Masaryk University
Faculty of Informatics

Similarity of programming problems

Bachelor's Thesis

Dominik Gmiterko

Brno, Spring 2018

This is where a copy of the official signed thesis assignment and a copy of the Statement of an Author is located in the printed version of the document.

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Dominik Gmiterko

Advisor: Radek Pelánek

Acknowledgements

These are the acknowledgements for my thesis, which can span multiple paragraphs.

Abstract

This is the abstract of my thesis, which can
span multiple paragraphs.

Keywords

similarity, metrics, programming, keyword2, …

Contents

Introduction

Tutoring systems are computer-based systems designed to introduce users into various domains. They usually have large amount of items which enables them to provide personalized experience. To maintain this large pool of items efficiently we need to be able to decide which items are useful and which are not.

Besides Introduction and Conclusion chapters, this thesis is structured into three additional chapters. First chapter talks in general about problem of measuring similarity of programming problems. It explains difference between program and programming problem, which data we have available and techniques used for measuring similarity of problems. Second chapter advances level deeper and describe everything what is specific to data we used. First part of chapter describes programming environment of Robotanik and data from it. Second part focuses in detail on metrics we used in experiments. Last chapter gives overview of implementation and usage of metrics and their evaluation.

1 Similarity

In this chapter we will talk in general about questions in learning systems, and computing their similarity. Most of the chapter focuses on explaining what kinds of data are available when comparing questions in learning systems and techniques to do so. Last section describes goals of the thesis.

A lot of research has been dedicated to similarity in many different fields computer science like bioinformatics (sequence alignment, similarity matrix of proteins), information retrieval (document similarity), plagiarism detection and many more.

One closely related area is recommender systems which differs from problem similarity only slightly. Both areas are distinguishing users and items. Only difference is that we know how well user did while solving specific item and recommender systems use rating of the items.

Main difference is that we can use more data about problem. We also have some problem statement and data about performance of students when solving problem.

1.1 Items

In this work we use the term "items" (problems, questions, assignments) when we refer to single entry in educational system which users can answer to. Since many aspects of this work are generally applicable we decided to use this general term. In some learning systems this can refer to simple choice from two options in another complex tasks which user solves in matter of minutes. On other side of the spectrum are systems for teaching introductional programming. Users tend to spend few minutes solving each task and there is fewer of them.

To further specify the context of our research, we will describe characteristics of items. For computing similarity of items it is most important knowing which data are available to us. Therefore we describe items by sources of data can be used for measuring similarity.

- **Item statement:** specification of the item that a learner should solve, e.g., as a natural language description of the task.
- **Item solutions:** details about solutions obtained from learners or sample solution to item.
- **Learner's performance:** for example item solving times, correctness of answer, number of attempts needed.

This description of item is broad enough to cover most of learning systems. In next chapters we will discuss two systems in particular - umimecesky a umimematiku.

1.2 Why is similarity of items useful

As we mentioned previously key part of learning solving of educational items.

1.3 Computing similarity of items

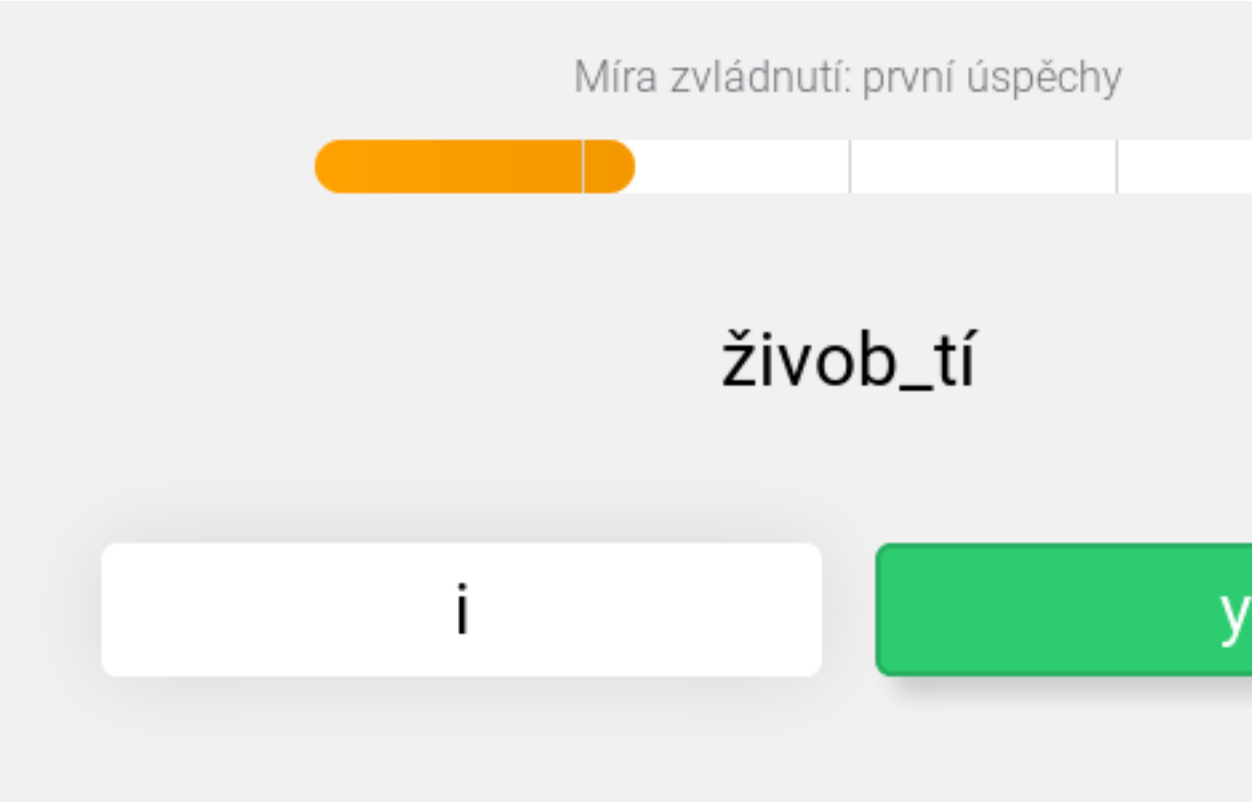
The general approach to measuring and using similarity of educational items

1.4 Used datasets

In our analysis we use both real data from educational system and simulated data. There is a reason why use both as only real-world data are useful for concluding any results. However evaluation of this data is often complicated as we do not know truth about many of their aspects. That's why we used simulated data for validating some of our conclusions. We will talk more in depth about how we generated simulated data in next chapter when describing their specific usage.

Most of used real-world data comes from system Umíme česky[1]. Later we have validated our results by also using data from its sibling

1. <https://umimecesky.cz/>



system Umíme matiku[2]. We think it is useful as data come from another context but are provided in same format and therefore can be used directly in previously created tools.

1.4.1 Umíme česky

Umíme česky[3] is system for practice of Czech grammar. System contains multiple exercise types, but in our analysis we use only one exercise - simple "fill-in-the-blank" with two possible answers. This type of exercise can be then viewed by student in multiple ways.

We focused only on "fill-in-the-blank" exercises but they can still be used to train many concepts of Czech grammar.

2. <https://umimematiku.cz/>
3. <https://umimecesky.cz/>

1.4.2 Umíme matiku

2 Evaulation

3 Conclusion

A An appendix

Here you can insert the appendices of your thesis.