Progetti

Data-mining

Come lavorare sul progetto

- L'**obiettivo principale**: usare il progetto per approfondire alcuni degli argomenti visti a lezione.
- **Importante**: dimostrare di aver capito l'articolo associato al progetto.
- Il progetto scelto deve essere svolto seguendo uno dei **percorsi** seguenti
  - **Ricerca**: approfondire l'aspetto algoritmico, cercare soluzioni simili, confrontare i risultati
  - **Sviluppo**: risolvere il problema usando gli strumenti visti: API Weka Java, Hadhoop. Porre l'accento su tempi di esecuzione e vantaggi implementativi.

Cosa consegnare

- Consegnare una relazione contenente:
  - Descrizione del problema
  - Descrizione del dataset
  - Algoritmi utilizzati
  - Ottimizzazioni particolari
  - Sistema utilizzato: Hadoop,Mapreduce, SO standard
  - Risultati ottenuti

Links

1) Alcuni link che possono essere utili per i progetti orientati allo *sviluppo* :
   - http://mahout.apache.org/
   - Weka and Hadoop http://markahall.blogspot.co.nz/2013/
   - Creare un cluster con più macchine http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html

     http://blog.cloudera.com/blog/2014/01/how-to-create-a-simple-hadoop-cluster-with-virtualbox/

Email Classification

- **Dataset**: Enron Email
  - This dataset contains data from about 150 users, mostly senior management of Enron, organized into folders.
  - May 7, 2015 Version of dataset (about 423 MB, tarred and gzipped).
  - URLs:
    https://www.cs.cmu.edu/~./enron/
    http://ceas.cc/2004/168.pdf
- **Preprocessing**: use this preprocessing on the original dataset

Project 1: Email Classification with Co-Training

- Paper: Kiritchenko, Svetlana, and Stan Matwin. *"Email classification with co-training."* Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research. IBM Corp., 2011. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.408.8821&rep=rep1&type=pdf
- Algorithms:
  - Support Vector Machines
  - Naive Bayes
- Task: Email Classification using Co-Training

Assegnato Sandroni

Project 2: Email Classification

- Paper: Brutlag, Jake D., and Christopher Meek. "Challenges of the email domain for text classification." ICML. 2000. http://research.microsoft.com/pubs/73532/AF1-1.pdf
- Algorithms:
  - Support Vector Machines
  - WEKA's Incremental Classifiers
- Task: Email Classification

Project 3: Learning Rules that Classify E-Mail

- Cohen, William W. "Learning rules that classify e-mail." AAAI spring symposium on machine learning in information access. Vol. 18. 1996. http://www.aaai.org/Papers/Symposia/Spring/1996/SS-96-05/SS96-05-003.pdf
- Algorithms:
  - method based on TF-IDF weighting,
  - Rule learning algorithm
- Task: Email Classification

Project 3.1: E-mail sorter

- Crawford, Elisabeth, Judy Kay, and Eric McCreath. "IEMS-the intelligent email sorter." ICML. Vol. 2. 2002. http://www.cs.cmu.edu/~ehc/papers/syd/iems.pdf
- Algorithms:
  - TF-IDF,
  - Rule learning algorithm
- Task: Email Sorter

Assegnato Zaccaria

Project 4: Using MapReduce for Email Classification

- Xu, Ke, et al. "A MapReduce based Parallel SVM for Email Classification." Journal of Networks 9.6 (2014): 1640-1647. http://www.ojs.academypublisher.com/index.php/jnw/article/viewFile/jnw0906164016 47/9522
- Algorithms:
  - Use mahout.apache.org Naïve bayes
- Task: Email Classification and MapReduce

Assegnato Rocca

Project 5:
E-mail authorship attribution

- Schmid, Michael R., Farkhund Iqbal, and Benjamin CM Fung. "E-mail authorship attribution using customized associative classification." Digital Investigation 14 (2015): S116-S126. http://www.sciencedirect.com/science/article/pii/S1742287615000572
- Algorithms:
  - Use WEKA Association Rules
- Task: Email Classification

Assegnato Procopio

anonymous web data

- **Dataset**: msnbc.com
  - This data describes the page visits of users who visited msnbc.com on September 28, 1999.
  - Number of users: 989818
  - Average number of vitis per user: 5.7
  - Number of URLs per category: 10 to 5000
  - The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports"
  - URLs: http://kdd.ics.uci.edu/databases/msnbc/msnbc.html

Project 6: Frequent Pattern Mining in Web Log Data

- Iváncsy, Renáta, and István Vajk. "Frequent pattern mining in web log data." Acta Polytechnica Hungarica 3.1 (2006): 77-90. http://arxiv.org/pdf/1301.7401.pdf
- Algorithms:
  - Expectation–Maximization
  - K-means
- Task: mining in web log data

Assegnato Squillaci

Project 7: Automatic recommendation of web pages

- Sumathi, C. P., R. Padmaja Valli, and T. Santhanam. "Automatic recommendation of web pages in web usage mining." Internat. J. on Computer Science and Engg 2.9 (2010). http://arxiv.org/pdf/1301.7401.pdf
- Algorithms:
  - Expectation Maximization clustering
  - ...
- Task: web usage mining

Project 8: Web User Session Clustering

- Poornalatha, G., and Prakash S. Raghavendra. "Web user session clustering using modified K-means algorithm." Advances in Computing and Communications. Springer Berlin Heidelberg, 2011. 243-252. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.269.5866&rep=rep1&type=pdf
- Algorithms:
  - Modified K-Means
  - K-Means
- Task: Web User Session Clustering

movie recommendation service

- **Dataset**: MovieLens 20M Dataset
  - This dataset (ml-20m) describes 5-star rating and free-text tagging activity from MovieLens, a movie recommendation service.
  - It contains 20000263 ratings and 465564 tag applications across 27278 movies.
  - These data were created by 138493 users between January 09, 1995 and March 31, 2015.
  - URLs:
    http://files.grouplens.org/datasets/movielens/ml-20m-README.html
    http://files.grouplens.org/datasets/movielens/ml-20m.zip
    (132 MB)

Project 9: recommendation algorithm

- Kim, Choonho, and Juntae Kim. "A recommendation algorithm using multi-level association rules." Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on. IEEE, 2003. http://ai.dgu.ac.kr/publication/pds/WI2003.doc
- Algorithms:
  - Association Rules
  -
- Task: recommendation using association rules

Assegnato Pistocchini

Project 10: Recommender System Using Naive Bayes Classifier

- Ghazanfar, Mustansar, and Adam Prugel-Bennett. "An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering." (2010). http://eprints.soton.ac.uk/268483/1/IMECS2010_MustansarAliGhazanfar.pdf

- Algorithms:
  - Naive Bayes
  -

- Task: Recommender System, Naive Bayes Classifier, Collaborative Filtering

Project 11: Collaborative Filtering

- CarlKadie, JohnS Breese DavidHeckerman. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering." Microsoft Research Microsoft Corporation One Microsoft Way Redmond, WA 98052 (1998). https://www.researchgate.net/profile/Carl_Kadie/publication/235357340_Empirical_Analysis_of_Predictive_Algorithms_for_Collaborative_Filtering/links/546386850cf2c0c6aec4d910.pdf
- Algorithms:
  - Decision tree
  -
- Task: Collaborative Filtering, Decision tree

Assegnato Bollini

Project 12: Collaborative Filtering as Classification problem

- Xia, Zhonghang, Yulin Dong, and Guangming Xing. "Support vector machines for collaborative filtering." Proceedings of the 44th annual Southeast regional conference. ACM, 2006. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.400&rep=rep1&type=pdf

Assegnato Smillovich

- Algorithms:
  - Support vector machines
  -
- Task: Collaborative Filtering, Support vector machines