TEHNICI AVANSATE *DATA MINING* UTILIZATE ÎN SISTEMELE INFORMATICE INTELIGENTE PENTRU ASISTAREA DECIZIILOR

ADVANCED DATA MINING TECHNIQUES USED IN INTELLIGENT DECISION SUPPORT SYSTEMS

Conf.univ.dr. Elena ŞUŞNEA*

În prezent, datele sunt omniprezente, iar volumul de date noi, care sunt generate şi stocate în fiecare zi, continuă să crească exponenţial. Organizaţia militară, ca multe alte organizaţii, colectează şi stochează volume mari de date. Colectarea, corelarea şi interpretarea *big data* într-o concepţie coerentă formează imaginea operaţională comună (*common operating picture – COP*). O astfel de imagine oferă comandanţilor soluţii, în orice moment, fiind esenţială pentru luarea deciziilor în timp util. Descoperirea informaţiilor valoroase, ascunse în seturile de date, este dificil de realizat. De aceea, transformarea acestor date în cunoştinţe prin utilizarea tehnicilor *data mining*, în special a tehnicilor specifice inteligenţei artificiale, poate fi o provocare. Utilizând aceste tehnici, putem extrage cunoştinţe care să constituie soluţii optime pentru problemele identificate.

At present, data are ubiquous and the amount of new data that is generated and stored every day continues to increase exponentially. The military organization, as many other organizations, collects and stores huge amounts of data. The coherent collection, correlation anf interpretation of big data shapes the common operating picture (COP). Such a picture provides commanding officers with solutions at any time, and it is essential for timely decision-making. Thus, transforming these data into knowledge using data mining techniques, especially artificial intelligence techniques may prove a challenge. Using these techniques, we can extract knowledge which can become optimal solutions for the identified problems.

Cuvinte-cheie: data mining; inteligenţă artificială; *big data*; sisteme informatice pentru asistarea deciziilor; imaginea operaţională comună.

***Keywords**: data mining; artificial intelligence; big data; decision support systems; common operating picture.*

Trăim într-o lume în care datele sunt colectate în cantităţi tot mai mari, acestea exprimând, într-o varietate de formate, comportamentul oamenilor şi al maşinilor şi captând rapid diverse niveluri de agregare. Pentru a evidenţia volumul imens de date generate, eterogenitatea acestora, viteza mare de colectare şi veridicitatea seturilor de date obţinute, se foloseşte frecvent termenul big data. Cele patru dimensiuni ale big data – volum, viteză, varietate şi veridicitate – definesc modelul 4V[1]. Acest model este susţinut de „tehnologiile *big data* care descriu o nouă generaţie de tehnologii şi arhitecturi, concepute pentru a extrage plusvaloare din volume imense de date, disponibile într-o mare varietate de formate, permiţând captarea, descoperirea şi/sau analiza la viteză foarte mare”[2].

Apariţia *big data* este consecinţa scăderii drastice a costurilor de stocare pentru 1 GB, de la un milion de dolari, în 1967, la 0,2 dolari, în prezent[3], şi a creşterii capacităţii de stocare. Tranziţia de la generarea datelor cu ajutorul creionului şi hârtiei la generarea datelor cu ajutorul computerului a constituit un pas important către *big data*. În 1990 a fost lansat primul proiect de digitizare a colecţiilor deţinute de Biblioteca Congresului Statelor Unite, care a avut ca obiectiv transformarea unui număr de 160 de milioane de obiecte în format digital[4]. Volumul de date obţinut a fost estimat la 235 TB[5], ceea ce înseamnă aproximativ 1,59 miliarde de pagini Web, în timp ce, în noiembrie 2017, doar motorul de căutare Google indexa în jur de 50 de miliarde de pagini Web[6]. Paginile indexate

UniversitateaNaţională de Apărare „Carol I”
e-mail*: esusnea@yahoo.com*

conțin date referitoare la tranzacții financiare, mesaje, fotografii etc. Acest exemplu evidențiază dimensiunile 4V ale *big data*.

Utilizarea tehnicilor *data mining* în *big data* sprijină extragerea cunoștințelor necesare procesului decizional. În primul rând, un volum mai mare de date permite o viziune mai cuprinzătoare asupra trecutului, prezentului și viitorului, permițând elaborarea unor descrieri sau predicții. În al doilea rând, viteza mare de generare a datelor poate fi utilă factorilor de decizie, deoarece datele sunt actualizate în timp real. În al treilea rând, varietatea formatelor sub care sunt prezentate datele poate contribui la o analiză mai nuanțată a problemei. Nu în ultimul rând, pe măsură ce veridicitatea datelor se îmbunătățește, crește și încrederea factorului de decizie că informațiile obținute sunt exacte, autentice și coerente.

Natura complexă și dinamică a actualului mediu de securitate impune, adesea, luarea deciziilor în timp real. În aceste medii decizionale complexe, care sunt caracterizate de *big data*, este important ca factorii de decizie să ia decizii proactive. Comparativ cu bazele de date tradiționale, big data oferă noi oportunități pentru descoperirea cunoștințelor ascunse în seturile de date. Un instrument util, în acest sens, este sistemul informatic pentru asistarea deciziilor. Un astfel de sistem utilizează tehnologii care analizează rapid cantitățile mari de date de diferite tipuri (de exemplu, date structurate din baze de date relaționale și date nestructurate, cum ar fi imagini, videoclipuri, e-mailuri, date despre interacțiuni sociale) dintr-o varietate de surse pentru a produce un flux de informații acționabile.

Voi prezenta, în continuare, tipurile de aplicații *data mining* care contribuie la dezvoltarea unor sisteme inteligente pentru asistarea deciziilor. Mai întâi, voi sublinia interesul manifestat în domeniul militar pentru *big data*, apoi voi evidenția rolul tehnicilor *data mining* în analiza datelor care sprijină procesul decizional, iar la final voi reliefa câteva aspecte ale rețelelor neuronale artificiale.

Revoluția *big data* în domeniul militar și sistemele informatice inteligente pentru asistarea deciziilor

În ultimii ani, tot mai multe persoane din diverse domenii de activitate – precum informaticienii, fizicienii, economiștii, matematicienii, oamenii politici, sociologii etc. – au devenit din ce în ce mai interesate de avantajele pe care le oferă *big data*.

Pentru a spori capacitățile de apărare, domeniul militar acordă, de asemenea, o atenție deosebită acestei tendințe[7]. De exemplu, Departamentul Apărării al Statelor Unite investește anual 250 de milioane de dolari în programe de tip *big data*, cum ar fi XDATA, CyberInsider Threat (CINDER), Anomaly Detection at Multiple Scales (ADAMS), Insight, Mind's Eye, Machine Reading, Mission-oriented Resilient Clouds, Programming Computation on Encrypted Data (PROCEED) și Video and Image Retrieval and Analysis Tool (VIRAT)[8]. În general, specialiștii din domeniul militar colectează cantități masive de date, prin senzori inteligenți, prin supraveghere și prin recunoaștere (ISR)[9], de la diverse entități implicate în luptă și despre diverse evenimente care au loc pe câmpul de luptă.

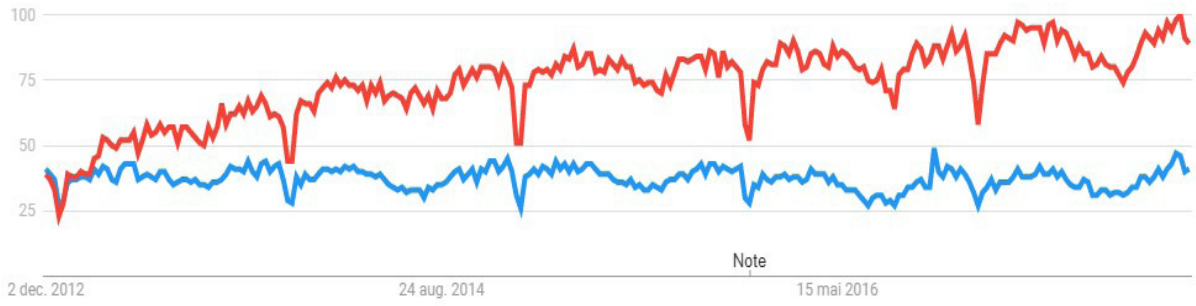

Fig. 1 *Nivelul de interes pentru subiectele big data și data mining (în ultimii cinci ani)*[10]

După colectarea și stocarea datelor, cea mai mare provocare nu este doar managementul acestora, ci și analiza și extragerea informațiilor cu valoare semnificativă pentru procesul decizional.

Reușita unei misiuni implică existența unui sistem informatic inteligent care să asiste decidenții militari în recunoașterea factorilor-cheie ai spațiului de luptă, în evaluarea variantelor decizionale și în

alegerea cursului de acțiune în cel mai scurt timp. Arhitectura acestor sisteme informatice are la bază trei componente principale: o bază de date sau un *data warehouse*, o bază de modele și instrumente analitice și interfața cu utilizatorul.

Cu două decenii în urmă, în multe domenii de activitate se colectau și se raportau rezumate ale datelor folosind tehnici din statistică și interogări ale bazelor de date. Recent însă, a avut loc o schimbare de paradigmă. Creșterea volumului și a detaliilor informațiilor captate, creșterea volumului de date multimedia, expansiunea rețelelor de socializare și apariția Internetului obiectelor (Internet of things, IoT) au necesitat noi tehnologii pentru stocare și analiză. Prin urmare, începând cu anul 2012, interesul pentru cele două domenii științifice, *big data* și *data mining,* a fost într-o continuă creștere (figura 1).

Datele sunt omniprezente, iar volumul de date noi, care sunt generate și stocate în fiecare zi, continuă să crească exponențial. La nivel mondial, există câteva organizații specializate în culegerea datelor, de exemplu NASA și CIA, unde volumul de date colectat zilnic este de aproximativ 1 TB.

În domeniul militar, ca multe alte domenii de activitate, se generează, se colectează și se stochează volume mari de date. Prin digitalizarea câmpului de luptă se generează o mare cantitate de date specifice, caracterizate de cele patru dimensiuni ale *big data*. Mediul operațional în care se desfășoară în prezent conflictele, chiar și cele care au loc în zonele cele mai izolate, nu a generat niciodată date atât de multe și atât de complexe. Sarcina de a analiza aceste date pentru a identifica informațiile acționabile sau cunoștințele necesare procesului decizional devine tot mai dificilă, deoarece majoritatea sistemelor informatice clasice nu reușesc să depășească problema analizei *big data*. De aceea, potențialul militar al datelor trebuie exploatat în mod eficient cu tehnici avansate *data mining*.

Sistemele informatice au fost studiate intens în domeniul militar[11], deoarece era deosebit de important să se înțeleagă punctele forte și slăbiciunile acestor sisteme, pentru a îmbunătăți procesul de luare a deciziilor. Un sistem informatic inteligent pentru asistarea deciziei trebuie să fie un „sistem antropocentric și evolutiv care are rolul de a implementa funcțiile unui sistem de asistare umană care ar fi necesar pentru a ajuta factorul de decizie să depășească limitele și constrângerile pe care le poate întâlni atunci când încearcă să rezolve o problemă decizională”[12].

Tranziția de la sisteme informatice clasice pentru asistarea deciziilor la sisteme inteligente s-a realizat având în vedere că „este necesar să se dezvolte sisteme informatice care să fie nu doar precise și ușor de utilizat, dar, de asemenea, să stimuleze utilizatorii să dobândească noi competențe, să adopte noi stiluri de lucru și să-și dezvolte talentul și creativitatea”[13]. Generarea și accesarea volumelor mari de date, prezentate în diverse formate – text, imagini, clipuri video și semnale generate de o multitudine de platforme –, este dificil de exploatat, mai ales în timp real, din cauza volumului și a lipsei de omogenitate. Acest lucru a făcut incompatibilă analiza datelor cu metodele tradiționale specifice bazelor de date și statisticii.

Progresele uriașe înregistrate de tehnologia hardware, cum ar fi dezvoltarea de senzori miniaturizați, dispozitive GPS (Global Positioning System), pedometre și accelerometre, și de aplicațiile software, precum platformele social media, miniblogging etc., care pot fi utilizate pentru a genera și partaja diferite tipuri de date, au generat cantități extraordinare de date în timp real. Scăderea costurilor acestor dispozitive și aplicații software în mod constant, în ultimii ani, și creșterea eficienței tehnologiilor de colectare a datelor au influențat procesul de colectare a datelor de către sistemele informatice inteligente pentru asistarea deciziilor.

Colectarea, corelarea și interpretarea *big data* într-o concepție coerentă formează imaginea operațională comună (common operating picture, COP). O astfel de imagine oferă comandanților soluții în orice moment, fiind esențială pentru luarea deciziilor în timp util. Descoperirea informațiilor valoroase ascunse în *big data* este dificil de realizat. De aceea, transformarea acestor date în cunoștințe prin utilizarea tehnicilor *data mining*, în special a tehnicilor specifice inteligenței artificiale, constituie o provocare.

Tipuri de aplicaţii *data mining*

Inovaţiile din domeniul tehnologiei informaţiilor au făcut posibilă achiziţionarea şi stocarea în baze de date a unor cantităţi mari de date. Multe domenii de activitate, printre care şi domeniul militar, devin din ce în ce mai dependente de colectarea,

stocarea şi procesarea datelor. Totuşi, abundenţa datelor colectate face dificilă găsirea unor informaţii esenţiale care să corespundă unui anumit scop. În anii ’90, a apărut un nou domeniu de cercetare care a sprijinit analiza informaţiilor extrase din datele existente în bazele de date, denumit *data mining*.

Data mining este procesul care constă în descoperirea unor modele, corelaţii sau tendinţe nebănuite, de o importanţă deosebită, prin analiza datelor stocate în baze de date, utilizând tehnologii de recunoaştere a formelor, precum şi tehnici din statistică şi din inteligenţa artificială. În încercarea de standardizare a procesului *data mining*, s-a remarcat, în mod deosebit, modelul CRISP-DM (CRoss-Industry Standard Process for Data Mining), dezvoltat de un mare consorţiu de companii europene, Integral Solution Ltd., NCR, DaimlerChrysler, OHARA. Modelul a fost sprijinit prin programul ESPRINT, iniţiat de Comisia Europeană, şi constă în parcurgerea a şase etape: înţelegerea aplicaţiei, înţelegerea datelor, pregătirea datelor, modelare, evaluare şi implementare (figura 2).

În funcţie de obiectivele pentru care sunt analizate datele, avem următoarele trei tipuri de aplicaţii *data mining*: analiză descriptivă, analiză exploratorie a datelor, analiză predictivă.

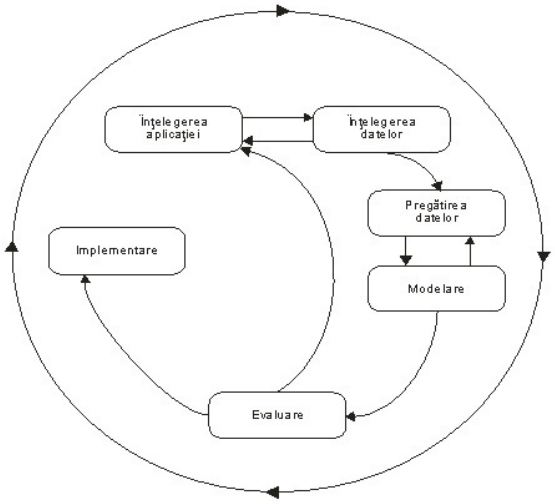

Fig. 2 *Etapele modelului CRISP–DM*[14]

Modelarea descriptivă are ca scop descrierea tuturor datelor existente în set. Astfel de descrieri includ modele pentru distribuţia de probabilitate a datelor (estimarea densităţii), partiţionarea spaţiului p-dimensional în grupuri (analiza şi segmentarea grupurilor, cunoscută sub denumirea de *clustering*) şi modele care descriu relaţia dintre variabile (modelarea dependenţelor). În analiza segmentării, de exemplu, scopul este de a grupa înregistrări similare. Împărţirea înregistrărilor în grupuri omogene se face astfel încât obiectele cu similaritate mai mare să aparţină aceluiaşi grup. Numărul de grupuri este ales de către utilizator; nu există noţiunea de număr corect. Aceasta contrastează cu analiza de grup (*cluster*), în care scopul este de a descoperi grupuri naturale din date. Modelarea descriptivă a fost folosită în *clustering* (recunoaşterea formelor prin împărţirea unei imagini digitale în regiuni distincte, cu scopul detectării marginilor acestora, sau recunoaşterea obiectelor) şi în segmentare.

Analiza exploratorie a datelor, după cum sugerează şi numele, are ca obiectiv explorarea datelor, prin izolarea caracteristicilor relevante, prin identificarea potenţialelor structuri ale caracteristicilor şi prin generarea de ipoteze plauzibile pentru a explica structura. Procedurile analizei exploratorii sunt completate cu proceduri de modelare şi de testare a ipotezelor, de direcţionare sau revizuire a analizei ca răspuns la informaţiile neaşteptate obţinute din date. În general, tehnicile utilizate sunt interactive şi vizuale şi se bazează, în special, pentru seturile de date de dimensiuni mici, pe histograme ale variabilelor continue sau discrete, diagrame boxplot, partiţii de date etc. Noţiuni, precum dimensiune şi detaliu, devin foarte importante, în contextul în care datele cu rezoluţie mică pot fi prezentate cu riscul de a nu observa unele detalii importante.

Modelarea predictivă a captat atenţia în mod special, deoarece aceasta încearcă să îmbunătăţească procesul decizional prin luarea unor decizii raţionale, bazate pe date. Analizele predictive realizate cu tehnici din inteligenţă artificială (precum reţelele neuronale artificiale, algoritmii genetici) transformă modul în care se iau deciziile.

Prin urmare nu este surprinzător faptul că *data mining* şi, în general, paradigmele bazate pe date au fost aplicate cu succes într-o varietate de aplicaţii militare care permit extragerea cunoştinţelor. În acest sens, se disting două tipuri de cunoştinţe: cunoaştinţe globale şi cunoştinţe locale. Cunoştinţele globale sunt necesare pentru a determina direcţiile de acţiune pe care să îşi concentreze atenţia decidenţii militari, în timp

ce cunoștințele locale sunt utile pentru evaluarea validității unei anumite alternative bazate pe un anumit set de constatări. Cunoștințele globale sunt reprezentate ca o rețea de indicatori relevanți între alternative și caracteristici.

Ponderile prin care se evaluează puterea relevanței și se determină direcțiile globale (ipotezele) pentru analiza situației sunt atașate la legăturile acestei rețele. Pentru cunoștințele locale, se pare că, în majoritatea problemelor practice, ar fi necesare mai multe tehnici de reprezentare pentru a caracteriza adecvat alternativele prin caracteristicile lor relevante.

Tehnici inteligente din domeniul *data mining*: rețele neuronale artificiale

Rețelele neuronale artificiale s-au folosit din ce în ce mai mult, în ultimii ani, pentru rezolvarea problemelor complexe. Acestea reprezintă o variantă atrăgătoare pentru analiza datelor din domeniul militar, deoarece sunt capabile să construiască un model care „emulează comportamentul unui comandant de armată pe câmpul de luptă atunci când întâlnește diferite situații, folosind două configurații diferite de rețele neuronale – perceptronul multistrat (MLP) și rețeaua probabilistică neuronală (PNN)”[15]. Modelul este deosebit de util pentru comandanți, deoarece sprijină luarea deciziilor pe câmpul de luptă, dezvoltarea de noi strategii și managementul resurselor.

Rețelele neuronale artificiale au la bază sisteme inteligente inspirate din rețelele neuronale biologice. Acestea sunt capabile să învețe din exemple și să generalizeze cunoștințele achiziționate. Deși lungul curs al evoluției i-a dat creierului uman multe caracteristici, o serie dintre acestea nu se regăsesc încă în modelele actuale de rețele neuronale artificiale. Calculatoarele actuale surclasează oamenii din punctul de vedere al calculului numeric și al manipulării simbolurilor asociate. Totuși omul poate rezolva, fără mare dificultate, probleme complexe de percepție, de genul recunoașterii unei persoane într-un loc aglomerat după trăsăturile feței acesteia. Explicația este dată de faptul că arhitectura sistemului nervos biologic este complet diferită de arhitectura artificială Von Neumann, ceea ce afectează tipul de probleme pe care fiecare dintre modele îl poate rezolva.

În 1943, McCulloch & Pitts au propus un model de neuron artificial care avea la bază o unitate cu prag binar. Acest model matematic calculează suma ponderată a celor *n* intrări și generează un răspuns binar *y*. Răspunsul este *1,* dacă suma depășește un anumit prag *u*, în caz contrar fiind egală cu *0*:

$$y = \theta\left(\sum_{j=1}^{n} w_j x_j - u\right)$$

unde θ *( )* reprezintă funcția treaptă unitate în *0*, iar ω_j sunt ponderile sinapselor asociate valorii intrării x_j. Pentru simplificarea notării, se consideră pragul *u* ca o altă pondere $w_0 = -u$ asociată intrării constante $x_0 = 1$. Ponderile pozitive corespund sinapselor excitatoare, în timp ce ponderile negative sunt inhibatoare. Modelul McCulloch-Pitts a fost generalizat în diferite moduri.

Concluzii

În această lucrare am prezentat tehnicile avansate de analiză a datelor care pot fi folosite în cadrul sistemelor inteligente pentru asistarea deciziilor. *Data mining* a primit o atenție considerabilă în domeniul militar, la nivel internațional, fiind lansate diferite programe militare care au ca obiectiv exploatarea datelor, cu scopul extragerii cunoștințelor necesare procesului de luare a deciziilor

Într-o lume dominată de conexiuni și de mobilitate, cantitățile de date generate sunt uriașe. În domeniul militar, asemenea multor altor domenii de activitate, sursele generatoare de date variază de la dispozitive inteligente și rețele de socializare la imagini digitale, la sisteme de geolocație și multe altele. Caracteristicile acestor date definesc *big data* și sunt reunite în modelul 4V (volum, viteză, varietate și veridicitate).

Datele sunt materii prime foarte valoroase pentru organizația militară. Cu toate acestea, volumul foarte mare, eterogenitatea și viteza de generare pot face procesul de luare a deciziilor extrem de complex. Dezvoltarea unor instrumente inteligente pentru asistarea deciziilor, prin includerea unor tehnici de analiză *data mining*, în particular a rețelelor neuronale artificiale, ar permite factorilor de decizie să utilizeze în mod eficient toate aceste date în timp real. Timpul real este cea mai mare provocare pentru crearea imaginii operaționale comune. Pentru a răspunde acestei provocări, sunt dezvoltate noi tehnici *data mining* și

algoritmi tot mai puternici de învățare supervizată și nesupervizată, care permit o putere mai mare de procesare și o analiză în timp a datelor.

NOTE:

1 http://www.ibmbigdatahub.com

2 J. Gantz, D. Reinsel, *Extracting Value from Chaos*, IDC iview, 2011, pp. 9-10.

3 L. Mearian, *CW@50: Data storage goes from $1M to 2 cents per gigabyte*, Computerworld, 2017.

4 https://memory.loc.gov/ammem/about/index.html

5 J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh & A.H.Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, 2011.

6 http://www.worldwidewebsize.com

7 I.J. Donaldson, S. Hom, T. Housel, *Visualization of big data through ship maintenance metrics analysis for fleet maintenance and revitalization*, Naval Postgraduate School, Monterey, California, USA, 2014.

8 R. King, *U.S. government spending on big data to grow exponentially*, 2013, http://www.biometricupdate.com/201308/u-s-government-spending-on-bigdata-to-grow-exponentially

9 O. Savas, Y. Sagduyu, J. Deng, J. Li, *Tactical big data analytics: challenges, use cases, and solutions*, ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 4, 2014, pp. 86-89.

10 https://trends.google.com

11 M. Ben-Bassat, *Knowledge requirements and management in expert decision support systems for (military) situation assessment (Technical report)*, U.S. Army Research Institute for the Behavioral and Social Sciences, 1983.

12 F. Filip, B.C. Zamfirescu, C. Ciurea, *Computer-Supported Collaborative Decision-Making*, Springer International Publishing, 2017.

13 F. Filip, *Towards more humanized real-time decision support systems*. In: Balanced Automation Systems; Architectures and Design Methods (L. M. Camarinha – Matos and H. Afsarmanesh, eds.). Chapman & Hall, London, 1995, pp. 230- 240.

14 http://www.crisp-dm.org

15 G.S. Gill, J.S. Sohal, *Battlefield Decision Making: A Neural Network Approach*, Journal of Theoretical and Applied Information Technology, vol. 4, no. 8, pp. 697-699, 2009.

BIBLIOGRAFIE

Ben-Bassat M., *Knowledge requirements and management in expert decision support systems for (military) situation assessment* (Technical report), U.S. Army Research Institute for the Behavioral and Social Sciences, 1983.

Donaldson I.J., Hom S., Housel T., *Visualization of big data through ship maintenance metrics analysis for fleet maintenance and revitalization*, Naval Postgraduate School, Monterey, California, USA, 2014.

Filip F., *Towards more humanized real-time decision support systems*. In: Balanced Automation Systems; Architectures and Design Methods (L. M. Camarinha – Matos and H. Afsarmanesh, eds.). Chapman & Hall, London, 1995.

Filip F., Zamfirescu B.C., Ciurea C., *Computer-Supported Collaborative Decision-Making*, Springer International Publishing, 2017.

Gantz J., Reinsel D., *Extracting Value from Chaos*, IDC iview, 2011.

Gill G.S., Sohal J.S., *Battlefield Decision Making: A Neural Network Approach*, Journal of Theoretical and Applied Information Technology, vol. 4, no. 8, 2009.

King R., *U.S. government spending on big data to grow exponentially*, 2013, http://www.biometricupdate.com/201308/u-s-government-spending-on-bigdata-to-grow-exponentially

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, 2011.

Mearian L., *CW@50: Data storage goes from $1M to 2 cents per gigabyte*, Computerworld, 2017.

Savas O., Sagduyu Y., Deng J., Li J., *Tactical big data analytics: challenges, use cases, and solutions*, ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 4, 2014.

http://www.ibmbigdatahub.com

http://www.crisp-dm.org

http://www.worldwidewebsize.com

https://trends.google.com

https://memory.loc.gov/ammem/about/index.html