In the second part of research we will construct our graph –base classifier and evaluate It by using 5 fold validation technique, our main goal is train our classifier for answering a simple question if user like /or dislike a new subject based on his passed features and subject.

# Research steps

## Step 1 – crawling pinterest and saving data

<most content will be copy form thesis proposal >

At end of the first part of the research we successfully collect data of 51,092 users in pinterest , all content we save in local xml format files.

## Step 2 - parsing data and creating graph

In the second phase research we will focused on analyzing the data and represent into large scale mathematical graph which allow us to build and represent connections between subject for retrieving a valid recommendation.

In this research our resources was:

* 1 X64 i5 Machine with 4 GB Ram, 500GB hard drive for storage.
* 1 x64 i5 Machine with 16 GB ram 2Tb hard drive for storage
* Both machine are running under windows 7 x64 version ,connected to internet buy using university infrastructure, both of those machine are service .

### Build graph – using traitFinder and Neo4j:

At the first trait finder tool design to build the graph by using Neo4j – for improve performance and tool portability traitFinder was include all Neo4j library , by this way traitFinder was able to run at any platform with valid x86 JVM .

TraitFinder tool have many modes when one of them is the live mode , which allow to user to crawl pinterest website while building graph using Neo4j, although this solution work it suffer from several problems :

* In case machine was suddenly shutdown graph DB files descriptor and locks can be damage and damaged Graph Db.
* It’s take more machine resources
* Need to implement recover mode in case of crash

For avoiding those problem we decided to create saving algorithm and validation – so when the graph is pass 5000 users trait’s finder will activate automatic saving , in the meanwhile traits finder will save users and there subject in simple csv format:

The format of the csv is very simple and contain users and subjects only saved in the following format:

| Users | Subjects |

| userx | subject 1 | subject 2 |…… |subjectN

For example – 2 users

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| user | subject |  |  |  |
| Oz | Food | Animals | Cars | Sports |
| Ran | Games | Cars |  |  |