

Contents lists available at ScienceDirect

### Information Processing and Management

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



# Introducing linked open data in graph-based recommender systems



Cataldo Musto\*, Pierpaolo Basile, Pasquale Lops, Marco de Gemmis, Giovanni Semeraro

Universita degli Studi di Bari "Aldo Moro", Department of Computer Science, Via E. Orabona, Bari 470126, Italy

#### ARTICLE INFO

# Article history: Received 5 August 2015 Revised 21 November 2016 Accepted 4 December 2016

Keywords:
Recommender systems
PageRank
Graphs
Linked open data
Feature selection
Diversity

#### ABSTRACT

Thanks to the recent spread of the Linked Open Data (LOD) initiative, a huge amount of machine-readable knowledge encoded as RDF statements is today available in the so-called LOD *cloud*. Accordingly, a big effort is now spent to investigate to what extent such information can be exploited to develop new knowledge-based services or to improve the effectiveness of knowledge-intensive platforms as Recommender Systems (RS).

To this end, in this article we study the impact of the *exogenous knowledge* coming from the LOD cloud on the overall performance of a graph-based recommendation framework. Specifically, we propose a methodology to automatically feed a graph-based RS with features gathered from the LOD cloud and we analyze the impact of several widespread feature selection techniques in such recommendation settings.

The experimental evaluation, performed on three state-of-the-art datasets, provided several outcomes: first, information extracted from the LOD cloud can significantly improve the performance of a graph-based RS. Next, experiments showed a clear correlation between the choice of the feature selection technique and the ability of the algorithm to maximize specific evaluation metrics, as *accuracy* or *diversity* of the recommendations. Moreover, our graph-based algorithm fed with LOD-based features was able to overcome several baselines, as collaborative filtering and matrix factorization.

© 2016 Elsevier Ltd. All rights reserved.

#### 1. Introduction

In 2007, the Linked Open Data project (Bizer, 2009) was launched to stimulate researchers and organizations publishing their data in RDF<sup>1</sup> format and adopting shared vocabularies, in order to express an agreed semantics and interlink the data to each other. Nine years later, 150 billions<sup>2</sup> of RDF triples and almost 10,000 linked datasets are available on the Web, thus representing a rapidly growing piece of the big data puzzle (Manyika et al., 2011).

These interconnected RDF statements form a huge decentralized knowledge base, called Linked Open Data (LOD) cloud (see Fig. 1). The LOD cloud covers many topical domains, ranging from government and geographical data to structured

<sup>\*</sup> Corresponding author.

*E-mail addresses*: cataldo.musto@uniba.it (C. Musto), pierpaolo.basile@uniba.it (P. Basile), pasquale.lops@uniba.it (P. Lops), marco.degemmis@uniba.it (M. de Gemmis), giovanni.semeraro@uniba.it (G. Semeraro).

<sup>1</sup> http://www.w3.org/TR/rdf-concepts/.

<sup>&</sup>lt;sup>2</sup> http://stats.lod2.eu/.

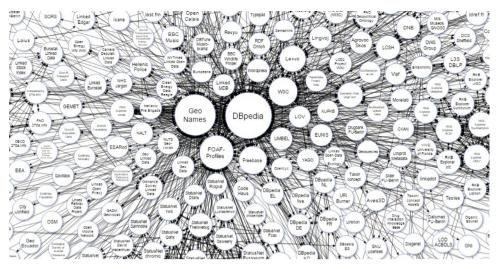


Fig. 1. The core of the Linked Open Data Cloud.

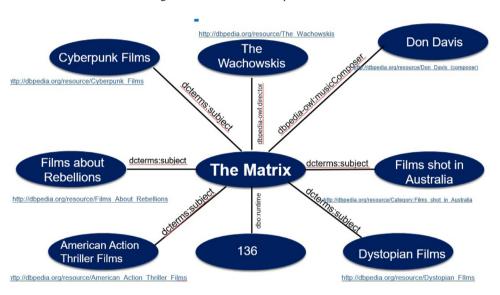


Fig. 2. A (tiny) portion of the properties describing the movie "The Matrix".

information about media (movies, books, etc.) and life sciences. The typical *entry point* to these data is DBpedia (Auer et al., 2007), the RDF mapping of Wikipedia which is commonly considered as the *nucleus* of the emerging *Web of Data*.

Due to the enormous availability of such machine-readable knowledge, a big effort is now spent to investigate whether and how knowledge-intensive services and applications, as Recommender Systems (RS) (Kantor, Rokach, Ricci, & Shapira, 2011), may benefit of this *plethora* of data. By considering a typical pipeline carried out by a RS, a very straightforward use of the information encoded in the LOD cloud regards the enrichment of the representation of the *items* to be recommended as well as of the *preferences of the target user*. For example, the movie *The Matrix* is described in the LOD cloud by means of a huge set of properties (see Fig. 2).

Clearly, some of the properties (as the *director* of the movie, the *year* or the *composer*) are quite trivial, but many others are very fine-grained and can actually enrich the representation of the items by automatically injecting new and useful knowledge. Accordingly, thanks to these novel data points, user preferences and tastes can be better modeled: as an example, it is possible to infer that a user interested in *The Matrix* may be also interested in *Dystopian* or *Cyberpunk* movies.

Similarly, also recommendation algorithms can be boosted by exploiting the non-trivial connections encoded in the LOD cloud. As shown in Fig. 3, the information encoded in DBpedia allows to discover that both *The Matrix* and *The Lost World: Jurassic Park* share some unexpected connections: indeed, by just sifting through the LOD cloud it emerges that both movies, mapped to the URIs http://dbpedia.org/resource/The\_Matrix and http://dbpedia.org/resource/The\_Lost\_World:Jurassic\_Park, respectively, share the common feature category:Films\_shot\_in\_Australia, which is encoded through the property dotterms:subject.

### Download English Version:

## https://daneshyari.com/en/article/4966449

Download Persian Version:

https://daneshyari.com/article/4966449

<u>Daneshyari.com</u>