A Survey on Domain-Specific Languages for Machine Learning in Big Data

Ivens Portugal

Agenda

- Introduction
  - Big Data and Machine Learning
- Research Problem and Goal
- Approach
- Research Progress
- Results and Contributions

Introduction: Big Data

- Definition (Tanaka, 2013)
  - Relates to datasets whose size is beyond the ability of typical database software to capture, store, manage, and analyze.
- 3V (Gartner, 2012)
  - Volume, Velocity, Variety
- Example (Chen, 2014)
  - 267 million transactions in Walmart per day
  - 3 billion pieces of content generated on Facebook per day
  - 30 petabytes of image data generated by Large Synoptic Survey Telescope (LSST) per day
  - 60 terabytes of data generated by Large Hadron Collider (LHC) per day
  - 1 day = 24 * 60 * 60 = 86400 seconds

Introduction: Machine Learning

- Definition (Simon, 2013)
  - A field of study that give computers the ability to learn without being explicitly programmed.
- Created in 1950s, popular since 1990s.
- Algorithms (Ullman, 2012)
  - Bayesian Network, k-Means, Clustering, Logistic regression, Support vector machine, Neural network, and many more.

Research Problem and Goal

- Machine learning in Big Data
    - More data, more learning, more research, more results
- Why isn't everybody using?
- Goal:
    - Let's make it easy to develop.
    - Let's survey and analyze the languages being used
        - GPL - C, C++, Java, UML
        - DSL - SQL, Matlab, HTML

Approach

- DSL - classification (Van Deursen, 2000; Fowler, 2010)
  - Requirements, Programming, Modeling
  - Textual, Graphical
  - Internal, External
  - Dynamically typed, Statically typed
  - Declarative, Functional
  - Translation (Compilation), Interpretation
  - (External) Target Platform and Execution Engine
  - (Modeling) Descriptive, Prescriptive model

Approach

- DSL
  - Scala (Scala, 2015)
  - OptiML (Sujeeth, 2011)
  - VisuML (Breuker, 2014)
  - Infer.net (Minka, 2015)
  - ScalOps (Weimer, 2011)

- Others
  - HiveQL (Thusoo, 2009)
  - Pig Latin (Olston, 2008)
  - Salang (Sawmill, 2015)
  - SCOPE (Chaiken, 2008)
  - Spark (Spark, 2015)

Approach

DSL	Requirements/ Programming/ Modeling	Textual/ Graphical	Internal/ External	Dynamically/ Statically typed	Declarative/ Functional	Translation/ Interpretation	Target Platform	Execution Engine	Descriptive/ Prescriptive model
ScalOps									
OptiML									
Scala									
VisuML									
...									

Research Progress

DSL	Requirements/ Programming/ Modeling	Textual/ Graphical	Internal/ External	Dynamically/ Statically typed	Declarative/ Functional	Translation/ Interpretation	Target Platform	Execution Engine	Descriptive/ Prescriptive model
ScalOps	Programming	Textual	Internal (Scala)	Statically typed	Declarative	Translation	-	-	-
OptiML	Programming	Textual	Internal (Scala)	Statically typed	Functional	Translation	-	-	-
Scala	Programming	Textual	Internal (Java)	Statically typed	Functional	Translation	-	-	-
VisuML	Modelling	Textual	Internal (several)	-	-	-	-	-	Descriptive
...									

Results and Contributions

- Results
  - Identify strengths and weaknesses of languages
  - Identify features that languages have to ease ML in BD system development
- Contributions
  - Better understanding of the development of these systems
  - Beginners may better choose a language to start developing, modeling or gathering requirements for this domain

References

- Breuker, D. (2014). Towards Model-Driven Engineering for Big Data Analytics--An Exploratory Analysis of Domain-Specific Languages for Machine Learning. In System Sciences (HICSS), 2014 47th Hawaii International Conference on (pp. 758-767). IEEE.

- Chaiken, R., Jenkins, B., Larson, P. Å., Ramsey, B., Shakib, D., Weaver, S., & Zhou, J. (2008). SCOPE: easy and efficient parallel processing of massive data sets. Proceedings of the VLDB Endowment, 1(2), 1265-1276.

- Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347.

- Fowler, M. (2010). Domain-specific languages. Pearson Education.

- Minka, T., Winn, J.,Guiver, J., Webster, S., Zaykov, Y., Yangel, B., Spengler, A., Bronskill, J. (2014). Infer.NET 2.6, Microsoft Research Cambridge, 2014. http://research.microsoft.com/infernet.

- Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A. (2008). Pig latin: a not-so-foreign language for data processing. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1099-1110). ACM.

- Phil Simon (2013). Too Big to Ignore: The Business Case for Big Data. Wiley. p. 89. ISBN 978-1-118-63817-0.

- Rajaraman, A., & Ullman, J. D. (2012). Mining of massive datasets (Vol. 77). Cambridge: Cambridge University Press.

References

- Sawmill (2015). http://www.sawmill.net/cgi-bin/sawmill8/docs/sawmill.cgi?dp+docs.technical_manual.salang+webvars.username+samples+webvars.password+sawmill. Accessed on November 24th, 2015.

- Scala (2015). http://www.scala-lang.org. Accessed on November 24th, 2015.

- Spark (2015). http://www.spark-2014.org. Accessed on November 24th, 2015.

- Sujeeth, A., Lee, H., Brown, K., Rompf, T., Chafi, H., Wu, M., ... & Olukotun, K. (2011). OptiML: an implicitly parallel domain-specific language for machine learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 609-616).

- Tanaka, T., (2013). Big Data Application Technology: An Overview.

- Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., ... & Murthy, R. (2009). Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment, 2(2), 1626-1629.

- Van Deursen, A., Klint, P., & Visser, J. (2000). Domain-Specific Languages: An Annotated Bibliography. Sigplan Notices, 35(6), 26-36.

- Weimer, M., Condie, T., & Ramakrishnan, R. (2011). Machine learning in ScalOps, a higher order cloud computing language. In NIPS 2011 Workshop on parallel and large-scale machine learning (BigLearn) (Vol. 9, pp. 389-396).

A Survey on Domain-Specific Languages for Machine Learning in Big Data

Ivens Portugal