LEARNING ANALYTICS

Petr Grolmus

Abstrakt

Příspěvek stručně shrnuje a popisuje problematiku Educational Data Mining a Learning Analytics, mapuje společné i rozdílné prvky obou vědních oblastí. Dále je představeno základní rozdělení metod běžně používaných pro zpracování dat získaných z e-learningových systémů spolu s příklady k jednotlivým typům. Kromě odkazů na zahraniční publikace jsou zmíněny i přínosné práce k problematice od českých autorů. Následně jsou představeny výzkumné cíle a hypotézy mé disertační práce spolu s volbou a zdůvodněním zvolených výzkumných metod.

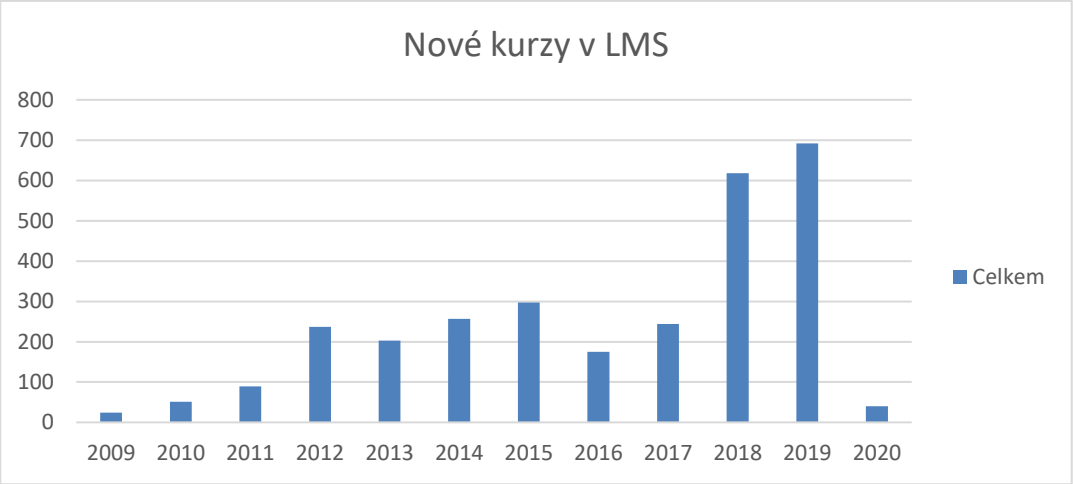
Klíčová slova

Learning analytics, educational data mining, learning management system, big data, big data analysis, prediction methods, classifiers, regressors, structure discovery methods, clustering, factor analysis, domains structure discovery, social network analysis, relationship mining, correlation mining, association rule mining, discovery with models, predcition study success.

1 Zpracování dat z e-learningových systémů

Za dlouhou dobu používání e-learningových systémů (LMS) se v nich a v logovacích souborech webserverů, na kterých jsou tyto LMS provozovány, nashromáždilo obrovské množství dat o přístupech a chování uživatelů. Od počátku byl kladen hlavní důraz výhradně na primární cíl LMS, kterým je předávání informací a vzdělávání. Tato často velmi obsáhlá data o chování uživatelů během procesu učení a konání testů zůstávala nevyužita a v podstatě jen zabírala cenné zdroje serveru. S rostoucím trendem využívání LMS, jež lze vysledovat i z dat systému Moodle provozovaném na Západočeské univerzitě v Plzni (ZČU) – viz obrázek 1, roste velkou rychlostí i množství uložených údajů.

Zájem o tyto provozní údaje lze pozorovat až v posledních letech a je významnou měrou spojen s rozvojem výzkumných oblastí Educational Data Mining (EDM) a o něco později i Learning Analytics (LA). Oblast EDM byla na vzestupu v letech 2008-2009 (Romero, Ventura, 2013), přestože její počátky lze datovat již do roku 2005 (Romero, Ventura, 2007). Mladší oblast LA lze datovat do let 2010-2011 (Ferguson, 2012; Juhaňák, Zounek, 2016).



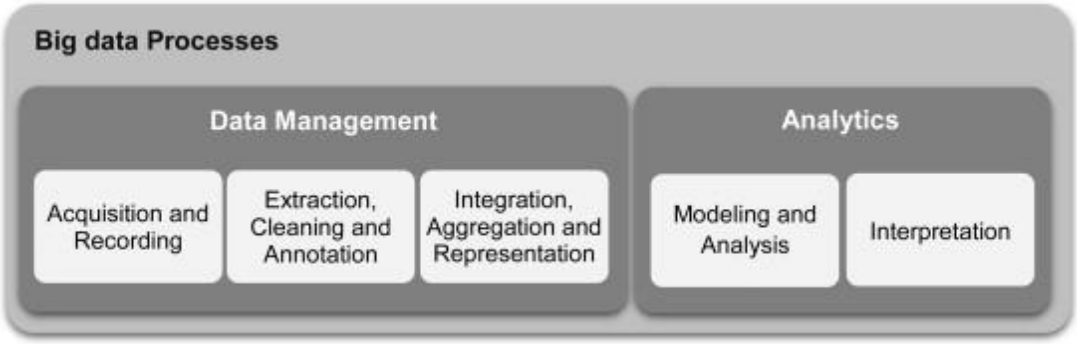
Obrázek 1. Rostoucí tendence počtu kurzů za rok

Obě výzkumné oblasti EDM a LA se v mnoha ohledech liší, co však mají společné je, že se snaží vyzískat maximum informací a souvislostí z dat generovaných a uložených v rámci samotného LMS, a také se snaží o zlepšování výuky, lépe chápat problémy ve vzdělávání a hledat postupy, jak tyto problémy minimalizovat a jak jim předcházet (Siemens, Baker, 2012). Za tímto účelem používají různé analytické a data mining metody, které umožňují získat důležité informace a poznatky o tom, jak se studenti v těchto systémech chovají, učí, plní úkoly a skládají testy (Juhaňák a kol., 2019).

(Siemens a kol., 2011) definují oblast LA jako:

"Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. Learning analytics are largely concerned with improving learner success."

Stejný zdroj uvádí, že LA je speciální reprezentací aplikace "*Big Data*" a analýzy v oblasti vzdělávání. Činnosti spojované s Big Data bývají obecně děleny na dvě základní části: data management a samotnou analýzu (Gandomi, Haider, 2015). Data management je dále členěn na získávání a zaznamenávání dat, extrakci a čištění dat a jejich následnou integraci, agregaci a reprezentaci. S takto získanými daty se pak provádí modelování, analýza a následná interpretace výsledků – viz obrázek 2.



Obrázek 2. Zpracování Big Data, zdroj: Gandomi, Haider, 2015

1.1 Hlavní rozdíly LA a EDM

Společné prvky LA a EDM jsou diskutovány výše. Jelikož se jedná o dvě oblasti výzkumu se svými komunitami, je třeba zdůraznit též jejich hlavní rozdíly. (Baker, Siemens, 2014) identifikují následující tři základní rozdíly.

Výzkumníci EDM se zajímají o automatické metody vyhledávání v datech z LMS systémů, zatímco výzkumnici na poli LA používají více metody řízené rozhodováním člověka při prozkoumávání shodných dat. Tento rozdíl přibližně sleduje rozdíl mezi dolováním dat a analýzou výzkumných dat v širším záběru vědecké literatury. Automatické metody vyhledávání v EDM pomáhají získávat nejlepší možnou předpověď, zatímco člověkem řízené metody v LA mohou vést k více popisným a srozumitelným modelům problematiky.

Dalším rozdílem je, že výzkumníci EDM kladou důraz na modelování konkrétních konstrukcí a vztahů mezi nimi. Výzkumníci LA se věnují spíše celkovým náhledům a popisům systémů a většímu porozumění konstrukcí.

Jako třetí hlavní rozdíl (Baker, Siemens, 2014) uvádějí, že výzkumníci v oblasti EDM se orientují na aplikace automatizované adaptace, jako je například podpora studentů pomocí vzdělávacího softwaru, který identifikuje potíže studenta a automaticky se přizpůsobí jeho znalostem nebo zkušenostem. Výzkumníci v oblasti LA hledají způsoby jak identifikovat studenty, kteří mají potíže s učebním materiálem, aby následně upozornili učitele nebo tutora kurzu, že konkrétní student potřebuje více podpory ve výuce.

2 Klíčové metody EDM/LA

Metodologie používané v EDM a LA pochází z mnoha zdrojů, ale největší inspiraci přebírají z oblastí výzkumu *data miningu* a analýzy dat obecně, a dále pak z psychometrie a dřívějších postupů používaných při měření vzdělávání. V mnoha případech specifika dat z LMS vedla k vytvoření nových postupů v EDM/LA odvozených od obecného *data miningu* (Baker, Siemens, 2014).

(Baker, Inventado, 2014) dělí metody EDM/LA do následujících čtyř základních tříd.

2.1 Metody predikce

U metod predikce je hlavním cílem vyvinout model, který odvozuje část dat, tzv. *predikovanou proměnnou* (obdoba závislé proměnné v tradiční statistické analýze) z kombinace ostatních částí dat, tzv. *prediktorovou proměnnou* (obdoba nezávislé proměnné v tradiční statistické analýze). Predikce se tedy snaží odkrývat vazby mezi jednotlivými proměnnými v množině dat.

V EDM jsou nejčastěji používanými typy predikce *klasifikátory* (např. Smith a kol., 2012) a *regressory* (např. Jayaprakash, 2014), které mají dlouhou historii v oblasti *data miningu* a umělé inteligence. U klasifikátorů se predikovaná proměnná roztřiďuje do dvou nebo více kategorií. Populární metody klasifikace a regrese jsou rozhodovací

stromy (*decision trees*), náhodné lesy (*random forrests*), rozhodovací pravidla (*decision rules*), lineární regrese (*stepwise regression*) a logistická regrese (*logistic regression*).

2.2 Structure Discovery – rozkrývání vazeb

Metody rozkrývání vazeb (*structure discovery*) zkoušejí nalézt vazby-závislosti v datech bez předchozí *a-priori* znalosti struktury dat. V tomto se zásadně liší od předchozí skupiny metod predikce, kde je nutná znalost formátu základních dat - proměnných. Oblíbené metody pro rozkrývání vazeb jsou *clustering, factor analysis, domain structure discovery* a *social network analysis*.

U *clusteringu* jsou data shlukována do množin na základě nějaké podobnosti. Většinou dopředu neznáme počet těchto množin, jsou vytvářeny v průběhu zpracování dat. Pokud by byly množiny dopředu známy, pak by se jednalo o dříve popisovanou klasifikaci. *Faktorová analýza* hledá skryté vazby mezi proměnnými a používá se hlavně pro ověření předpokládaného rozdělení. *Domain structure discovery* např. zjišťuje, které položky mapují specifické dovednosti nebo znalosti mezi studenty. Může být využito i pro vyhledávání expertů pro danou oblast znalostí. *Social network analysis* slouží k rozkrývání vazeb a interakcí mezi účastníky kurzu (např. Černý, 2018).

2.3 Relationship Mining – dolování/hledání vztahů

Cílem relačního dolování je objevní vztahů mezi proměnnými v množině dat s velkým počtem proměnných. Formou relačního dolování může být i hledání nejsilněji závislé proměnné na vybrané proměnné zájmu, tj. hledání nejsilnějších závislostí proměnných. Nejvyužívanějšími metodami hledání vztahů jsou *association rule mining, sequential pattern mining* a *correlation mining*.

Association rule mining zkouší hledat v proměnných jasné vazby KDYŽ-PAK. *Sequential pattern mining* zkouší hledat i jen dočasné asociace mezi událostmi, může být využito pro motivační analýzu konkrétního chování. *Correlation mining* se snaží dohledat zřetelnou pozitivní nebo negativní vazbu mezi proměnnými. Metoda je známá ze statistiky.

2.4 Discovery with model

U metody *discovery with model* je model sledovaného jevu připraven pomocí predikce, shlukováním (*clustering*) nebo znalostním inženýrstvím (v tomto případě je model vyvíjen hlavně pomocí lidských úvah než automatizovaných metod). Vytvořený model se pak používá jako součást druhé analýzy nebo nového odvozeného modelu. Tento způsob je hlavně využíván v oblasti LA, než v EDM, kde nalézají větší uplatnění dříve představené metody.

3 LA/EDM u nás i ve světě

Přestože LA i EDM jsou poměrně mladé obory, jejich důležitost rapidně stoupá s celosvětovou snahou o zvýšení kvality výuky. Metody LA a EDM poskytují nové metriky pro hodnocení úspěšnosti studentů, srozumitelnosti studijních materiálů, oblíbenosti předmětů nebo i učitelů. Lze jimi též hledat experty pro konkrétní oblasti.

Pro širokou škálu možných sledovaných nebo hledaných jevů v edukačních datech a velkou množinu použitelných metod a postupů nelze příliš dobře vzájemně poměřovat úspěšnost použitých metod právě z důvodu často rozdílných cílů.

Zajímavostí zcela jistě je zvýšený nárůst publikací v loňském roce o identifikaci studentů ohrožených studijní neúspěšností. Neméně zajímavou skutečností je, že tyto publikace pocházejí hlavně od výzkumných týmů jejichž mateřský jazyk není angličtina (např. Mouaici a kol., 2019; Falcão a kol. 2019; Simanca a kol. 2019).

Počet publikací k tématu LA/EDM je ve světě významně vyšší než v našem prostředí, jak dokládá i seznam literatury použité v tomto příspěvku. Přesto i u nás v České republice již lze nalézt výzkumníky v těchto oblastech. Příkladem může být přehledová studie (Juhaňák, Zounek, 2016), dále pak navazující práce (Juhaňák a kol., 2019), která v detailu mapuje chování studentů při online testování. Lze též zmínit práci (Černý, 2018), která se zabývá analýzou sociálních sítí v online vzdělávání.

4 Výzkumné cíle

Ve své disertační práci **chci prokázat, že na základě analýzy vzorců chování studentů v LMS lze vytipovat množinu studentů ohrožených studijní neúspěšností.** K tomuto cíli bych rád využil poslední diskutovanou skupinu metod – *discovery with model*.

Domnívám se, že v akademickém prostředí je možné vyžít skutečnosti, že se studijní předměty vyučované v po sobě jdoucích letech příliš, nebo dokonce vůbec, nemění. To samozřejmě platí i pro většinu elektronických kurzů v LMS, kde většinou dochází pouze k opravám chyb v textu a jen výjimečně k doplnění o nové poznatky.

Mám v plánu využít dat nasbíraných v předchozích letech u vhodně vytipovaných předmětů pro vytvoření modelu pomocí shlukování (*clusteringu*), při rozdělení na studenty, kteří byli úspěšní, a kteří neúspěšní. U dat opakujícího se kurzu lze pak pomocí metod LA určit, zda konkrétní student pomocí nalezených charakteristik inklinuje ke skupině dříve úspěšných nebo neúspěšných studentů.

„Vhodně vytipovaný předmět“ je takový předmět, který je vyučován v LMS, každý rok jej absolvuje několik desítek nebo stovek studentů, přímo v kurzu jsou vykonávány testy, které jsou bodově hodnocené a vhodná je též vzájemná interakce studentů a tutorů přes diskusní fórum kurzu. Nutností je znalost výsledného hodnocení studenta v kurzu, zda uspěl nebo neuspěl.

Naopak, nevhodný je úzce profilovaný předmět, který studují jen jednotky studentů, v LMS je jen část studijních materiálů a kurz vyžaduje jen minimální interakci studenta.

Formulace hypotéz disertační práce:

- Na základě dat z dřívějších akademických let lze u vybraných kurzů pomocí metod LA vytvořit vhodný model pro identifikaci studentů ohrožených studijní neúspěšností v aktuálním akademickém roce.
- Přesnost použitého postupu bude srovnatelná s dříve publikovanými jinými metodami, které predikují studenty ohrožené studijní neúspěšností.

Literatura

- BAKER, R., INVENTADO, P.S., 2014. *Educational Data Mining and Learning Analytics.* In: *Design of Learning Analytics Experiences*, p. 61-75.
- BAKER, R., LINDRUM, D., LINDRUM, M. J., & PERKOWSKI, D., 2015. *Analyzing Early At-Risk Factors in Higher Education e-Learning Courses*. In *Proceedings of the 8th International Conference on Educational Data Mining, Madrid, Spain, Jun 26-29, 2015*.
- BAKER, R., SIEMENS G., 2014. *Educational Data Mining and Learning Analytics.* In: *The Cambridge Handbook of the Learning Sciences*, p. 253-272.
- ČERNÝ, M., 2018. *Vybrané přístupy k učení se od druhých v online prostředí.* In: *ProInfo, Vol. 10, No. 2*, p. 147-166.
- FALCÃO T.P., a kol., 2019. *Students' Perceptions about Learning Analytics in a Brazilian Higher Education Institution*. In: proceedings of conference ICALT 2019, pp. 204-206.
- FERGUSON, R., 2012. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), p. 304–317.
- GANDOMI, A., HAIDER, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, volume 35, issue 2*, p. 137-144.
- JAYAPRAKASH, S. M., MOODY, E. W., LAURÍA, E. J., REGAN, J. R., & BARON, J. D., 2014. Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics, 1(1), p. 6-47.
- JUHAŇÁK, L., ZOUNEK, J., 2016. Analytika učení: nový přístup ke zkoumání učení (nejen) ve virtuálním prostředí. *Pedagogická orientace*, 26(3), p.560–583.
- JUHAŇÁK, L., ZOUNEK, J. ROHLÍKOVÁ, L., 2019. Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. In: *Computers in Human Behavior*, vol. 92, pp. 496-506, ISSN: 0747-5632.
- MOUAICI M., a kol., 2019 Detection of learners at risk of failure in online professional training. 3rd Annual Learning & Student Analytics Conference: An Ethical Vision of Learning Analytics Individuals VS Community, Nancy, France.
- ROMERO, C., VENTURA, S., 2013. Data mining in education. In: *WIREs Data Mining and Knowledge Discovery*, 3, p.12–27.
- ROMERO, C., VENTURA, S., 2007. Educational data mining: a survey from 1995 to 2005. In: *Expert Systems with Applications*, 33(1), p.135–146.
- SIEMENS, G. a kol., 2011. Open Learning Analytics: an integrated & modularized platform, http://solaresearch.org/wp-content/uploads/2011/12/OpenLearningAnalytics.pdf.
- SIEMENS, G., BAKER, R. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (p. 252–254).

- SIMANCA F., a kol., 2019. Identifying Students at Risk of Failing a Subject by Using Learning Analytics for Subsequent Customised Tutoring. [online] Available at: https://www.mdpi.com/2076-3417/9/3/448/htm> [Accessed 4 Jan. 2020]
- SMITH, V. C., LANGE, A., & HUSTON, D. R., 2012. Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses. *Journal of Asynchronous Learning Networks*,16(3), p.51-61.

Ing. Petr Grolmus
Západočeská univerzita v Plzni
Pedagogická fakulta, KVD
Klatovská třída 51
306 14 Plzeň
e-mail: indy@civ.zcu.cz;