

# Content-based recommender system using linked data and structured tags

## Description

The goal is to build and measure the accuracy of a content-based (CB) recommender systems in the movies domain. The experiments will be performed on the RDF version of the 1 million ratings MovieLens dataset (ML1M) which is linked to DBpedia<sup>1</sup>.

Example 1:

<i>movieId   title   DBpedia resource</i> 1   Toy Story (1995)   <a href="http://dbpedia.org/resource/Toy_Story">http://dbpedia.org/resource/Toy_Story</a>
---------------------------------------------------------------------------------------------------------------------------------------------------------------

Di Noia et al. [1] built a content-based recommender by computing similarities of movies represented as vectors of selected properties (Vector space model VSM).

The goal of the project is to also consider tags given by users to the movies.

Example 2:

<i>userId   movieId   Tag   Timestamp</i> 17647::1::animation::1140064168 21374::247::based on a true story::1174957603 23172::5508::breaking the fourth wall::1219931735
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Tags are not available in ML1M, but in the 10M ratings MovieLens dataset (ML10M). However, movies in ML1M are a subset of the movies in ML10M and by joining both sources on the movie id it is possible to obtain the tags.

The strategy for extending the movie profile should be roughly as follows:

- Apply entity recognition techniques extract structured information from the tag and link it to DBpedia. This could for example achieved with DBpedia spotlight<sup>2</sup>;
- Use these new identified DBpedia resources of the tag to extend the movies' profiles;

Then the following similarity metrics should be evaluated individually:

- The originally proposed VSM metric of [1];
- One or more metrics proposed by the student which takes into account the structured and not structured information of the tags;

The student should then choose how to combine the metrics. Mean Absolute Error, Root Mean Squared Error, precision and recall are mandatory evaluation metrics.

## References

[1] Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker: Linked open data to support content-based recommender systems. I-SEMANTICS 2012:1-8

---

<sup>1</sup> <http://sisinflab.poliba.it/semanticweb/lod/recsys/datasets/>

<sup>2</sup> DBpedia Spotlight is a tool for automatically annotating mentions of DBpedia resources in text.  
<https://github.com/dbpedia-spotlight/dbpedia-spotlight>