Student Projects for AIM3

Database Systems and Information Management Group, TU Berlin

Overview

1. **Porting algorithms from Mahout to PACT**
2. **Adding Graph Analysis to Mahout**
3. **Goals and Timeline**

Porting RowSimilarityJob

RowSimilarityJob

- computes the pairwise similarities of the rows of a sparse matrix using a similarity metric like cosine, ...

Usecase

- find similar text documents or Amazon-like suggestions of „people who like {x} also like {y}“

Task (2 Students)

- port Mahout's implementation to PACT
- compare code complexity and running time

Porting Naive Bayes

Naive Bayes

- classifier (learn decision making from examples)
- based on Bayes Theorem, assumes feature independence (→ naive)

Usecase

$$P(c|D) = \frac{P(D|c)\,P(c)}{P(D)}$$

- detect spam mails, ...

Task (2 Students)

- port Mahout's implementation to PACT
- compare code complexity and running time

Comparing K-Means

K-Means

- clustering (group data points into sets of similar points)
- simple, iterative algorithm

Usecase

- group news by topics, find users with similar taste

Task (1 Student)

- PACT implementation already available
- compare code complexity and running time

Porting Collocation Identification

Collocation

- find sequences of words or terms that co-occur more often than would be expected by chance

Usecase

- find lexical units in a text that can be used as features in a vectorized representation

Task (2 Students)

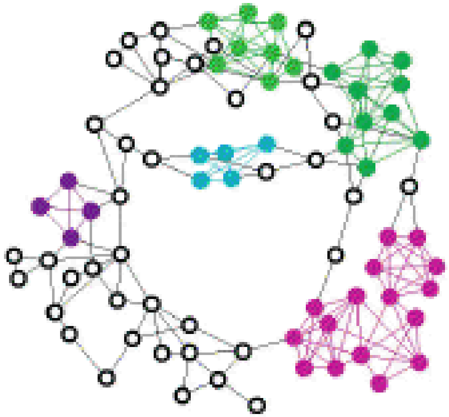
- Port Mahout's implementation to PACT
- compare code complexity and running time

Network Analysis

Finding k-trusses in a graph

- Augment edges with degrees, enumerate triangles
- find trusses (truss = relaxation of a clique)

Usecase



- finding groups of highly interconnected people in a social graph

Task (2 Students)

- Add an implementation to Mahout
- explain code complexity and running time

Network Analysis

PageRank

- Used to measure the „relative importance“ of a node in a network

Usecase

- ranking in web search

Task (1-2 Students)

- Port the Pegasus implementation to Mahout
- explain code complexity and running time

Goals and Timeline

Goals

- PACT projects: create 3 slides about your results: algorithm definition, anatomy of PACT implementation, comparison with Hadoop implementation
- Mahout project: have your patch contributed, create explanatory slides

Timeline

- 10min presentation with roadmap in 2 weeks (27. May)
- first working prototype code (10. May)
- slides with results in six weeks (24. June)
- final presentation/report towards the end of this term