Linked Open Data-enabled Recommender Systems: ESWC 2014 Challenge on Book Recommendation

Tommaso Di Noia[1], Iván Cantador[2], Vito Claudio Ostuni[1]

[1] Department of Electrical and Electronic Engineering, Politecnico di Bari, Italy
{tommaso.dinoia, vitoclaudio.ostuni}@poliba.it
[2] Department of Computer Science, Universidad Autónoma de Madrid, Spain
ivan.cantador@uam.es

Abstract. In this chapter we present a report of the ESWC 2014 Challenge on Linked Open Data-enabled Recommender Systems, which consisted of three tasks in the context of book recommendation: rating prediction in cold-start situations, top N recommendations from binary user feedback, and diversity in content-based recommendations. Participants were requested to address the tasks by means of recommendation appraoches that made use of Linked Open Data and semantic technologies. In the chapter we describe the challenge motivation, goals and tasks, summarize and compare the nine final participant recommendation approaches, and discuss their experimental results and lessons learned. Finally, we end with some conclusions and potential lines of future research.

1 Introduction

People generally need more and more advanced tools that go beyond those implementing the canonical search paradigm for seeking relevant information. A new search paradigm is emerging, where the user perspective is completely reversed: from finding to being found. Recommender systems may help to support this new perspective, because they have the effect of pushing relevant items (movies, videos, music albums, books, job offers, etc.), selected from a large space of possible options, to potentially interested users [12]. To achieve this objective, recommendation methods generally rely on data referring to three types of entities: users, items, and their relations.

Recent developments in the Semantic Web community offer novel strategies to represent data that may improve the current state of the art on recommender systems, in order to move towards a new generation of systems that fully understand the user preferences (tastes, interests, and goals), item features (e.g., domain attributes, categories, and related concepts), and contextual signals (e.g., time, location, mood, and social company) they deal with.

More and more semantic data are published following the Linked Open Data principles [1,2] (LOD), which enable to set up links between entities in different knowledge sources, by connecting information in a single global data space: the Web of Data [4]. Today, the Web of Data includes different types of knowledge represented in a homogeneous form, both sedimentary (encyclopedic, cultural, linguistic, common-sense) and real-time (news, data streams, etc.) types.

This knowledge might be useful to interlink diverse information about users, items, and their relations, and implement reasoning mechanisms that can support and improve the recommendation process. Hence, the primary goal of the ESWC 2014 Challenge on Linked Open Data-enabled Recommender Systems[3] was twofold. On the one hand, we wanted to create a link between the Semantic Web and the Recommender Systems communities. On the other hand, we aimed to show how Linked Open Data and semantic technologies can boost the creation of a new breed of knowledge-enabled and content-based recommender systems. In particular, we focused on the particular scenario of book recommendation, and stated three tasks, namely rating prediction in cold-start situations, top N recommendations from binary user feedback, and diversity in content-based recommendations. Participants were requested to address the tasks by means of recommendation approaches that made use of Linked Data and semantic technologies.

In the remainder of the chapter, we describe the challenge dataset (Section 2), tasks (Section 3), and evaluation protocol (Section 4), summarize and compare the nine final participant recommendation approaches (Section 5), and present the obtained experimental results (Section 6) and derived conclusions (Section 7) in the challenge.

2 Challenge Dataset

The challenge tasks were conducted on the DBbook dataset, which was built upon the LibraryThing dataset[4], and relies on user preferences (ratings in the [0, 5] integer interval) for books retrieved from the Web. As explained in [6], the books available in the original rating dataset were mapped to their corresponding DBpedia URIs, allowing participants extract semantic features from DBpedia [1] and other Linked Open Data repositories, which could be exploited by their recommendation approaches in the challenge tasks.

The final mapping contained 8170 DBpedia URIs. For each task, the dataset was split into a training set and a test set. In the former, user ratings were provided to build the recommender systems, while in the latter ratings were removed, since they were used in an eventual evaluation stage.

---

1 Linking Open Data, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

2 Linked Data, http://linkeddata.org

3 ESWC 2014 Challenge on Linked Open Data-enabled Recommender Systems, http://challenges.2014.eswc-conferences.org/index.php/RecSys

4 LibraryThing dataset, http://www.macle.nl/tud/LT

For Task 1 – rating prediction in cold-start situations –, the training dataset contained the numeric values of the user ratings. A total of 75559 ratings from 6181 users for 6166 distinct books were provided as training data.

For Task 2 – top N recommendations from binary user feedback – and Task 3 – diversity in content-based recommendations –, in contrast, the ratings were given in a binary scale, where 1 meant that a book was relevant for a user, and 0 otherwise. In this case, a total of 72372 ratings from 6181 users for 6733 distinct books were provided as training data.

3 Challenge Tasks

3.1 Rating Prediction in Cold-start Situations

This task dealt with the rating prediction problem, in which a system is requested to estimate the value of unknown numeric ratings that a target user would assign to available items, indicating whether she likes or dislikes them.

In order to favor the proposal of content-based, LOD-enabled recommendation approaches, and limit the use of collaborative filtering strategies, this task aimed to predict ratings in cold-start situations, that is, predicting ratings for users with a few past ratings, and predicting ratings of items rated by a few users.

Participants were asked to exploit the ratings provided as training data, in addition to semantic features freely chosen and extracted from Linked Data repositories, in order to estimate missing ratings of a test set. Estimated ratings were submitted in the format `userID \t itemID \t rating`.

Recommendation approaches were evaluated on the test set *Te* by means of the well known Root Mean Square Error (RMSE), which measures the differences between actual ratings $r_{u,i}$ and predicted ratings $p_{u,i}$ of users u and items i :

$$RMSE = \sqrt{\frac{1}{|Te|} \sum_{(u,i,r_{u,i}) \in Te} \left(p_{u,i} - r_{u,i}\right)^2}$$

3.2 Top N Recommendations from Binary User Feedback

This task dealt with the top N recommendations problem, in which a system is requested to find and recommend a limited set of N items that best match a user profile, instead of correctly predicting the ratings for all available items.

Similarly to Task 1, in order to favor the proposal of content-based, LOD-enabled recommendation approaches, and limit the use of collaborative filtering strategies, this task aimed to generate ranked lists of items, in cold-start situations, for which no graded ratings were available, but binary ones.

Participants were asked to complete the user-item pairs in the test set by adding the correspondent relevance score according to the format `userID \t itemID \t score`. These relevance scores were used by an evaluation service to form a Top 5 item recommendation list for each user. This means that for each user, only items in the test set were considered to form the top 5 recommendation list. The evaluation metric for this task was the F-measure@5:

$$F\text{-}measure@5 = Precision@5 \cdot Recall@5$$

where

$$Precision@5 = \frac{1}{|U|} Precision_u@5 \qquad Precision_u@5 = \frac{1}{5}\sum_{i=1}^{5} rel_{i,u}$$

$$Recall@5 = \frac{1}{|U|} Recall_u@5 \qquad Recall_u@5 = \frac{1}{R_u}\sum_{i=1}^{5} rel_{i,u}$$

being U the set of users, $rel_{i,u}$ the binary relevance value of the item with the i-th highest predicted rating for user u, and R_u the set of u's relevant items.

3.3 Diversity in Content-based Recommendations

A very interesting aspect of content-based recommender systems, and then of LOD-enabled ones, is giving the possibility to evaluate the diversity of recommended items in a straight way. This is a very popular topic in content-based recommendations, which usually suffer from over-specialization.

In this task, the evaluation was made by considering a combination of both accuracy of the recommendation list, and the diversity of items belonging to it. Given the domain of books and the challenge focus on Linked Data, we considered diversity with respect to two properties: `http://dbpedia.org/ontology/author` and `http://www.w3.org/2004/02/skos/core#subject`.

Participants were asked to submit a top 20 recommendations list for each user. The submitted lists had to be computed considering all unrated items of each user, and selecting the 20 items with highest predicted ratings. Similarly to Task 2, in this case, the line format of the submission file was `userID \t itemID \t score`.

In this task, the evaluation metric was a combination of accuracy and diversity. In particular, F-measure@20 was used for measuring accuracy, and Intra-List Diversity ILD@20 [15] for diversity. ILD@20 was defined based on ILS@20:

$$ILS@20 = \frac{1}{|U|} ILS_u@20 \qquad ILS_u@20 = \frac{1}{2}\sum_{i \in L_u^{20}} \sum_{j \in L_u^{20}} sim(i,j)$$

where L_u^{20} is the list of 20 items recommended to user u with highest predicted ratings, and $sim(i,j) \in [-1,+1]$ is a content-based similarity between items i and j. The final ranking is computed as follows. First, F-measure@20 and ILD@20 alone were used to form two initial rankings. Then, a final ranking was produced by considering each participant's score as the mean of her rank positions in the two initial rankings.

4 Evaluation Protocol

For Task 1, the training and test sets were available at the following URLs:

- `root_url/DBbook_train_ratings.zip`
- `root_url/task1_useritem_evaluation_data.tsv.zip`

where `root_url` has to be replaced by
`http://sisinflab.poliba.it/semanticweb/lod/recsys/2014challenge`.

For Task 2 and 3, the training and test sets were available at the following URLs:

- `root_url/DBbook_train_binary.zip`
- `root_url/task2_useritem_evaluation_data.tsv.zip`

The training sets were provided as tab-separated value files, in which each line had the format `userID \t itemID \t rating`, and the test sets were also provided as tab-separated value files, but having the line format `userID \t itemID`.

To evaluate their approaches, participants were asked to submit a file containing the rating predictions or recommendations to an evaluation system using the web form available at `http://193.204.59.20:8181/eswc2014lodrecsys/`.

Alternatively, participants could also submit their results using a Java client available at `root_url/lodrecsys2014challenge_evaluation.jar`, by launching the following command:

```
java -jar lodrecsys2014challenge_evaluation.jar
                                    TaskNumber GroupID filePath
```

5 Participant Approaches

During the challenge, 14 approaches participated in Task 1, 24 approaches participated in Task 2, and 12 approaches participated in Task 3. Finally, 9 distinct approaches completed the challenge, by taking part in one or more of the challenge tasks, and being described in a paper accepted by three program committee members in a blind review process. In the following, we describe and compare the final participant approaches, and highlight their lessons learned.

SemWex1. Hybrid Recommending Exploiting Multiple DBpedia Language Editions [11]
By Ladislav Peska, and Peter Vojtas (Charles University in Prague, Czech Republic)

This is a hybrid recommendation approach that is based on a content-based extension of the matrix factorization method for collaborative filtering. The approach incorporates item features into the matrix to factorize, and generates item latent factor vectors from the latent factors of the items features.

A total of 100 features and 35K feature-value pairs were obtained from RDF data associated to books and writers in DBpedia. The final features were generated from different transformations of the original ones:

- Discretizing numeric feature values;
- Grouping equivalent features, e.g., the `precededBy` and `notableWork` properties were unified into a `similarWork` property;
- Manually annotating authors and genres with metadata, such as serious or fun literature, male or female target audience, and literary genre clusters;
- Extending the items categories by 3 levels of super categories (obtained through the `skos:broader` property);
- Filtering low informative features.

To increase the diversity of generated recommendations, the approach applies a heuristic that selects the book (item) with the highest rating of each author from the top N recommendations list.

One of the lessons learned with this approach is that super categories were often too general to provide valuable information.

The approach took the 4th position in Task 1, the 4th position in Task 2, and the 4th position in Task 3.

helloWorld. **A Hybrid Multi-Strategy Recommender System Using Linked Open Data** [13]
By Petar Ristoski, Eneldo Loza Mencía, and Heiko Paulheim (University of Mannheim, Germany)

This is a hybrid recommendation approach that uses stack regression and rank aggregation techniques to combine recommendations from several methods:

- Content-based recommendation methods that use different sets of item features;
- User- and item-based heuristic collaborative filtering methods using the cosine similarity function;
- Popularity-based recommendation method that returns global item popularity scores, which are independent of the target user, and are computed with the books average scores in Amazon and the number of ingoing/outgoing links with Wikipedia and other datasets.

For each book, the considered features were:

- the direct (YAGO) types of the book,
- the direct categories of the book,
- the super categories of the book categories,
- all books written by the book author,
- the genres of the book and author's books,
- the writers who influenced or were influenced by the book author, and
- a bag of keywords extracted from the Wikipedia abstract of the book.

These features were extracted from DBpedia, the RDF Book Mashup dataset (`http://datahub.io/dataset/rdf-book-mashup`), the British Library

Bibliography (`http://bnb.data.bl.uk`), and the DBTropes catalogue (`http://dbtropes.org`).

To increase the diversity of generated recommendations, the approach applies a heuristic that filters out books whose authors and categories already appear as metadata of books in the top N recommendations list.

The main lessons learned with this approach are that item popularity allowed increasing accuracy, and hybridization allowed increasing diversity.

The approach took the 1st position in Task 1, the 2nd position in Task 2, and the 1st position in Task 3.

IDEAL. Exploring Semantic Features for Producing Top N Recommendation Lists from Binary User Feedback [2]
By Nicholas Ampazis, and Theodoros Emmanouilidis (University of the Aegean, Greece)

This is a content-based recommendation approach that uses a feature vector representation for users and items. The approach computes similarities between the items liked by the user in the past and the reminder items, to suggest those with highest similarities. Several similarities were tested, namely the cosine, Euclidean distance, and Tanimoto similarities. The best performing was Tanimoto similarity, which is computed as the ratio between the size of the intersection of two vectors by the size of their union.

The features used for testing the approach were the book authors and categories, extracted from DBpedia.

In order to account for recommendation diversity, the approach generates an initial list with the top 50 recommended items. Next, it measures the pair-wise similarities of the 50 items. Finally, it selects for the top 20 recommendation list, those items that more frequently exhibit the lowest similarities with the other items.

A lesson learned with this approach is that even simple content-based similarities and diversification strategies may obtain good recommendation results.

The approach took the 5th position in Task 3.

UNIBA. Aggregation Strategies for Linked Open Data-enabled Recommender Systems [3]
By Pierpaolo Basile, Cataldo Musto, Marco De Gemmis, Pasquale Lops, Fedelucio Narducci, and Giovanni Semeraro (University of Bari, Italy).

This is a hybrid recommendation approach that consists of a linear combination of recommendations from some (depending on the challenge task) of the following methods:

- Popularity-based recommendation method, in which item popularity is computed as the ratio between the number of positive ratings perceived by the item and the total number of ratings (positive and negative) of the item;
- Enhanced Vector Space Model (eSVM) with negation, which is a content-based method based on an incremental dimensionality reduction technique;

- Page Rank with priors (PR), in which a personalization vector may be used for assigning different initial weighs to certain nodes liked/disliked by the user;
- Random Forest (RF), which is an ensemble classification method that consists of several decision trees built with different training items and features;
- Logistic Regression (LR), which is a classification method that builds a linear model based on a transformed target variable.

The above methods used a combination of the following features:

- Keywords extracted from Wikipedia descriptions and DBpedia abstracts of the books;
- DBpedia concepts appearing in the book description and abstract;
- DBpedia properties of the books, in particular, the 10 most frequent properties (http://dbpedia.org/ removed for brevity); `ontology/wikiPageWikiLink`, `http://purl.org/dc/terms/subject`, `property/genre`, `property/publisher`, `ontology/author`, `property/followedBy`, `property/precededBy`, `property/series`, `property/dewey`, `ontology/nonFictionSubject`.

To account for recommendation diversity, the approach applies the PageRank algorithm with different priors:

- 80% of the initial weight evenly distributed to those nodes that correspond to books liked by the user (0 for those disliked by the user);
- 10% of the initial weight evenly distributed to those nodes that do not correspond to books;
- 10% of the initial weights proportionally distributed to those nodes that correspond to books not rated by the user; the weight distribution is done according to a diversity score, which is an average of similarity and novelty metrics.

The main lessons learn with this approach were:

- Very simple methods based on SVM and probabilistic models are capable of obtain accurate recommendation;
- The usefulness of semantic data was evident in recommendation methods based on classifiers;
- The application of a graph-based ranking algorithm on a semantic network built with DBpedia concepts and properties allowed diversifying recommendation lists;

The approach took the 2nd position in Task 1, the 1st position in Task 2, and shared the 2nd position with UIMR-NUIGalway in Task 3.

UIMR-NUIGalway. SemStim at the LOD-RecSys 2014 Challenge [7]
By Benjamin Heitmann, and Conor Hayes (National University of Ireland - Galway, Ireland)

This is a graph-based recommendation algorithm based on Constrained Spreading Activation (CSA), which uses generic constraint functions for the activation, restart, and termination of the weight propagation process. The approach is executed on a semantic graph where source nodes are associated to concepts (books) liked by the user, and target nodes are associated to books not rated by the user. The reminder nodes are associated to concepts associated to book categories, properties, and Wikipedia disambiguation and redirect links.

The approach was only tested in the diversity task. In the cases in which the approach generated recommendation lists with less than 20 items, randomly selected items were aggregated to the lists.

Similarly to other approaches that made use of graph-based algorithms, this approach performed well when providing diversity in recommendation lists.

The approach shared the 2nd position with UNIBA in Task 3.

UniMannheim. Popular Books and Linked Data: Some Results for the ESWC'14 RecSys Challenge [14]
By Michael Schuhmacher, and Christian Meilicke (University of Mannheim, Germany)

This team tested two approaches. The first approach was a naive, non personalized recommendation approach based on the items popularity computed on the training dataset according to the top rated items, and without making use of any external knowledge. The second approach was a hybrid method composed of a Naive Bayes classifier that was built with item features on user neighbor clusters. In this approach, other classifiers (Support Vector Machines, Linear Regression, and Decision Trees) were also tested.

The used features were:

- DBpedia properties: genre (`dbo:literaryGenre`), Wikipedia subjects (`dcterms:subject`), YAGO types (`rdf:type`), authors (`dbo:author`, `dbo:writer`), book series (`dbo:series`), publisher (`dbo:publisher`).
- DBpedia categories: the Wikipedia categories of each book plus their immediate (1 level) super categories, obtained via the `skos:broader` and the `dbo:wikiPageWikiLink` (Wikipedia links) properties.
- 30 manually defined categories (e.g., *science fiction*, *fantasy*, *horror*, and *philosophy*), each of them assigned to a book if certain pattern (usually the category name) was found in the book abstract (`dbo:abstract`), genre (`dbo:literaryGenre`, `dbp:genre`), or subject (`dcterms:subject`).
- Expanded categories, selected based on the highest Dice similarity between the values of `dcterms:subject`, `dbo:literaryGenre`, and `dbp:genre` properties, e.g., `Literary_history` and `History_of_literature`.

The approach was only tested on the top N recommendations for binary user feedback task. Since it did not perform well isolated, their results were combined with a user-based collaborative filtering method, which generated user neighbor clusters on which the classifier was executed.

The main lessons learn with this approach were:

- The popularity-based baseline achieved competitive recommendation results;
- The user aggregation methods showed a significant influence on the overall performance;
- There was a marginal contribution from each feature to the overall performance, especially from the expanded categories.

The approach took the 3rd position in Task 2.

VUAgroup. Semantic Pattern-based Recommender [9]
By Valentina Maccatrozzo, Davide Ceolin, Lora Aroyo, and Paul Groth (VU University Amsterdam, The Netherlands)

This approach extracts semantic patterns from DBpedia, and exploits such patterns for user modeling and recommendation purposes. For instance, a user who liked books written by Jack Kerouac, may be interested in a book written by Ernest Hemming, since the former *influenced* the latter.

The approach uses the patterns to automatically (via SPARQL queries) build semantic paths between the user's rated books and other books. The user's ratings for the unrated items are computed by means of personalized positive/negative scores assigned to patterns and books.

The approach was only tested in the top N recommendations for binary user feedback task, achieving reasonable performance results without any setup and exploration of alternative configurations or adaptations.

The approach took the 5th position in Task 2.

LDOS. Increasing Top 20 Diversity through Recommendation Post-processing [8]
By Matevz Kunaver, Tomaz Pozrl, Stefan Dobravec, Andrej Kosir, and Uros Droftina (University of Ljubljana, Slovenia)

This approach is a rule-based recommendation method that represent each book with some of 17 DBpedia features (author, year of publishing, type, etc.), and Dublin Core categories; each item having on average 5 different categories.

The approach applies a post-processing method to generated recommendations in order to increase diversity. Specifically, it applies a *leave one out* technique measuring the ILD@19 metric on the 40 top ranked items. The approach sorts the recommendation list in ascending order by the ILD@19 value, and excludes the item whose absence has the smallest impact on the diversity of the recommendation list. This process is iteratively done until discarding 20 items.

The main lesson learned with this approach is that the followed diversification strategy, which aims to optimize ILD (the evaluation metric), indeed increases diversity, but entails a high loss of precision and recall.

The approach shared the 5th position with IDEAL in Task 3.

UniAndes1. Hybrid model rating prediction with Linked Open Data for Recommender Systems [10]
By Andrés Moreno, Christian Ariza-Porras, Paula Lago, Claudia Lucía Jiménez-Guarín, Harold Castro, and Michel Riveill (Universidad de los Andes, Colombia)

This approach is a switched hybrid recommendation method that maintains different models in parallel, and reports to the user the rating predictions and recommendations of the model with highest confidence. Specifically, it uses a collaborative filtering strategy (the SVD++ matrix factorization algorithm) when enough ratings are present, and uses a content-based recommendation strategy otherwise.

Additionally, the approach clusters the feature values to reduce the dimensionality of the user and item profiles. For the content-based recommendation method, the used features were the book authors, categories, literary genres, and the subject property.

The approach took the 3rd position in Task 1.

Table 1 depicts a comparison of the challenge final participant approaches. For each approach, we show:

- The type of **hybridization technique** used (if any), based on Burke's hybrid recommender system taxonomy [5]: *feature combination* (putting features from different recommendation data sources into a single method), *feature augmentation* (using output from a recommendation method as input to another), *mixed hybridization* (jointly presenting recommendations from several methods), *weighting hybridization* (combining the recommendation scores from several methods), and *switching hybridization* (using some criterion to switch between recommendation methods).
- The type of the underlying **recommendation method(s)**, e.g., *content-based*, *collaborative filtering*, and *popularity-based*.
- The nature of the content- and semantic-based **features** exploited by the recommendation methods, such as *book attributes (title, author, genres, etc.)*, *Wikipedia categories*, and *text keywords*.
- The type of **diversification strategy** applied (if any), namely *pre-processing* if the approach itself is modeled to provide diversity in generated recommendations, and *post-processing* if the approach makes use of a strategy to diversify generated recommendations.

It can be seen that 5 out of the 9 approaches used some type of hybridization technique, without a predominant one existing among the participants. It seems that those techniques that combine recommendations from different methods (*mixed* and *weighting*) performed better than the others. Regarding the recommendation methods, we note that exploiting item popularity information helped to increase accuracy in cold-start situations (Task 1), and graph-based approaches achieved both high accuracy and diversity (Task 3). Moreover, as stated by some of the authors, the use of the books Wikipedia categories and super-categories was not a relevant feature to improve recommendations. In contrast, extending user and item profiles by means of keywords extracted from the book abstracts and descriptions may help dealing with binary user feedback (Task 2). Finally, we note that all except one approach applied a post-processing diversification strategy. In this context, those strategies aimed to avoid repetitions of book authors and genres within the recommendation lists achieved the best results in Task 3; optimizing ILD of recommendation lists alone was not a good solution to the task, since it did not account for the loss of accuracy.

Table 1. Comparison of the challenge final participant approaches. The superscript * indicates that the participant shares rank position with other participant(s).

Approach	Hybridization technique	Recommendation methods	Features	Diversification strategy	Ranking		
					T1	T2	T3
SemWex1 [11]	Feature combination	Content-based extension of matrix factorization	Attributes Extended categories Manual metadata	Post-processing (non repeated authors)	4	4	4
helloWorld [13]	Mixed	Content-based User-based collaborative filtering Item-based collaborative filtering Popularity-based	Attributes Extended categories Abstract keywords Popularity	Post-processing (non repeated authors and genres)	1	2	1
IDEAL [2]	-	Content-based	Authors Categories	Post-processing (non similar books)			5*
UNIBA [3]	Weighting	Content-based Graph-based (PageRank) Machine learning (RF, LR) Popularity-based	Most popular attributes Description keywords Description concepts	Pre-processing (diversity scores on graph nodes)	2	1	2*
UIMR-NUIGalway [7]	-	Graph-based (CSA)	Attributes Categories Disambiguation links Redirect links	-			2*
UniMannheim [14]	Feature augmentation	User-based collaborative filtering Machine learning (Naive Bayes) Popularity-based	Attributes Extended categories	-		3	
VUAgroup [9]	-	Semantic pattern-based	Attributes	-		5	
LDOS [8]	-	Rule-based	Attributes Categories	Post-processing (filtering books via relative ILD values)			5*
UniAndes1 [10]	Switching	Content-based Matrix factorization	Authors Literary genres Categories	-	3		

6 Results of the Participant Approaches

Overall, 14 teams participated in Task 1, 24 in Task 2 and 12 in Task 3. Among them, 15 submitted a paper describing the approach they adopted for competing in the challenge, and 9 of them were selected by the program committee and chairs as final participants. Those final participants are the ones just presented in Section 5, and were the ones who were considered in computing the final rankings to determine the winner for each task. In the following, we discuss the results achieved by those final participants in the three tasks. Such results are shown in Table 2.

6.1 Results of the Rating Prediction in Cold-start Situations Task

The best performing participant in this task was *helloWorld* who achieved the lowest RMSE score. *UNIBA* and *UniAndes1* ranked second and third, respectively. As we can note, the difference between *UNIBA* and *UniAndes1* is limited, while the gap between those groups and *SemWex1*, who ranked fourth, is quite marked. Looking at Table 1 we can see that both the two best performing approaches used hybridization strategies based on recommendation combinations.

6.2 Results of the Top N Recommendations from Binary User Feedback Task

In this task, instead, the best performing approach was the one adopted by *UNIBA* which achieved a F-measure of 0.57151. *helloWorld* ranked at the second position with a score only 3×10^{-5} lower than the highest one. Then, *UniMannheim* ranked third and the other participants to follow. Also in this case, the two best performing approaches were the ones based on an ensemble of several different recommendation methods.

6.3 Results of the Diversity in Content-based Recommendations Task

In this task, the best performing participant was again *helloWorld*, which obtained the best ILD and F-measure values of 0.04816 and 0.4846, respectively. At the second position there were two participants: *UNIBA* and *UIMR-NUIGalway*. The first got a F-measure value of 0.04813, and ranked fourth in the ILD ranking with a ILD score of 0.47169. While the second ranked third in the F-measure ranking with a score of 0.04129 and third in the ILD ranking with a score of 0.47603.

Looking at the individual metrics alone, the best approaches in terms of accuracy were the ones proposed by *helloWorld* and *UNIBA* in accordance also with their results in Task 2. The differences in scale between the F-measure scores in Task 2 and Task 3 are due to the different evaluation protocols. Particularly, in Task 2, each user recommendation list had to be generated considering only test items, while in Task 3, considering all items except the ones in the user training data.

Regarding diversity, the highest scores were achieved by *helloWorld* and *SemWex1* who both adopted a post-processing diversification strategy aimed to avoid repetitions of book authors and genres within the recommendation lists.

Table 2. Participant Results. The superscript * indicates that the participant shares rank position with other participants.

Approach	Task 1		Task 2		Task 3		
	Ranking	RMSE	Ranking	F-measure@5	Ranking	F-measure@20	ILD@20
SemWex1 [11]	4	0.93686	4	0.55396	4	0.01989	0.48025
helloWorld [13]	1	**0.86322**	2	0.57148	1	**0.04816**	**0.48460**
IDEAL [2]	-	-	-	0.53312	5*	0.03479	0.44471
UNIBA [3]	2	0.87422	1	**0.57151**	2*	0.04813	0.47169
UIMR-NUIGalway [7]	-	-	-	-	2*	0.04129	0.47603
UniMannheim [14]	-		3	0.56070	-	-	-
VUAgroup [9]	-	-	5	0.51622	-	-	-
LDOS [8]	-	-	-	-	5*	0.03085	0.45489
UniAndes1 [10]	3	0.87871	-	-		-	-

7 Conclusions from the Challenge

The Linked Open Data-enabled Recommender Systems Challenge at ESWC 2014 was among the first attempts to bring together the two communities of Recommender Systems and Semantic Web. The high number of participants and the quality of results obtained by the different teams show that there is an increasing interest in the topic, as well as that recommender systems have been recognized as a potential killer application for the exploitation of Linked Open Data.

What emerges by looking at the different approaches proposed by the participants is that the best performing techniques, with respect to the provided dataset, for rating prediction and top-N recommendations use an ensemble of several different recommendation methods, while post-processing results are very effective in increasing the diversity of the recommendation list.

We think that there is still room to better exploit both the semantics encoded in LOD datasets and the connections among items to improve the quality of recommendation results both in terms of accuracy and in terms of diversity, in the future, novelty and serendipity. We also believe that contextual semantic data, e.g., coming from data streams, can be easily integrated with the information currently available in LOD datasets to build a new wave of context-aware recommender systems.

Acknowledgements

We thank all participants for their interest in the challenge, their submissions, and presentations and discussion during the conference. We also thank the program committee members for their valuable reviews of submissions, and Valentina Presutti and Milan Stankovic for their help with the organization of the challenge.

References

1. Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z. (2007) DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, pp. 722-735.
2. Ampazis, N., Emmanouilidis, T. (2014) Exploring Semantic Features for Producing Top N Recommendation Lists from Binary User Feedback. Semantic Web Evaluation Challenges 2014. Springer.
3. Basile, P., Musto, C., De Gemmis, M., Lops, P., Narducci, F., Semeraro, G. (2014) Aggregation Strategies for Linked Open Data-enabled Recommender Systems. Semantic Web Evaluation Challenges 2014. Springer.
4. Bizer, C., Heath, T., Berners-Lee, T. (2009) Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5(3), pp. 1-22.
5. Burke, R. D. (2002) Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12(4), pp. 331-370.
6. Di Noia, T., Mirizzi, R., Ostuni, V. C., Romito, D., Zanker, M. (2012) Linked Open Data to Support Content-based Recommender Systems.In: Proceedings of the 8th International Conference on Semantic Systems, pp. 1-8.
7. Heitmann, B., Hayes, C. (2014) SemStim at the LOD-RecSys 2014 Challenge. Semantic Web Evaluation Challenges 2014. Springer.
8. Kunaver, M., Pozrl, T., Dobravec, S., Kosir, A., Droftina, U. (2014) Increasing Top 20 Diversity through Recommendation Post-processing. Semantic Web Evaluation Challenges 2014. Springer.
9. Maccatrozzo, V., Ceolin, D., Aroyo, L., Groth, P. (2014) Semantic Pattern-based Recommender. Semantic Web Evaluation Challenges 2014. Springer.
10. Moreno, A., Ariza-Porras, C., Lago, P., Jiménez-Guarín, C. L., Castro, H., Riveill, M. (2014) Hybrid Model Rating Prediction with Linked Open Data for Recommender Systems. Semantic Web Evaluation Challenges 2014. Springer.
11. Peska, L., Vojtas, P. (2014) Hybrid Recommending Exploiting Multiple DBpedia Language Editions. Semantic Web Evaluation Challenges 2014. Springer.
12. Ricci, F., Rokach, L., Shapira, B., Kantor, P. B. (2011) Recommender Systems Handbook. Springer.
13. Ristoski, P., Mencía, E. L., Paulheim, H. (2014) A Hybrid Multi-Strategy Recommender System Using Linked Open Data. Semantic Web Evaluation Challenges 2014. Springer.
14. Schuhmacher, M., Meilicke, C. (2014) Popular Books and Linked Data: Some Results for the ESWC'14 RecSys Challenge. Semantic Web Evaluation Challenges 2014. Springer.
15. Ziegler, C.-N., McNee, S. M., Konstan, J. A., Lausen, G. (2005) Improving Recommendation Lists through Topic Diversification. In: Proceedings of the 14th International Conference on World Wide Web, pp. 22-32.