Now it is time to focus a little more on the evaluation techniques for a recommender system.

As we anticipated before, the first category of the evaluation techniques is composed by the online ones,

while the second category is made up of the offline ones.

But let's now spend more time analyzing these two categories.

We now start analyzing different online evaluation techniques.

The first online evaluation technique we

consider is the direct user feedback. We can ask some real uses to define their level of satisfaction about the recommender system.

 For example, we can investigate the satisfaction using questionnaires.

Is this technique a good choice?

Well, yes, but there are two problems. First, the size of the sample should be meaningful and secondly, the opinion expressed by the user could not be reliable.

Another possibility is to monitor the online behavior of the users and to apply the so called A/B testing.

The core idea is to compare the behavior of users who receive recommendations, set A with the behavior of the users

who do not receive recommendations, set B. For example, if we consider an e-commerce, we would expect as a result that users who receive recommendation buy more products. A/B testing is a powerful method of evaluation, but difficult to set up.

In addition, its results may be difficult to interpret.

If users do not follow recommendations, what could be the problem? Is it because of a lack of

relevance or a lack of diversity?

Another way is to use the controlled experiments technique.

In a controlled experiment, a mockup application is made available to a group of potential users.

Users are asked to use for awhile the application, then they are asked to give their opinion about the recommendations received.

This method is less expensive than other methods, but there are some problems.

The application is not a real application, users are not real users and their opinions can be not reliable

since they aren't as motivated as the real user of a real online system.

For example, if I decide to volunteer for testing a movie recommender system and it suggests me to watch the movie Star Wars, probably I would watch 10 minutes of it and give a positive rating.

On the other hand, if I was a real user, I would think twice before I waste my time if I am not really convinced.

The last online evaluation technique is based on crowdsourcing.

It consists in asking people after an offer of some kind of compensation to answer an online questionnaire expressing their opinion about the mockup of an application.

This technique is powerful since there can be a large crowd of volunteers.

But how can we trust the opinions of the volunteers as maybe they were just interested into the compensation, and so now they are giving random answers?

Typically, these users are the less reliable, so we need very strong statistical test in order

to understand the reliability of their answers.