Master Thesis

Recommender System for Galaxy Tools and Workflows

(Find similar tools and predict next tools in workflows)

Anup Kumar

Advisor:	Dr. Björn Grüning
Examiners:	Prof. Dr. Rolf Backofen, Prof. Dr. Wolfgang Hess

University of Freiburg
Faculty of Engineering
Department of Computer Science
Chair for Bioinformatics

July, 2018

Writing period

09. 01. 2018 – 09. 07. 2018

Examiners

Prof. Dr. Rolf Backofen and Prof. Dr. Wolfgang Hess

Advisor

Dr. Björn Grüning

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date

Signature

Abstract

This study undertakes two tasks in order to build a recommendation system for Galaxy tools and workflows - one is to find similarity in tools and another is to predict next tools in workflows.

To find similarities in the tools, we need to extract information about each tool under multiple categories like a tool's name, description, input and output file types and help text. We take into account these categories one by one and compute similarity matrices. Each row in a similarity matrix keeps similarity scores of one tool against all other tools. These similarity scores depend on the similarity measure used to compute the score between a pair of tools. We compute three such similarity matrices, one each for input and output file types, name and description and help text. Now, we are posed with a question of how to combine these matrices? One simple solution would be to take an average of the corresponding rows in each of the three matrices for one tool. This would give us one final row containing the tool's similarity with all the other tools given the categories of information. But assigning equal importance factor to all the corresponding rows in matrices might not be optimal? To find and optimal combination, we use optimization and learn importance weights (3 positive real numbers which sum up to 1) on these rows for each tool. Based on the similarity measure, we assign a true similarity value. We take an array of 1.0 as true value because we use jaccard index and cosine similarity as similarity measures.

Next task deals with the prediction of next or future tools for workflows. Wouldn't it be convenient to leaf through a set of next possible tools as a guide while creating workflows? It can assist the less experienced Galaxy users in creating workflows who could be unsure about which tools come next. In addition to this benefit, it can also curtail the workflow creation time. To achieve that, we need to learn the connections among tools in order to be able to predict the next possible ones. To preprocess each workflow, we compute all the paths bridging the starting and ending tools. These paths contain a set of connected tools. We follow a classification approach to predict the next tools for a given tools or a set of connected tools. For classification, we use a

variant of neural networks (long short-term memory) which performs well for learning long range dependencies (tools connections). We report the accuracy as precision.

Zusammenfassung

German version is only needed for an undergraduate thesis.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

1.0.1 What is Galaxy?

1.0.2 Galaxy tools

1.0.3 Tool categories

1.0.4 Motivation

2 Approach

2.1 Data preprocessing

2.1.1 Tools attributes

2.1.2 Data extraction

2.1.3 Clean data

Refine tokens

2.2 Word embeddings

- Latent Semantic Indexing
- Paragraph Vectors

2.3 Similarity measures

Jaccard index

Cosine similarity

2.4 Optimization

Gradient descent

Backtracking line search