Schema-aware feature selection in Linked Data-based recommender systems (Extended Abstract)*

Corrado Magarelli[1], Azzurra Ragone[1], Paolo Tomeo[2], Tommaso Di Noia[2], Matteo Palmonari[1], Andrea Maurino[1], Eugenio Di Sciascio[2]

[1]University of Milan Bicocca, P.zza Dell'Ateneo Nuovo, 1, 20126 Milano, Italy
{corrado.magarelli,azzurra.ragone,matteo.palmonari,andrea.maurino}@unimib.it
[2]Polytechnic University of Bari, Via Orabona, 4, 70125 Bari, Italy
{paolo.tomeo,tommaso.dinoia,eugenio.disciascio}@poliba.it

Abstract. Semantics-aware recommendation engines have emerged as a new family of systems able to exploit the semantics encoded in unstructured and structured information sources to provide better results in terms of accuracy, diversity and novelty as well as to foster the provisioning of new services such as explanation. In the rising of these new recommender systems, an important role has been played by Linked Data (LD). However, as Linked Data is often very rich and contains many information that may result irrelevant and noisy, an initial step of feature selection may be required in order to select the most meaningful portion of the original dataset. Many approaches have been proposed in the literature for feature selection that exploit different statistical dimensions of the original data. In this paper we investigate the role of the semantics encoded in an ontological hierarchy via schema-summarization when exploited to select the most relevant properties for a recommendation task.

1 Introduction

In the last years we have witnessed a flowering of semantics-aware solutions for Recommender Systems (RSs) exploiting information held in knowledge graphs, as the ones available in the Linked Data (LD) Cloud. Several approaches using LD to build RSs have been proposed in the literature. However, almost no one tackles the issue of *automatically* selecting the best subset of LD-based features. Usually, the feature-selection process is done manually by choosing the properties more "suitable" for the scenario taken into account. For example, in a scenario related to movies, properties as `dbo:starring` or `dbo:director` look more relevant than `dbo:releaseDate` or `dbo:distributor`. As well as for the music domain, properties as `dbo:genre` and `dbo:writer` look more important than `dbo:producer` or `dbo:recordedIn`. However, without an automatic feature selection process, the human intervention is required every time a new domain

* An extended version of this paper has been published in [7]

is chosen, while it could be good to have a general way to select properties regardless of the domain. In machine learning tasks there is the need to perform a selection of features and this could not be straightforward when attributes are embedded in a knowledge graph. In many graph-based recommendation systems the knowledge exploration starts from the data and goes on following the relations between entities, without taking into account the knowledge lying in the ontology and then in its class hierarchy. In this paper we investigate how ontological schema summarization could be used as a feature selection technique for LD-based recommender systems when features are represented by RDF properties and compare the results with other "classical" techniques for feature selection.

2 Feature selection and recommender systems

When dealing with recommender systems, a relevant task is to determine the impact of a particular feature selection technique on the behavior of the underlying algorithms. Indeed, some techniques can improve the accuracy of the recommendation, some improves the diversity while others can provide a good trade-off between diversity and accuracy. Among all the different feature selection techniques available in the literature, in our experimental setting, we initially selected *Information Gain*, *Information Gain Ratio*, *Chi-squared test* and *Principal Component Analysis* as their computation can be adapted to categorical features, as the LD ones. Then, the features selected from each technique have been used as input for two recommendation algorithms based on graph-kernels [6]: entity-based and path-based. Experimental results showed *Information Gain* as the best performing technique[1]. Information Gain (IG) is defined as the expected reduction in entropy occurring when a feature is *present* versus when it is *absent*. For a feature f_i, IG is computed as [5]:

$$IG(f_i) = E(I) - \sum_{v \in dom(f_i)} \frac{|I_v|}{|I|} * E(I_v)$$

where $E(I)$ is the value of the entropy of the data, I_v is the number of items in which the feature f_i (e.g. *starring for movies*) has a value equal to v (e.g. *Al Pacino* in the movie domain), and $E(I_v)$ is the entropy computed on data where the feature f_i assumes value v. The IG of a feature f_i is higher as the lower is the value of the entropy $E(I_v)$. Features are ranked according to their IG and the top-k ones are returned.

Schema summarization for feature selection. Linked Data summarization is the process of extracting a summary of an input linked data set, such that this summary is smaller (in size) than the input data, but retains information useful for certain tasks. Relevance-oriented summaries capture subsets of the input data sets and/or ontologies. These subsets are estimated to be more relevant

[1] The interested reader may refer to `https://github.com/sisinflab/SAC2017/FeatureSelection` for results obtained with other feature selection techniques

for the users according to multidimensional relevance criteria [10]. Vocabulary-oriented summaries describe the usage of vocabularies, e.g., ontologies, used in a dataset. These summaries are usually defined so as to be complete, i.e., to provide information about every element of the vocabulary/ontology used in the data set [9]. Vocabulary-oriented summaries that provide complete descriptions of vocabulary usage may support feature selection by providing relevant information about every possible feature, i.e., property, in the data set.

In this paper we use summaries produced by a vocabulary-oriented summarization framework named ABSTAT[2]. It takes a linked data set and - when available - one or more ontologies used in this data set as input, and returns a summary. The summary consists in a set of *patterns* having the form $\langle C, P, D\rangle$, with C and D being types, i.e., concepts or datatypes, and P being an RDF property. We refer to C and D as source and target types, respectively. Each pattern $\langle C, P, D\rangle$ tells that there exist some instance of type C linked to some instance of type D through the property P. For example, a pattern $\langle$`dbo:Film`, `dbo:starring`, `dbo:Actor`$\rangle$ tells that there are instances of `dbo:Film` linked to instances of type `dbo:Actor` through the property `dbo:starring` in the data set. The summary is complete for relational assertions in an RDF data set, i.e., assertions about individuals: for every relational assertion $\langle x, p, y\rangle$ that exists in the data set, at least one pattern is generated, i.e., every such assertion is represented by at least one pattern. The generation of these patterns is based on explicit typing assertions, e.g., $\langle$`dbr:Tom_Cruise`, `rdf:type`, `dbo:Actor`$\rangle$ or on implicit typing assertions (for literals), e.g., `1962-01-01`$^{\texttt{xsd:date}}$ extracted from the dataset. Differently from other approaches that also extract vocabulary-based patterns from linked data sets [4, 3], ABSTAT applies a pattern minimalization technique leveraging the relations between types defined in the ontologies (when the ontologies are used in the summarization process). Additional information provided in summaries and of major importance for feature selection is pattern *frequency*, which counts the occurrences of patterns in the data set. For example, $\langle$`dbo:Film`, `dbo:starring`, `dbo:Actor`$\rangle$[10662] tells that 10662 instances of `dbo:Film` are linked to instances of type `dbo:Actor` through the property `dbo:starring` in the data set[3].

3 Evaluation

For evaluating the quality of a recommendation algorithm, given a particular feature selection technique, we use four metrics, as each one of them measures a different dimension in the final result. To evaluate recommendation **accuracy**, we use Precision and Mean Reciprocal Rank (MRR). While *Precision@N* is a metric denoting the fraction of relevant items in the top-N recommendations,

[2] ABSTAT summaries for several datasets can be explored at `http://abstat.disco.unimib.it:8880/`

[3] For more details about the summarization process, the impact of minimalization on the size of extracted summaries, the use of ABSTAT summaries to support data set understanding, and the services through which summaries are accessible via web interfaces we refer to [9].

Entity-based Graph kernel	Top-K features	Precision@10	MRR@10	itemCov@10	aggrEntropy@10
IG	**5**	0.02327	0.15578	0.54262	8.96
	10	0.01734	0.13599	0.90658	10.24
	15	0.02055	0.14685	0.91989	10.19
ABSTAT	**5**	0.02035	0.14694	0.54953	9.12
	10	0.01651	0.13705	0.64346	9.42
	15	**0.02062***	0.13757	0.67417	9.42
Path-based Graph kernel	**Top-K features**	**Precision@10**	**MRR@10**	**itemCov@10**	**aggrEntropy@10**
IG	**5**	0.02266	0.16248	0.58971	9.12
	10	0.01518	0.13221	0.88252	10.26
	15	0.01387	0.13069	0.89762	10.25
ABSTAT	**5**	0.02026	0.15310	0.54825	9.13
	10	0.01519	0.13331	0.57461	9.33
	15	**0.01726***	**0.13510***	0.62606	9.46

Table 1. Experimental results using the entity-based and the path-based Graph kernel recommendation algorithms. In bold the configurations where ABSTAT outperforms IG (the * symbol indicates that the differences between ABSTAT and the IG baseline are statistically significant with $p - value < 0.001$ according to the paired t-test.)

MRR computes the average reciprocal rank of the first relevant recommended item, and hence results particularly meaningful when users are provided with few but valuable recommendations (i.e., Top-1 or Top-3)[8]. To evaluate **aggregate diversity**, we consider *catalog coverage*, i.e., the percentage of items in the catalog recommended at least once and *aggregate entropy* [1]. The former is used to assess the ability of a system to cover the item catalog, namely to recommend as many items as possible. While the latter measures the distribution of the recommendations across all the items, showing whether the recommendations are concentrated on a few items or are better distributed.

The evaluation of the two feature selection methods, IG and ABSTAT, has been done via the well-know **Movielens** 1M dataset. In order to enrich it with information from Linked Data, we started from a dump of the DBpedia dataset[4] and we limited it to the movie domain by linking movies in Movielens dataset with their corresponding DBpedia entries. Table 1 shows the results for entity-based and path-based graph kernel algorithms [6], respectively. When selecting only the first 5 features, the two feature selection methods, IG and ABSTAT, show good values of accuracy, but lower values of aggregate diversity, especially in term of coverage. This is not really surprising as with a lower number of features, the system does not have enough diversified information to select more items and the effect of the popularity bias is stronger. Increasing the number of features the value of diversity increases at the expense of the accuracy. However, a good balance remains between accuracy and diversity thus showing a good trade-off between the two [2]. The implementation of the recommendation algorithm presented in this work and all the experimental results are available `https://github.com/sisinflab/SAC2017`.

[4] `http://downloads.dbpedia.org/2015-10/`

References

1. G. Adomavicius and Y. Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5), May 2012.
2. P. Castells, N. J. Hurley, and S. Vargas. Novelty and diversity in recommender systems. In *Recommender Systems Handbook*. Springer US, Boston, MA, 2015.
3. T. Gottron, M. Knauf, A. Scherp, and J. Schaible. ELLIS: interactive exploration of linked data on the level of induced schema patterns. In *Proceedings of the 2nd International Workshop on Summarizing and Presenting Entities and Ontologies.*, CEUR Workshop Proceedings, 2016.
4. N. Mihindukulasooriya, M. Poveda-Villalón, R. García-Castro, and A. Gómez-Pérez. Loupe - an online tool for inspecting datasets in the linked data cloud. In *Proceedings of the ISWC 2015 Posters & Demonstrations Track*, CEUR Workshop Proceedings, 2015.
5. C. Musto, P. Lops, P. Basile, M. de Gemmis, and G. Semeraro. Semantics-aware graph-based recommender systems exploiting linked open data. In *Proceedings of the 24th Conference on User Modeling Adaptation and Personalization, UMAP 2016*, 2016.
6. V. C. Ostuni, S. Oramas, T. Di Noia, X. Serra, and E. Di Sciascio. Sound and music recommendation with knowledge graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2016.
7. A. Ragone, P. Tomeo, C. Magarelli, T. Di Noia, M. Palmonari, A. Maurino, and E. Di Sciascio. Schema-summarization in linked-data-based feature selection for recommender systems. In *Proceedings of the Symposium on Applied Computing*, SAC '17, pages 330–335. ACM, 2017.
8. Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, N. Oliver, and A. Hanjalic. Climf: learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 2012.
9. B. Spahiu, R. Porrini, M. Palmonari, A. Rula, and A. Maurino. ABSTAT: ontology-driven linked data summaries with pattern minimalization. In *Proceedings of the 2nd International Workshop on Summarizing and Presenting Entities and Ontologies (SumPre 2016) co-located with ESWC.*, volume 1605 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.
10. G. Troullinou, H. Kondylakis, E. Daskalaki, and D. Plexousakis. RDF Digest: Efficient Summarization of RDF/S KBs. In *ESWC*, 2015.