University of Louisville: College of Business

Data, Text, and Web Mining

CIS 445-01: Data Mining

Dillon Wallace
8-30-2017

Data mining is a massive field of computer science that includes statistics, domain knowledge of the field that is utilizing data mining techniques, and many other fields that are tightly integrated for generating new knowledge from existing data a company or organization has available to them. Data mining is comprised of many different types of data, three of them are: Data, Text, and Web. Data is generally stored in databases, data warehouses, or is data generated from a transactional system; Text is generally unstructured narrative data that humans can read and understand, but computers have a harder time deciphering what the text means; Web mining is generally when web spiders/scrappers download massive amounts of website data. Applying Data Mining techniques to the collected data can generate knowledge that can allow business to get an edge over the competitors in their domain. Almost all business can gain some benefit by applying data mining techniques to the data they have access to, and learn about their industry or clients.

Data mining has many different techniques that experts can apply to extract meaningful information from the data. Some of these techniques include: neural networks, regressions, fuzzy logic, and cluster analysis. Neural Networks is a very interesting topic in computer science/data mining, it is computer software that is able to functions or mimic the way neurons work in a human brain. Regressions are statistical formulas that generate predictive models based on data. Fuzzy logic is a way of analysis that has “many truths” (the computer uses a binary 0 or 1 for true or false). Cluster analysis is a way to spot outliers or group data in special ways. There is a lot of different software packages or programming languages, some free or commercial that is for data analysis and data mining for the techniques above: R (programming language with packages for data mining), Python (also programming language with packages for data mining), MATLAB (programming language and packages), SAS (language and packages), and there are

many others. When all or some of these tools are combined with these various techniques available for data mining, experts can then process datasets and give vary interesting results to managers, customers, employees, or investors/donors, that can help make decisions or increase purchases/donations.

Text data mining is where analysts process narratives or sentences of text (which is very much UNSTRUCTURED, this is very important to note because in databases most information is structured and every column has a set meaning), and mine them for useful information. One very hard part of analyzing text is the structure, trying to figure out what is what and what everything means in the context of a sentence and what do shorten/abbreviations and jargon mean. The English language and all languages have very different meanings for the same word and to a computer it would be hard to figure out the meaning, because until it is taught it does not know what to associate with all the words and how to recognize and process sentence structure. For example, the sentence, "Dr. Zurada is the best at data mining", has a lot more into it than we realize at first glance. The subject is Dr. Zurada, the linking verb in the sentence is- is, which links the subject to the predicate, best is an adjective or adverb that indicates the subject's level, data mining is a noun that is a process of creating new knowledge by using data. The computer must learn all of this and how the English (or any other) language semantics, grammar, and how the language's dictionary work. An example of a source of text data would be a medical chart, which the medical field has a lot of jargon, there are even entire classes explaining the words, prefixes, and suffixes. Most clinical charts use clinical narratives, which is full of information that clinical staff do not want to break up into different fields such as check boxes or dropdowns because they can use equipment that can translate their words into the computer for them. These charts have a lot of valuable information such as: height, weight, diagnosis, problems, situations

and other information depending on the field, for example, in the substance abuse sector the charts must have the age of first use, what they use, frequency of use, cravings and urges, how the client is feeling, and other various things needed for substance use assessments. Processing clinical information this way can create trends for clients, and determine if the program is helping or if they need additional help. A free open source clinical data analysis tool is cTAKES which utilizes Apache UIMA which is a text processor that allows computers to have sentence detection, tokenization, part-of-speech tagging, dependency parsing, named entity recognition, concept normalization, assertions detection and UMLS (Unified Medical Language System) [which is published by the National Library of Medicine and DHHS] to process batches of clinical documentation, which has been utilized by the Mayo clinic. Something really interesting was it could process if the patient or family has history of disease (from the clinical assessment) and able to deal with positives or negatives for test results or diagnosis, and even process clinical timelines (patient was sick on Friday, worsen Sunday, better Monday), and drugs the patient was taken or given (along with the information attached like when, how it was given, time, etc). This has many applicable uses in the healthcare field, with huge health systems all using the same Electronic Health Record to store all of their clinical information they have a wealth of data that they may not know is tapped that may even be able to help public health and society as a whole.

Another type of data that can be mined is web data. Web crawlers/spiders browse every website it can get access to scrape the entire website (download all information and process the text and crawl all hyperlinks and process the words) to get an accurate picture of the website and all the words and information it contains. Web documents are formatted with HTML or XML which contain information between tags. This information is all wrapped up between tags inside tags (<body> <h1>Title</h1> Body </body>) which all have to be

unwrapped and processed in order for search engines to be able to index all the words. These words are then able to be looked up via a query or web search and providers like Google have algorithms to filter the websites and learn on input based on geographic area, and the key words that is used in queries to give the user a list of websites it thinks they will want. Some web mining techniques are used to track user's movements on a website in order to revamp it. For example, the company mouseflow has software that tracks the way you use your mouse in order to generate heat maps to tell designers what is working, what catches peoples eye, and what does work and what does not catch people's eyes to modify it so more people can use it more efficiently and make more purchases. This is very useful, especially if the store does not have a bricks-and-mortar and only has the online presence, for these companies this software could definitely have a make-it-or-break it scenario. Web mining has many application uses for all of the companies that have an online presence in order to bring in more customers, and interact with the website more efficiently.

Data mining is a branch of computer science that has a diverse range of other fields included in it. Data mining also involves various types of data that can be mined, two of which are text and web. Text data mining is a way the computer can break down sentences and turn it into data that can be mined for new information, which can be vary applicable in certain industries such as healthcare. Web data mining is the way computers process web pages or actions on a web page in order to generate new knowledge about their user's habits. Almost all industries can learn new trends or get more information from utilizing the many data mining tools, some free (R, cTAKES) or some paid (SAS, Fuzzy Logic Toolkit), which can give any industry that competitive advantage against a competitor that can bring in more revenue, clients,

or improve existing systems (manufacturing, supply chain management) which can increase productivity/reduce cost.

Works Cited:

http://dmr.cs.umn.edu/Papers/P2004_4.pdf

https://vector.childrenshospital.org/2013/12/ctakes-turning-clinical-notes-into-knowledge/

http://ctakes.apache.org/

https://uts.nlm.nih.gov/

http://dmr.cs.umn.edu/Papers/P2004_4.pdf

http://liacs.leidenuniv.nl/~bakkerem2/dbdm2007/05_dbdm2007_Data%20Mining.pdf

http://www.kdnuggets.com/

http://xplqa30.ieee.org/xpl/bkabstractplus.jsp?bkn=5769527

https://pdfs.semanticscholar.org/48ea/0d8f74f59505c3981f70e9f16c4f2b86b894.pdf

http://www.academia.edu/download/45679869/j.compbiomed.2005.08.00320160516-4867-5wllhl.pdf