Knowledge-Based Systems 136 (2017) 1–14

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Semantics-aware Recommender Systems exploiting Linked Open Data and graph-based features

Cataldo Musto*, Pasquale Lops, Marco de Gemmis, Giovanni Semeraro

Universita degli Studi di Bari "Aldo Moro", Department of Computer Science, Italy

ARTICLE INFO

Article history:
Received 10 April 2017
Revised 7 August 2017
Accepted 22 August 2017
Available online 23 August 2017

Keywords:
Recommender Systems
Linked Open Data
Semantics
Machine learning
Classifiers

ABSTRACT

The recent spread of Linked Open Data (LOD) fueled the research in the area of Recommender Systems, since the (semantic) data points available in the LOD cloud can be exploited to improve the performance of recommendation algorithms by enriching item representations with new and relevant features.

In this article we investigate the impact of the features gathered from the LOD cloud on a hybrid recommendation framework based on three classification algorithms, Random Forests, Naïve Bayes and Logistic Regression. Specifically, we extend the representation of the items by introducing two new types of features: *LOD-based features*, structured data extracted from the LOD cloud, as the *genre* of a movie or the *writer* of a book, and *graph-based features*, computed on the ground of the topological characteristics of both the *bipartite* graph-based representation connecting users and items, and the *tripartite* representation connecting users, items and properties in the LOD cloud.

In the experimental session we assess the effectiveness of these novel features; results show that the use of information coming from the LOD cloud could improve the overall accuracy of our recommendation framework. Finally, our approach outperform several state-of-the-art recommendation techniques, thus confirming the insights behind this research.

1. Introduction

According to its original vision [7], the goal of the *Semantic Web* was to make machine-readable the whole knowledge available on the Web. This enormous effort, that should have been carried out by stimulating the adoption of shared languages as RDF[1] or OWL[2] and protocols as URI, would have enabled a common framework allowing data to be shared and reused across applications, enterprises, and communities.

Unfortunately, more than fifteen years later the full vision of the Semantic Web has yet to be fully accomplished. Some considerable progress towards this direction has been obtained after the recent spread of the Linked Open Data (LOD) initiative [8], whose goal is to stress and emphasize the importance of publishing and making data *publicly* available and *linked* one to each other.

According to recent statistics,[3] thanks to the collaborative effort behind the LOD initiative, 150 billions of RDF triples and almost 10,000 linked datasets are now available in the so-called LOD cloud, a huge set of interconnected semantic datasets whose *nucleus* is commonly represented by DBpedia [1], the RDF mapping of Wikipedia that acts as a *hub* for most of the RDF triples made available in the LOD cloud. Such RDF triples represent, in a structured form, semantic information covering many topical domains, such as geographical locations, people, companies, books, scientific publications, films, music, TV and radio programs, genes, proteins, drugs, online communities, statistical data, and so on.

As an example for the musical domain, Fig. 1 shows a tiny portion of the properties, available in the LOD cloud, that describe the band *The Coldplay*. Such features range from very basic information, such as the fact that the band has its hometown in *London*, or that *Chris Martin, Jonny Buckland, Guy Berryman*, and *Will Champion* are their members, to more interesting and less trivial data points, as the fact that the group won a *Grammy Award* or plays *Pop music*. All these properties are freely available and can be easily gathered by using the SPARQL query language.[4]

* Corresponding author.
E-mail addresses: cataldomusto@gmail.com, cataldo.musto@uniba.it (C. Musto), pasquale.lops@uniba.it (P. Lops), marco.degemmis@uniba.it (M. de Gemmis), giovanni.semeraro@uniba.it (G. Semeraro).

[1] https://www.w3.org/RDF/.
[2] https://www.w3.org/OWL/.
[3] http://stats.lod2.eu/.
[4] https://www.w3.org/TR/rdf-sparql-query/.

http://dx.doi.org/10.1016/j.knosys.2017.08.015
0950-7051/

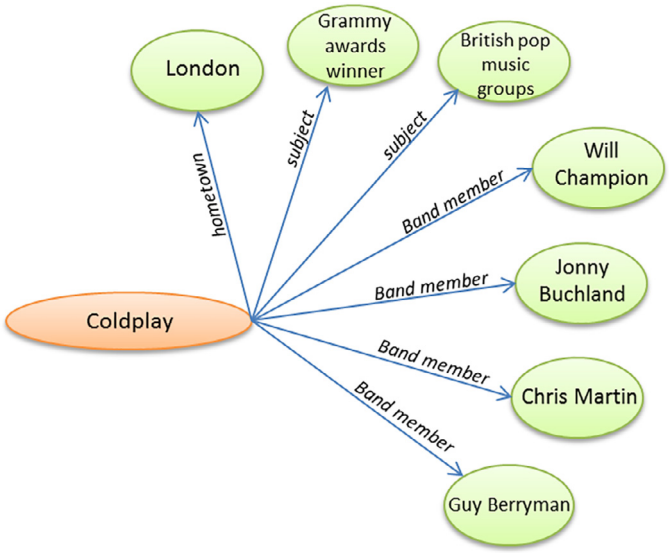

Fig. 1. A (tiny) portion of the properties, available in the LOD cloud, that describe the band *The Coldplay*.

This huge availability of semantics-aware machine-readable data attracted researchers and practitioners willing to investigate how such information can be exploited to develop new services and platforms, or to improve the effectiveness of existing algorithms. A very trending research line investigates the exploitation of these novel (semantic) data points in the area of Recommender Systems (RS) [24], since LOD can be effectively used to handle several problems RSs typically suffer from. *Content-based Recommender Systems* [15] for example, suffer from the well-known problem of *limited content analysis*, i.e. when limited or no features that describe the items to be recommended are available. The knowledge encoded in the LOD cloud can help to deal with this problem, since several features which are relevant for a recommendation task, as the *director* of a movie or the *genre* played by a band, can be gathered from the LOD cloud. This is a largely investigated research line, as we will show in the review of the literature in the area. Similarly, graph-based Recommender Systems can also benefit from such semantic data points. A recent survey on graph-based Recommender Systems is provided in [37], which also presents the data model they are based on. Basically, graph-based Recommender Systems model users and items as nodes in a graph and connect them according to the preferences of users on specific items. This model makes simpler the use of additional information related to users or items. In Fig. 2, the classic bipartite user-item graph representation modeling user-item preferences, as in classical *collaborative filtering algorithms*, can be easily extended by injecting in the graph the properties available in the LOD cloud that describe the items. Besides classical properties, items can be represented by very specific ones, which also allow to discover surprising connections. For example, as shown in Fig. 2 for the movie domain, both the movies *The Matrix* and *Moulin Rouge!* are *Australian films*, and these new information can in turn help to generate better (and maybe *unexpected*) recommendations.

According to these insights, it immediately emerges that RSs may tremendously benefit from the data points available in the LOD cloud. To this end, in this article we investigate the impact of such *exogenous knowledge* on the performance of a *hybrid* recommendation framework based on three classification techniques, Random Forests, Naïve Bayes and Logistic Regression. In this work we followed the hybridization strategy which is typically referred to as *feature combination* [11], i.e. items are represented in terms of different heterogeneous groups of features and are used as training examples to feed the classifiers. Such a model is then exploited to classify new and unseen items as *relevant* or *not relevant* for the target user.

The features we used can be roughly classified in three families: *basic features* (Section 3.1), *content-based features* (Section 3.2) and *topological features* (Section 3.3). *Basic features* include *popularity-based* features, as well as *collaborative features* built on the ground of the user preferences; *content-based features* include features extracted by processing the textual content describing the items, and *LOD-based properties* gathered from the LOD cloud, such as the *genre* of a movie or the *writer* of a book; *topological features* include *bipartite graph-based features* obtained by mining the bipartite graph connecting users to items they liked, and *tripartite graph-based features*which take into account the graph connecting users, items and properties gathered from the LOD cloud.

In the experimental session we assess the effectiveness of our framework by varying the sets of features used to represent items; results provide several interesting insights. Indeed, it emerges that the overall accuracy of the recommendation framework benefits from the introduction of *LOD-based* and *tripartite* graph-based features, and the proposed framework is able to overcome several state-of-the-art recommendation algorithms.

To sum up, the contributions of the paper can be summarized as follows:

- we developed a hybrid recommendation framework based on classification techniques, and we designed families of features to feed the framework. Novel types of features extracted from the LOD cloud have been taken into account besides classical ones, they have been properly combined, and tested on three different datasets;
- we investigated to what extent the injection of knowledge coming from the LOD cloud influences the performance of a recommendation framework based on classification techniques. We have contributed to shed more light on the influence of different item representations based on the knowledge coming from the LOD cloud on the accuracy of recommendations. We tested representations based on properties extracted from the LOD cloud and on topological characteristics of the graph connecting users, items and properties;
- we identified the subsets of features that maximize a specific evaluation metric in our recommendation setting. We tested the ability of specific features and their combination to identify the most relevant items and to correctly rank them in the recommendation list;
- we validated our methodology by evaluating its effectiveness with respect to several state-of-the-art baselines. We compared our approach with widespread and best-performing recommendation algorithms, and with approaches introduced in our previous research as well.

The rest of the paper is organized as follows: Section 2 analyzes related literature. The description of the different features we adopted in our recommendation framework is provided in Section 3, while the details of the experimental evaluation we carried out are described in Section 4. Finally, Section 5 sketches conclusions and future work.

2. Related work

This work investigates the use of features gathered from the LOD cloud in a recommendation framework based on classification techniques. The idea of *casting* the recommendation task to a classification one dates back to the late 90s and is due to Pazzani et al. [49], who proposed a news recommender system that adopted a Naïve Bayes classifier to learn user profiles. After that, the use of such techniques has been largely investigated, especially