 áttétes, 1180 tumoros, 12 normál) 5 féle szövet típusból, míg a GTEx-ből 11 215 normál mintát sikerült kinyerni melyek 51 féle szövet típusba vannak sorolva.

Összegzés: Munkánk során létrehoztunk egy adatbázist, mely összesen 56 938 minta transzkriptom-szintű adatait tartalmazza. A létrehozott web platform lehetővé teszi az adatbázis korlátlan felhasználását, amellyel elérhetővé válik a normál, tumoros és metasztatikus szövetek génexpresszió szintű összehasonlítása. Az oldal elérhető a <https://tnmplot.com/> felületen.