BIG DATA APPLICATION

Daniela Orozova, Milena Georgieva

Abstract: Big data is the new buzzword and it concerns enormously large, complex and quickly changing sets of data. Big data is data-driven method for extracting interesting patterns from the data sets, analyzing and using them for further development. This paper presents the Big Data paradigm and suggests some applications within the eLearning realm.

Keywords: Big Data, eLearning, Knowledge discovery.

ПРИЛОЖЕНИЯ НА „ГОЛЕМИТЕ ДАННИ”

Даниела Орозова, Милена Георгиева

Абстаркт: „Големи данни“ е нова модерна дума, която се отнася към изключително големи, сложни и бързо променящи се набори от данни. „Големите данни“ са подход за откриване на интересни модели на поведение в набора от данни, анализирането им и използването им в последващи разработки. Тази публикация представя парадигмата „Големи данни“ и разглежда някои техни приложения.

Ключови думи: „Големи данни“, Електронно обучение, Откриване на знания.

1. Парадигмата “големи данни”

За дума на годината през 2013 год. е обявен термина big data или големите данни. Големите данни са феномен, който набира огромна скорост и все по-често става част от нашето ежедневие [1, 9, 13]. Бойд и Крауфорд в публикацията „Критични въпроси към големите данни“ дефинират феномена в технологичен, аналитичен и митологичен аспект. Според [1] определящи са изчислителната мощ, точността на алгоритмите за събиране, анализиране, свързване и съпоставяне на големи набори от данни. Основната задача на „големите данни“ е събраното количество данни да се обработи по такъв начин, че да предостави смислена информация на потребителя. Анализът на данните цели откриване на корелации и модели на поведение скрити в данните, като се използват техники за разглеждане на данните в дълбочина по начин, по който до сега не е бил възможен [1, 9, 13].

Парадигмата “големи данни”, включва в себе си множество технологии, с помощта на които се съхранява и обработва информация от различни източници. Източниците на данни могат да бъдат: уеб сайтове, електронни магазини, банкови транзакции, сензори и датчици, социални мрежи, интернет търсачки, GPS координати и други. От друга страна за големи данни говорим, когато е необходимо да се използва сериозно количество изчислителна мощ и техники като изкуствен интелект, които да се прилагат към големи масивни от данни [16]. Работата с големите данни освен знания по статистика изискват познания в областта на програмирането, базите от данни (релационни и NoSQL), език R, невронни мрежи, визуализация и много други средства.

Традиционните бази от данни са групирани и подготвени за нуждите на конкретна цел. От друга страна, големите данни, са сформирани по начин, така че да бъдат гъвкави и да отговарят на по-сложни въпроси. Изследването на големите данни е мултидисплинарен подход, включвайки в себе си освен обработка на данните и тяхното съхранение, разнообразност и подвижност.

2. Началото

През 1944 Ф. Ридър изчислява, че библиотеките в американските университети се удвоява за шестнадесет години. Според [2] библиотеката на университета в Йел през 2040 година ще притежава 0.2 милиарда тома, което приблизително възлиза на 2 петабайта (PB) информация. Това изследване се счита за един от първите опити в областта на „големи данни“.

През 1997 год. М. Кокс и Д. Елсуърт публикуват изследване, чиято големина на данните е толкова голяма, че надхвърля капацитета на паметта и дисковете. Това е и първата публикация, където се споменава термина big data (големи данни) [2, 3].

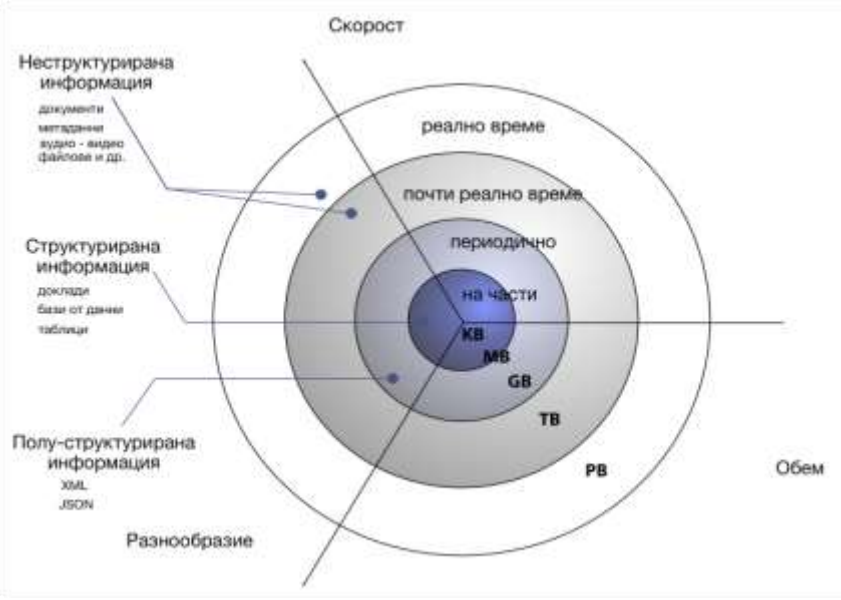
През 2000 година П. Лайман и Х. Р. Вариан в публикациите си [10, 11] определят количеството налична информация. Според резултатите получени от изследването към 1999 год. наличните данни възлизат на 1.5 екзабайта, което е около 1500 PB (петабайта) [2, 4, 12]. През 2013 год. е направено ново изследване, според което съхранената информация между 1999 год. и 2002 год. нараства с около 30% на година. Към 2014 год., по непотвърдени данни, съществуващите данни надхвърлят 1.2 зетабайта, което е около 1.3 трилион гигабайта.

През 2001 Д. Лейни публикува изследване със заглавия „3Д управление на данни: контролиране на данни обем, скорост и разнообразие“, където за пръв път се споменават трите V-та, залегнали в парадигмата на „големите данни“ [2, 4, 9].

3. Измерения на „големите данни”

Анализа на големите данни (Big Data) може да се дефинира като мултидисциплинарен процес на събиране, изчистване, анализиране и интерпретиране на различни по тип, произход и обем данни [3, 5, 6, 8]. Целта на един такъв анализ е откриване на скрити модели, корелации, дивиация и други интересни факти и събития скрити в самите данни.

Не съществува общоприета дефиниция за термина Големи Данни (**Big Data**). В по-старата литература се говори за три V-та, като през последните две години към обем (**Volume**), скорост (**Velocity**) и разнообразие (**Variety**), се добавя и достоверност (**Veracity**). На следващата фигура 1 е показана концепцията за три от тях (обем, скорост и разнообразие) [14, 15]. Измерението *достоверност* не е включена поради същността си.



Фигура 1. Концепция на измерения при “Big Data”

Обемът на обработваните данни е основната разлика между „големите данни“ и Data Mining парадигмата. Данни, които са толкова големи, че надхвърлят капацитета на релационните бази от данни се класифицират като „големи данни“ [2, 3, 5, 8, 12]. Скоростта на пристигане и обновяване на данните определя честотата на входиране на данните в платформата. Честотата на пристигане може да бъде: на части, през определен период от време, почти в реално време или в реално време.

Източниците на данни са от различен тип, докато генерираните данни са подчинени на различни стандарти. Често данните постъпват в системата във вид неподходящ за директна обработка и интеграция в платформа, което налага допълнителни стъпки за привеждане на данните във формат, отговарящ на дефинираните правила в платформата [8, 12, 13].

Данните се разделят в три класа според степента си на структурираност: структурирани, неструктурирани и полуструктурирани данни. Пример за структурирани данни са таблици, бази от данни, доклади генерирани от бази от данни и др. Неструктурираните данни се генерират от: интернет на събитията; интернет на хората (социалните мрежи - Facebook, Tweeter, LinkedIN И др.); интернет на нещата; интернет на локацията (мобилни телефони, смартфони, таблети и др.). Полуструктурирани данни се съхраняват в един от двата формата XML и JSON.

Четвърто измерение на „големите данни“ е достоверност на наличните данни. Данни генерирани в рамките на една организация предполагат достоверност и точност. За данни генерирани от външни източници не може да се гарантира достоверност, като това би довело до некоректност на резултатите на анализа.

4. Приложения на подхода „големи данни”

Пример за успешно прилагане на подхода „големи данни” е система за определяне на сферата на интереси на потребителите на сайтове като Netflix, Amazon, EBay и др. Алгоритмите използвани от тези сайтове са следят активността на потребителите и се опитват да определят диапазона на интересите на потребителя. На базата на събрани данни от други потребители се опитват да предвидят нови продукти, които биха представлявали интерес за клиента [3, 8].

Медицината е друго направление, която се цитира като подходяща сфера за прилагане на големите данни. Чрез събиране и обработване на информация за всеки отделен пациент ще се позволи разработване на персонализирано лечение, което ще повлияе положително върху състоянието на пациента.

Съществуват малък брой публикации в областта на големите данни и обучението. В сайта за безплатни курсове Coursera се предлага курс за „големи данни“ в образованието. Преподавателят Раян Бейкър - водещ учен в областта на анализа на данни, обръща внимание на алгоритми за обработване на вече събрани данни, като пренебрегва събирането, съхранението и изчистването на данните.

Книгата [12] се концентрира върху „големите данни“ в областта на образованието. В тази книга се дават примери за възможността да се използва парадигмата на големите данни в сфера на образованието. Авторите включват пример с курс в сайта Coursera на проф. Нг, преподавател по машинно обучение. На база на събраните данни за активността на студентите (във форума, класната стая, видео лекциите и поставени самостоятелни задачи) професорът определя кои лекции се възприемат по-лесно, къде студентите изпитват нужда от повторно преглеждане на предоставените материали, допълнителна информация и др.

С възхода на така наречените MOOC курсове за обучение (онлайн курсове подготвяни от водещи университети и предлагани от сайтове за обучение, често срещу

малка сума или безплатно) възниква и необходимостта от изучаване не само на способността на студентите да възприемат материала, но и подходите на преподаване, стила на тестовете, качеството на материалите, поредност на темите и др. [12]. Интересен обект на изследване е поредността на гледане на видео лекциите, колко и дали студентите гледат два или повече пъти едни и същи лекции, кои лекции се оказват по-трудни и кои по-лесни. Друг интересен аспект в обучението на студентите е проследяване на зададените самостоятелни работи, тестове, изготвяне на доклади и др. Този подход на обучение може да се разглежда като класически Data-Driven подход.

Освен работа с данните за всеки отделен студент, възможно е сравняване на група студенти и техните постижения и показатели с група студенти и техните показатели от предходни години [12]. По този начин би могло да се очертае обща тенденция в съответното направление или дисциплина, а преподавателят би могъл да коригира лекциите си спрямо нуждите и постиженията на студентите.

Литература:

[1] Boyd D., Crawford K., Critical questions for Big Data, 2012 уеб ресурс: http://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878
[2] Jifaa G., Linglingb Zh., 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014 Data, DIKW, Big data and Data science
[3] Mayer-Schoenberger M., Kenneth Cukier,Big Data Die Revolution, die unser Lebenveraendern wird, Redline Verlag, Munich, 2013
[4] Lyman P., Hal R. Varian, How Much Information?, 2000 уеб ресурс: http://www2.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf Senior
[5] Chikhale K., Data Mining: Exploring Big Data Analytics, Hadoop and Mapreduce, International journal of engineering science and research, уеб ресуср: http://www.ijesrt.com/issues%20pdf%20file/Archives-2014/August-2014/79.pdf
[6] M.P. van der Aalst W., Process Mining, Springer, 2011
[7] http://www.theverge.com/2013/12/26/5245008/amazon-sees-prime-spike-in-2013-holiday-season
[8] Jules J. Berman, Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information, Elsevier / Morgan Kaufmann, 2013
[9] Jifaa G., Linglingb Zh., 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014 Data, DIKW, Big data and Data science
[10] Lyman P., Hal R. Varian, How Much Information?, 2003 уеб ресурс: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf
[11] Lyman P., Hal R. Varian, How Much Information?, 2000 уеб ресурс: http://www2.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf Senior
[12] Mayer-Schoenberger M., Kenneth Cukier,Big Data , the future of education, Houghton Mifflin Harcourt, 2014
[13] O’Reilly media inc., Big Data now:2012 edition, O’Reilly, 2012
[14] http://www.gi.de/service/informatiklexikon/detailansicht/article/big-data.html
[15] http://velvetchainsaw.com/wp-content/uploads/2012/07/3VsBigData.jpg