Content-based recommendations via DBpedia and Freebase: a case study in the music domain

Phuong T. Nguyen, Paolo Tomeo, Tommaso Di Noia, Eugenio Di Sciascio

{phuong.nguyen, paolo.tomeo, tommaso.dinoia, eugenio.disciascio}@poliba.it

Polytechnic University of Bari - Bari (ITALY)

ISWC 2015 – 14th International Semantic Web Conference
October 11-15, 2015 in Bethlehem, Pennsylvania, USA

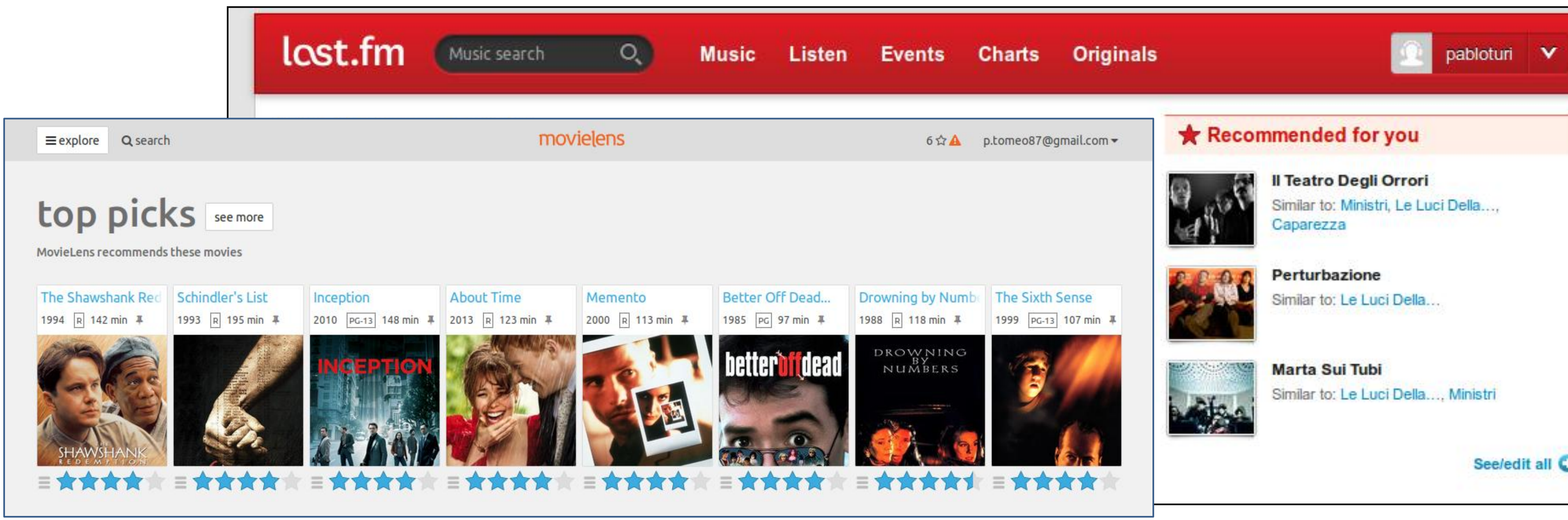
Introduction

- **Content-based Recommender Systems** base on the notion of similarity between items: obviously they need content
- **Web of Data** is an opportunity to foster knowledge-intensive applications
- Does the selection of the underlying knowledge graph affect the results of a recommendation engine?
  - Experiments with **DBpedia** and **Freebase**
- Evaluation in terms of **accuracy**, **sales diversity** and **novelty**

Recommender Systems

***Recommender Systems (RSs) are software tools* and *techniques providing suggestions for items to be of use to a user*.**

[F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. **Recommender Systems Handbook**. Springer, 2011.]



Recommender Systems

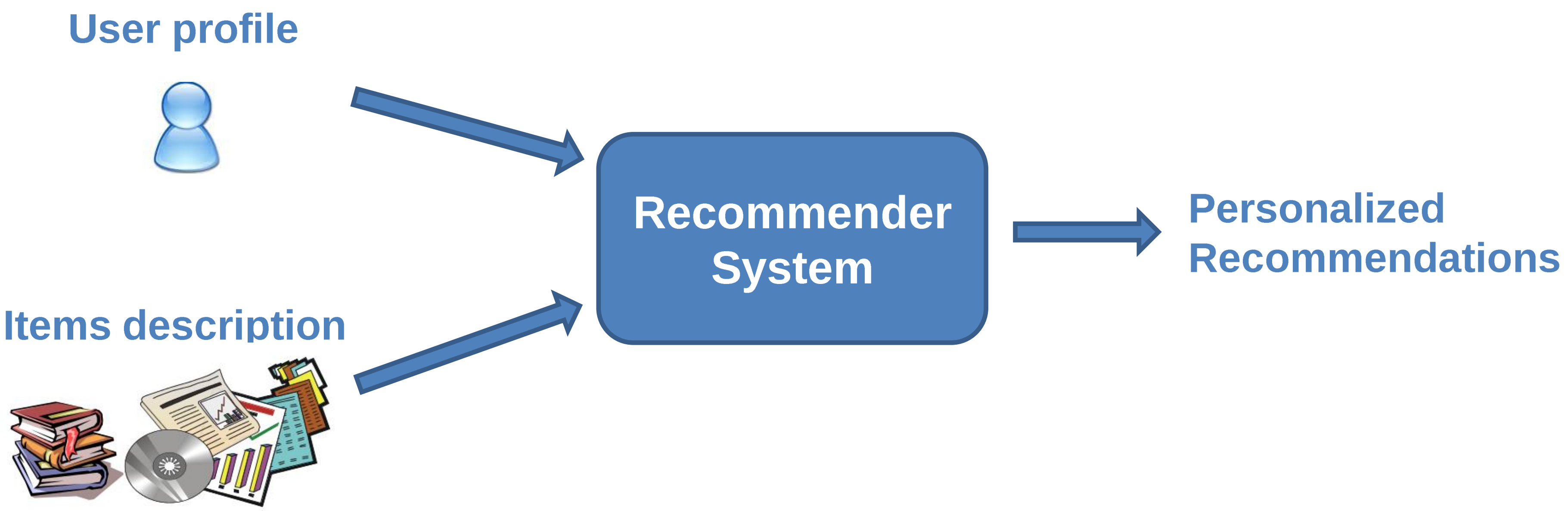
***Recommender Systems (RSs) are software tools* and *techniques providing suggestions for items to be of use to a user*.**

[F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. **Recommender Systems Handbook**. Springer, 2011.]

- **Content-based filtering**
- Collaborative filtering

Content-based RSs

Content-based RSs try to recommend items similar to those a given user has liked in the past. The recommendations are based upon a description of the items and a profile of the user's interests



Main drawback: Limited Content Analysis

Quality of CB recommendations depends on quantity and quality of the features explicitly associated to the items

We need domain knowledge and rich descriptions of the items

P. Lops, M. de Gemmis, G. Semeraro. **Content-based Recommender Systems: State of the Art and Trends.** In Recommender Systems Handbook: A Complete Guide for Research Scientists & Practitioners, Chapter 3, Springer, 2010.

Enrich Data model

Catalog Items

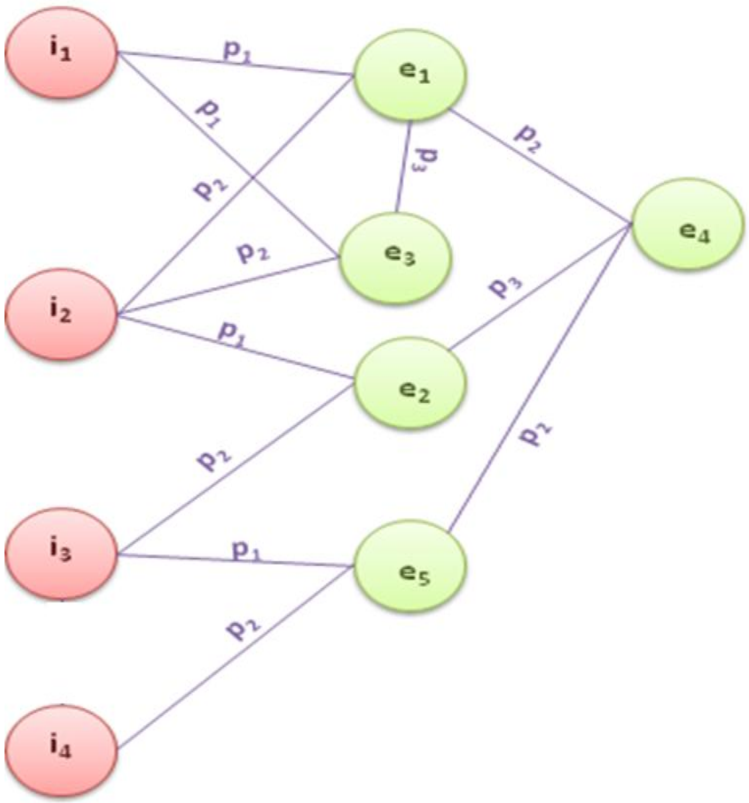
Knowledge Graph

1 – Mapping (Entity linking)
2 – Subgraph extraction

Enrich Data model

Catalog Items

Knowledge Graph

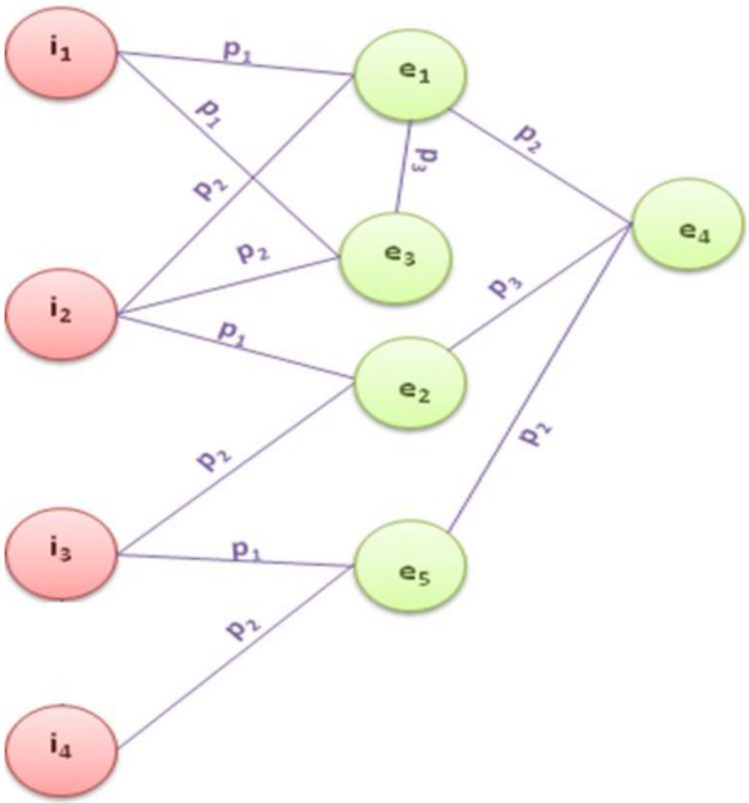


DBpedia

Freebase™

Feature-based Semantic Similarity

1 - describe resources as feature sets
2 - perform similarity calculation on them



one-hop features

i1 -> {e1, e3}   i3 -> {e2, e5}

i2 -> {e1, e3}   i4 -> {e5}

Feature-based Semantic Similarity Metrics

GbkSim
Graph-based Kernel

$$GbkSim(\alpha, \beta) = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \times \sqrt{\sum_{i=1}^{n} (b_i)^2}}$$

VsmSim
Vector Space Model

$$VsmSim_p(\alpha, \beta) = \frac{\sum_{i=1}^{n} a_{i,p} \times b_{i,p}}{\sqrt{\sum_{i=1}^{n} (a_{i,p})^2} \times \sqrt{\sum_{i=1}^{n} (b_{i,p})^2}}$$

FuzzySim
Fuzzy Semantic

$$FuzzySim(\alpha, \beta) = aggr(S_1, S_2, ..., S_n) = \sum_{j=1}^{n} b_j . \varphi_j(m)$$

Jaccard
Jaccard's index

$$Jaccard(\alpha, \beta) = \frac{|N_d(\alpha) \bigcap N_d(\beta)|}{|N_d(\alpha) \bigcup N_d(\beta)|}$$

Content-based Recommender System

k-nearest neighbors algorithm

$$P(u,\alpha) = \frac{\sum_{\beta \in neighbors(\alpha) \cap profile(u)} sim(\alpha,\beta) \cdot r(u,\beta)}{\sum_{\beta \in neighbors(\alpha) \cap profile(u)} sim(\alpha,\beta)}$$

Content-based Recommender System

k-nearest neighbors algorithm

Probability that user u likes α

$$P(u,\alpha) = \frac{\sum_{\beta \in neighbors(\alpha) \cap profile(u)} sim(\alpha,\beta) \cdot r(u,\beta)}{\sum_{\beta \in neighbors(\alpha) \cap profile(u)} sim(\alpha,\beta)}$$

Content-based Recommender System

k-nearest neighbors algorithm

Computed with one of the similarity metrics introduced before

Probability that user u likes α

$$P(u,\alpha) = \frac{\sum_{\beta \in neighbors(\alpha) \cap profile(u)} sim(\alpha,\beta) \cdot r(u,\beta)}{\sum_{\beta \in neighbors(\alpha) \cap profile(u)} sim(\alpha,\beta)}$$

Evaluation

Dataset

Subset of Last.fm hetrec-2011

- **1000 most popular artists and bands**
- **cold users removal (#ratings < avg)**
- **split 80-20% for each user**

Mapping

With DBpedia
http://sisinflab.poliba.it/semanticweb/lod/recsys/datasets/

With Freebase exploiting *owl:sameAs* in DBpedia

Selection of 20% most popular properties from both

	DBpedia Ontology	Freebase
# incoming properties	24	220
# outgoing properties	18	280

Evaluation Metrics

Accuracy	Precision Recall
Sales Diversity	Catalog coverage Entropy and Gini Index (Distribution)
Novelty	% Long-tail

Evaluation Setting

Four independent settings

one-hop	inbound and outbound features
	only outbound features

two-hop	inbound and outbound features
	only outbound features

Top-N varying N from 1 to 50

DBpedia vs Freebase

ACCURACY

Freebase beats DBpedia
except using VSM-Sim

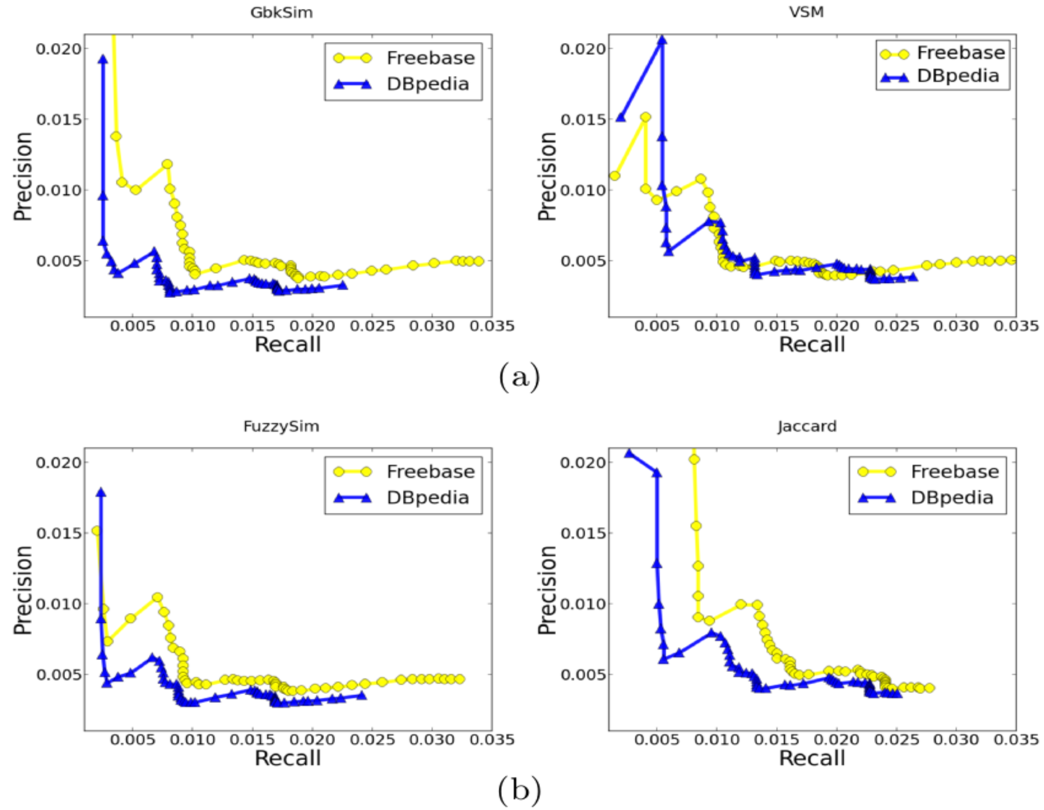
SALES DIVERSITY

Freebase gives better coverage

DBpedia better distribution

NOVELTY

DBpedia beats Freebase



One-hop and two-hop configurations obtain similar trends

Using both inbound and outbound properties gives better results

DBpedia vs Freebase

		Precision	Recall	Coverage	Entropy	Gini	%Long-tail
GbkSim	Top-10	Freebase	Freebase	Freebase	DBpedia	DBpedia	DBpedia
	Top-20	Freebase	Freebase	Freebase	DBpedia	DBpedia	DBpedia
	Top-30	Freebase	Freebase	Freebase	DBpedia	DBpedia	DBpedia
VsmSim	Top-10	Freebase	Freebase	Freebase	DBpedia	DBpedia	DBpedia
	Top-20	Freebase	DBpedia	DBpedia	DBpedia	DBpedia	DBpedia
	Top-30	Freebase	DBpedia	DBpedia	DBpedia	DBpedia	DBpedia
FuzzySim	Top-10	Freebase	Freebase	Freebase	DBpedia	DBpedia	DBpedia
	Top-20	Freebase	Freebase	Freebase	DBpedia	DBpedia	DBpedia
	Top-30	Freebase	Freebase	Freebase	DBpedia	DBpedia	DBpedia
Jaccard	Top-10	Freebase	Freebase	Freebase	Freebase	Freebase	DBpedia
	Top-20	Freebase	Freebase	Freebase	Freebase	DBpedia	DBpedia
	Top-30	Freebase	Freebase	Freebase	Freebase	Freebase	DBpedia

One-hop vs two-hop

			Precision	Recall	Coverage	Entropy	Gini	%Long-tail
GbkSim	Top-10	Freebase	+	+	−	+	+	−
		DBpedia	−	−	+	−	−	−
	Top-20	Freebase	+	+	−	+	+	+
		DBpedia	+	+	+	+	+	∼
	Top-30	Freebase	+	+	−	+	+	∼
		DBpedia	+	+	+	∼	+	−
VsmSim	Top-10	Freebase	−	−	+	+	+	−
		DBpedia	−	−	+	+	+	−
	Top-20	Freebase	−	−	+	+	+	−
		DBpedia	−	−	+	+	+	−
	Top-30	Freebase	−	−	+	+	+	−
		DBpedia	−	−	+	+	−	−
FuzzySim	Top-10	Freebase	−	−	−	+	+	−
		DBpedia	+	+	+	−	∼	∼
	Top-20	Freebase	+	+	∼	+	+	−
		DBpedia	+	+	+	∼	+	+
	Top-30	Freebase	+	+	−	+	+	−
		DBpedia	+	+	+	+	+	∼
Jaccard	Top-10	Freebase	−	−	+	+	∼	+
		DBpedia	−	−	+	+	+	−
	Top-20	Freebase	−	−	+	−	−	−
		DBpedia	−	−	+	+	+	−
	Top-30	Freebase	∼	∼	+	−	−	−
		DBpedia	−	−	+	+	+	∼

Discussion

Freebase brings higher accuracy and lower novelty

it is richer and has a strong crowd-sourced nature

DBpedia gives better distribution (Gini and Entropy)

but the coverage it provides is too low

Exploring up to two hops improves coverage and distribution but penalize novelty

increase of connections among items but in particular among the most popular

Conclusion

Comparison between DBpedia and Freebase for content-based recommedantions in terms of Accuracy, Sales Diversity and Novelty

We showed that the choice of the right dataset might affect the performance of the system

Future work

Same experiments using graph-based similarity metrics

Automated feature selection instead of most popular

Q & A

Thanks for your attention!