Informácia o riešení témy Big Data v projekte „Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológii a systémov“

Viera Rozinajová[1], Gabriela Kosková[1], Anna Bou Ezzeddine[1], Mária Lucká[1] , Mária Bieliková[1], Pavol Návrat[1]

[1] Fakulta informatiky a informačných technológií STU Bratislava,
Ilkovičova 2, 842 16 Bratislava
{rozinajova, koskova, ezzeddine, lucka,
bielikova, navrat}@fiit.stuba.sk

Abstract. Príspevok poskytuje informáciu o projekte s názvom Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov po jeho úvodnej faze riešenia. V závere identifikujeme hlavné úlohy, ktoré sa v projekte v ďalších fázach budú riešiť.

Slovenská technická univerzita (STU) v Bratislave je jedným z partnerov projektu s názvom *Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov*, ktorý je podporovaný agentúrou Agentúra Ministerstva školstva, vedy, výskumu a športu SR pre štrukturálne fondy EÚ. Cieľom projektu je vytvorenie medzinárodného centra excelentnosti a výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov. Projekt začal vo februári 2014 a čas jeho riešenia je naplánovaný do septembra 2015.

Základný výskum pre rozvoj inteligentných a bezpečných informačno-komunikačných systémov sa sústreďuje na výskumné témy:

- Inteligentná sieť – Smart grid,
- Bezpečnosť – Kryptografia,
- Veľké dáta – Big Data.

V rámci STU je do projektu zapojených viacero pracovísk: Ústav elektroenergetiky a aplikovanej elektrotechniky na Fakulte elektrotechniky a informatiky (FEI) STU v téme *Inteligentná sieť – Smart grid*, Ústav informatiky a matematiky FEI STU v téme *Bezpečnosť – Kryptografia* a Ústav informatiky a softvérového inžinierstva (ÚISI) na Fakulte informatiky a informačných technológií (FIIT) STU v téme *Veľké dáta – Big Data*. Tri riešené témy prepája oblasť energetiky, ktorá je vlastná téme *Inteligentná sieť – Smart grid*. Téma *Bezpečnosť – Kryptografia* sa venuje bezpečnosti prenosu dát v inteligentnej sieti. V rámci témy *Veľké dáta*, ktorá je zameraná na spracovanie, analýzu a vizualizáciu veľkých objemov dát sa venujeme analýze meraní z inteligentných meračov spotreby elektrickej energie.

Táto informácia sa týka výskumu na ÚISI FIIT STU a témy *Veľkých dát – Big Data*. Potrebu spracovávania veľkých dát v oblasti energetiky vyvolala snaha Európskej únie o zavedenie inteligentnej siete v rámci Európskej únie. Podľa vyhlášky MH SR č.358/2013 účinnej od 15. novembra 2013 má byť v Slovenskej republike inteligentnými meračmi spotreby elektrickej energie vybavených 80 % zo všetkých odberných miest, ktorých je rádovo dva milióny. Inteligentné merače zaznamenávajú informácie o odbere elektrickej energie v 15-minútových intervaloch a následne ich odosielajú do centrálneho informačného systému. Tieto merače umožňujú jednak každému spotrebiteľovi regulovať spotrebu energie (príkladom systému na meranie a vyhodnocovanie spotreby energie je systém Energomonitor (https://www.energomonitor.cz/)), jednak poskytujú podporu pre návrh nových spôsobov predikovania spotreby. Pri cieľovej konfigurácii prenášané a spracovávané dáta spĺňajú charakteristiky veľkých dát, ako je objem a rýchlosť pribúdania dát. Ďalšou dôležitou charakteristikou týchto dát je, že odbery pre jednotlivé odberné miesta, či distribučné skupiny je možné modelovať pomocou časových radov.

Výzvou témy veľkých dát vo všeobecnosti je spracovanie, analýza a vizualizácia veľkých objemov dát v reálnom čase s cieľom podporiť ďalšie rozhodovanie. V oblasti energetiky vzhľadom na charakter dát sú typickými úlohami:

- predikcia spotreby elektrickej energie pre distribučné spoločnosti, bilančné či iné skupiny alebo jednotlivcov,
- analýza vzťahu spotreby elektrickej energie a externých faktorov, napr. meteorologických či demografických údajov,
- dolovanie typických motívov a tvarov kriviek odberu elektrickej energie,
- klastrovanie odberných miest podľa charakteristík odberu,
- klasifikácia odberateľov elektrickej energie,
- detekcia anomálií.

V úvodnej fáze projektu sme analyzovali prístupy modelovania, spracovávania a predikovania časových radov klasickými prístupmi, ako sú regresia a analýza časových radov [6, 7, 8]. Modely jednoduchej a viacnásobnej regresnej analýzy môžeme zaradiť medzi kauzálne prognostické modely. V týchto modeloch uplatňujeme predpoklad, že medzi vstupmi a výstupmi systému existuje vzťah príčiny a dôsledku. Prognózovaná hodnota premennej je stanovená na základe identifikovaného kauzálneho vzťahu. Pomocou časového radu zaznamenávame zmenu skúmanej premennej v čase. Vychádzame z rozdelenia časového radu na štyri zložky: trendová, sezónna, cyklická a náhodná. V praktických úlohách často oddeľujeme a vyčíslujeme trendovú a sezónnu zložku a z takto upraveného modelu zisťujeme prognózu. Ak údaje časových radov vykazujú trend, používame nasledujúce metódy: Holtovo exponenciálne vyrovnanie, metódu kĺzavých priemerov, jednoduché regresné modely, modely trendu, rastové medely a autoregresné integratívne modely kĺzavých priemerov – ARIMA modely. V prípade identifikovania sezónnych znakov v modeli na stanovenie prognózy používame Wintersovo exponenciálne vyrovnanie, viacnásobnú regresiu, klasickú dekompozíciu, alebo ARIMA modely. Z pohľadu rizikového manažmentu má veľký význam identifikácia a kvantifikácia sezónnej zložky, čím zisťujeme odchýlku oproti očakávanej hodnote veličiny. Na prognózovanie budúcich hodnôt stacionárnych

časových radov používame naivné modely, kĺzavé priemery, jednoduché exponenciálne vyrovnanie a autoregresívne modely.

Po analýze klasických prístupov sme skúmali adaptívne prístupy založené na neurónových sieťach a metódach podporných vektorov [5, 6]. Zamerali sme sa na publikované práce z oblasti predikcie odberov elektrickej energie [1, 8]. Keďže kľúčovým problémom je výber vhodného spôsobu reprezentácie rýchlo pribúdajúcich údajov, analyzovali sme spôsoby reprezentácie časových radov z pohľadu redukcie dát a tiež potreby neustáleho rozširovania dátových súborov o nové namerané hodnoty [3]. Na podporu spracovávania dát v reálnom čase sme analyzovali prístupy paralelného a distribuovaného spracovávania veľkých dát . Práca s veľkými dátami vyžaduje aplikáciu netradičných spôsobov spracovania, vzhľadom na ich objem, rýchlosť zmien, distribuovanosť a špecifickú štruktúru. Skúmali sme možnosti použitia analytických nástrojov na spracovanie veľkých dát, akými sú MapReduce a Hadoop [2, 4]. Naštudovali sme základné vlastnosti a charakteristiky týchto softvérových nástrojov s cieľom posúdiť vhodnosť ich použitia v inteligentných sieťach. HDFS (Hadoop Distribuovaný File Systém) bol vytvorený na spoľahlivé uchovávanie dát a ich škálovanie, čím vytvára predpoklad rýchleho spracovania spolu so spoľahlivosťou.

Priebehové merania sa na Slovensku vykonávajú od 1.júla 2013. K dispozícii sme mali prvú vzorku dát, ktorých vlastnosti sme skúmali v prostredí pre štatistické výpočty R.

Vzhľadom na časovú náročnosť nie sú doposiaľ používané prístupy na modelovanie, analýzu či predikovanie odberov elektrickej energie vhodné na spracovávanie veľkých objemov dát, ktoré v tomto projekte riešime. Na analýzu a predikciu takýchto dát v reálnom čase plánujeme v ďalších fázach projektu riešiť:

- reprezentáciu dát pre rýchly prístup a spracovanie,
- predspracovanie a transformácia časových radov s cieľom získať vybrané charakteristiky,
- štatistické metódy analýzy dát,
- rýchle spracovanie dát s podporou paralelných a distribuovaných prístupov, resp. použitie inkrementálneho spracovania
- vizualizáciu dát podľa nájdených charakteristík.

English summary

The paper provides short information about recently started project ‘’International Centre of Excellence for Research of Intelligent and Secure Information-Communication Technologies and Systems.” Three main topics of the project include Smart grid, Cryptography and Big Data. Main challenge in the theme Big Data is to find models and methods suitable for real time processing, analysis and visualization of big data volumes coming from Smart Metering. Application of a new European Union Regulation concerning metering of electricity consumption to Slovakia requires to equip about 80% of electricity supply points with smart meters in the next two years. Smart meters will produce huge amount of data that must be properly stored and processed. We have studied classical methods suitable for analysis and

prediction of time series taking into account seasonal variations. We have analyzed suitability of these methods for real time processing of big data volumes arising from smart metering and we have investigated new models and methods. They include research in the area of adaptive methods based on neural networks and support vector machines and methods allowing incremental processing. We plan to use parallel and distributed software frameworks such as MapReduce with Hadoop [2,4] and to find proper representation of data that will enable online processing of data and visualization according to selected characteristics.

Poďakovanie: *Táto publikácia vznikla vďaka podpore projektu v rámci OP Výskum a vývoj pre projekt: „Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov“, ITMS: 2624012003, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.*

Literatúra

1. Aung, Z., Toukhy, M., Williams, J., Sanchez, A., Herrero, S.: Towards accurate electricity load forecasting in smart grids. In Proceedings of the 4th International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA) (2012) 51–57.
2. Dean J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. In Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, (2004).
3. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. In Proceedings of the VLDB Endowment 1 (2), (2008)1542–1552.
4. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Rcfile, XuZ.: A fast and space-efficient data placement structure in mapreduce-based warehouse systems. In ICDE, (2011) 1199–1208.
5. Hu, Z., Bao, Y., Xiong, T.: Electricity load forecasting using support vector regression with memetic algorithms. ScientificWorldJournal 2013:292575. doi: 10.1155/2013/292575. eCollection (2013).
6. Liu, N., Babushkin, V., Afshari, A.: Short-Term Forecasting of Temperature Driven Electricity Load Using Time Series and Neural Network Model. Journal of Clean Energy Technologies, Vol. 2, No. 4 (2014).
7. Moroshko, E., Crammer, K.: A Last-Step Regression Algorithm for Non-Stationary Online Learning. Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), Scottsdale, AZ, USA, (2013) 451–462.
8. Singh, A.K., Khatoon, I.S., Muazzam, M., Chaturvedi, D.K.: An Overview of Electricity Demand Forecasting Techniques, Network and Complex Systems, Vol.3, No.3 (2013)