

**Oskar Jarczyk**

*Department of Social Informatics*

*Polish-Japanese Institute of Information Technology, 2013*



**Programming skills on GitHub portal, thoughtful analysis with prediction mechanism**

**Abstract**

Is there a correlation between frequency of using different programming languages ? Experienced programmers intuition would suggest, that on GitHub portal repositories tagged as JavaScript projects are very often co-tagged as a code having CSS style sheets. How to prove it ? We propose a mechanism of predicting a repository on which GitHub registered developer will work on, only by analysis skills matrix.

**Categories and Subject Descriptors**

-

**General terms**

Machine learning, Big Data analysis

**Keywords**

*Machine learning, big data analysis, OSS (open source software),*

*SVM, decision trees, OSSD (open source software development)*

1. Problem definition
2. Data exploration
   1. Downloading user statistics

It’s possible to get a countable amount of GitHub logins which can be considered as the most active. With those logins we intend to get their skills associated with them on profile and their membership to repositories.



Rysunek 1: User pro-activity classification proposed by Paul Miller

* 1. Analyzing commits done by users

Decision tree

* 1. N count of etickets exists in a tree

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | … | R |
| A1 | 0,4 | 1 | 0,1 |  | R1 |
| A2 | 0,9 | 0,3 | 0 |  | R2 |

SVM learning