Linked Open Data for Semantics-Aware Recommender Systems

Pierpaolo Basile, Cataldo Musto

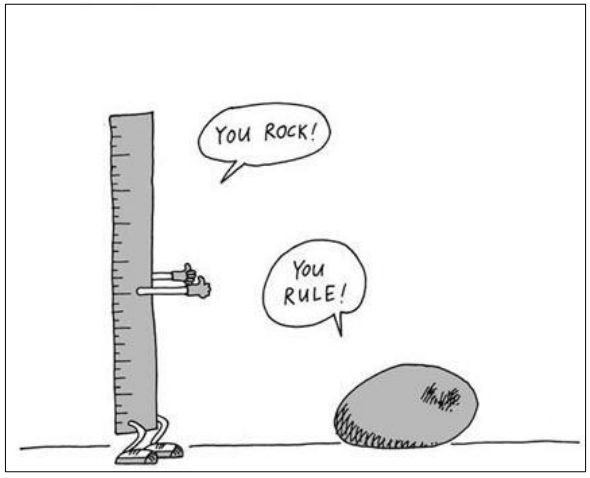
Tommaso Di Noia, Paolo Tomeo

tutorial@ESWC 2017 – Portoroz (Slovenia) – 29/05/2017

About us

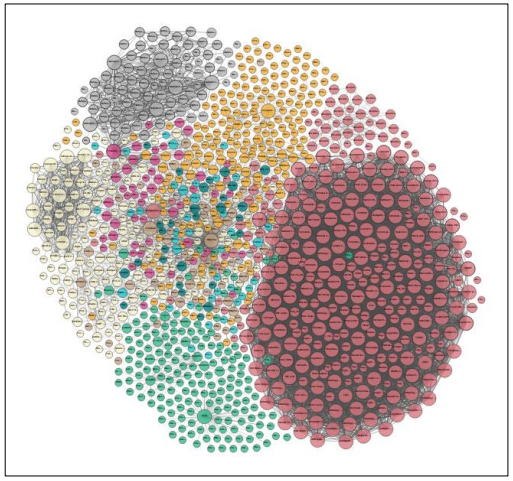
pierpaolo.basile@uniba.it
@basilepp

Semantics, NLP



tommaso.dinoia@poliba.it
@TommasoDiNoia

Linked Data, RecSys



cataldo.musto@.uniba.it
@cataldomusto

RecSys, User Modeling

in this tutorial

how to introduce semantics

in **recommender systems**?

Agenda

Why?

Why do we need **intelligent information access**?
Why do we need **content**?
Why do we need **semantics**?

How?

How to **introduce semantics**?
Basics of **Natural Language Processing**
Encoding **exogenous semantics** (top-down approaches)
Encoding **endogenous semantics** (bottom-up approaches)

What?

Recommender Systems based on **Distributional Semantics**
Recommender Systems on **Entity Linking techniques**
Recommender Systems based on **Linked Open Data**
Explaining Recommendations through **Linked Open Data**

Why?

Why do we need **intelligent information access**?

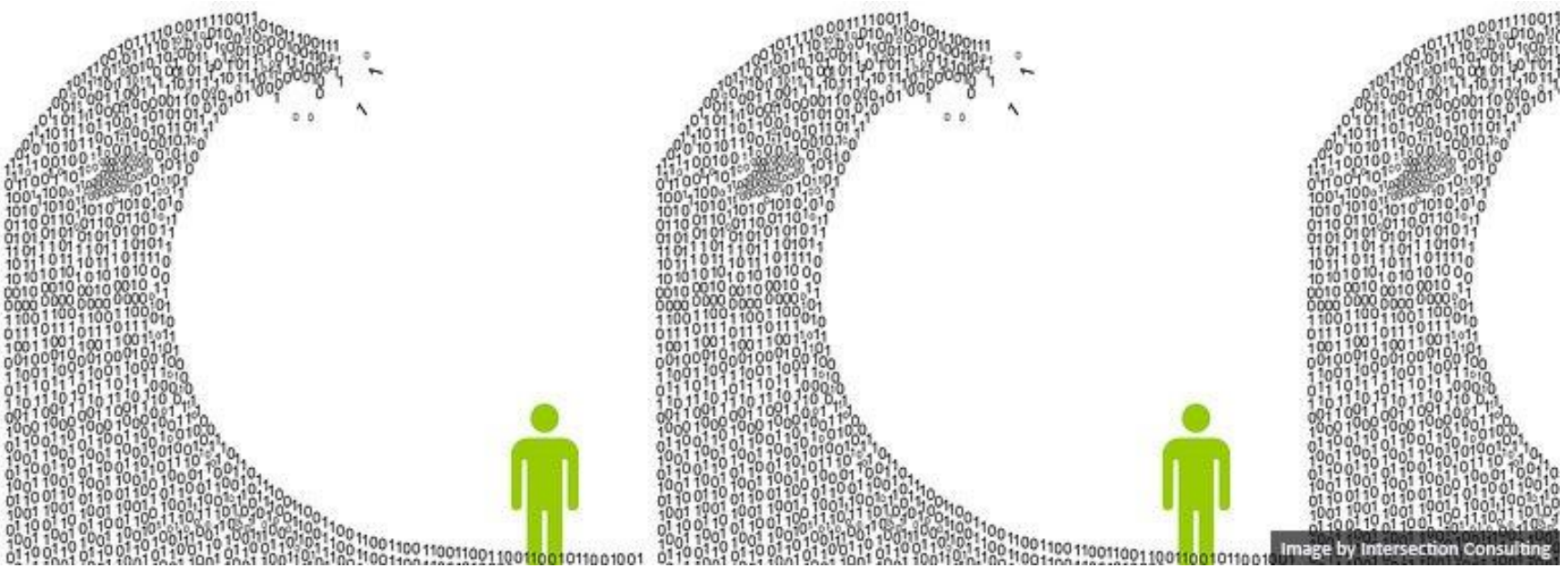
Sources: *Email:* 2013-2015: Radicati; *Facebook:* 2014 Qmee; 2015 Wishpond; *Google:* 2013 Statista; 2014 AdWeek; 2015 Internet Live Stats; *Instagram:* 2013 Tech Crunch; 2014 Nuke Suite; 2015 Instagram; *Twitter:* 2013 Internet Live Stats; 2014 Internet Live Stats; 2015 Tech Insider; *WordPress:* 2013 WordPress; 2014 WordPress; 2015 Internet Live Stats; *WhatsApp:* 2013 Fierce Mobile IT; 2014 Slash Gear; 2015 Expanded Ramblings; *YouTube:* 2013 Youtube Global Blog: 2014 Reel SEO.

physiologically

impossible

to follow the information flow

in **real time**



we can handle

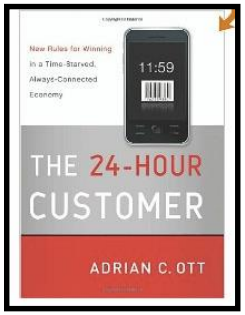
126 bits of information/day

we deal with

393 bits of information/day

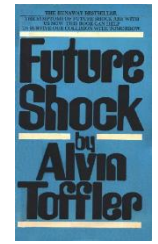
ratio: more than **3x**

(Source: Adrian C.Ott, The 24-hour customer)



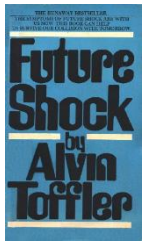
Information overload

(Appeared for the first time in «Future Shock» by Alvin Toffler, 1970)



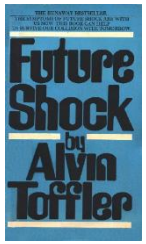
Information overload

(Appeared for the first time in «Future Shock» by Alvin Toffler, 1970)

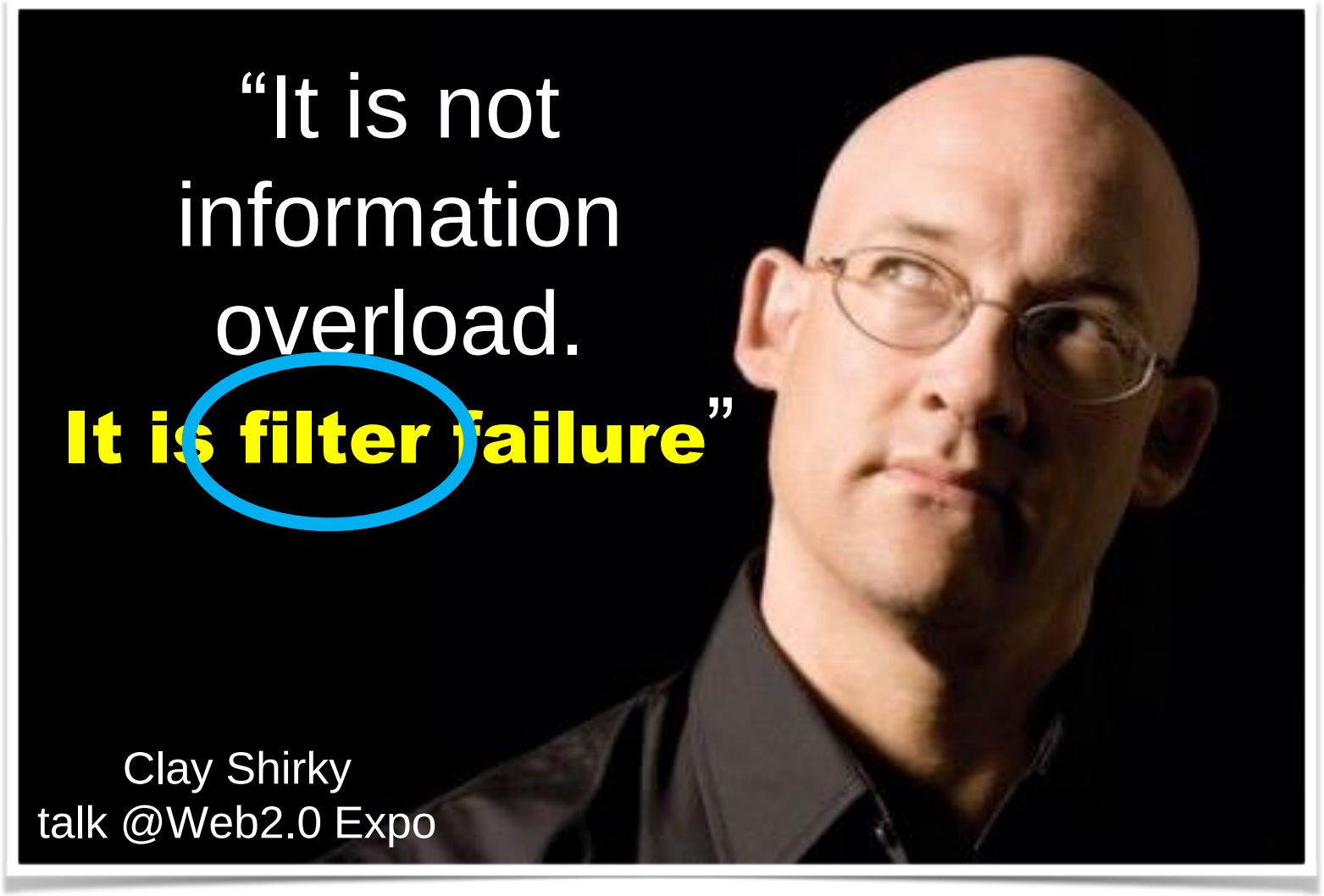


Information overload

(Appeared for the first time in «Future Shock» by Alvin Toffler, 1970)



Information overload

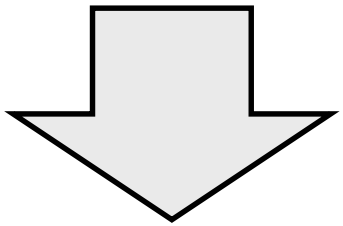


Challenge

To effectively cope with

information overload

we need to **filter** the information flow



We need technologies and algorithms for
intelligent information access

… and we already have some evidence!

Intelligent Information Access

success stories

Information Retrieval (Search Engines)

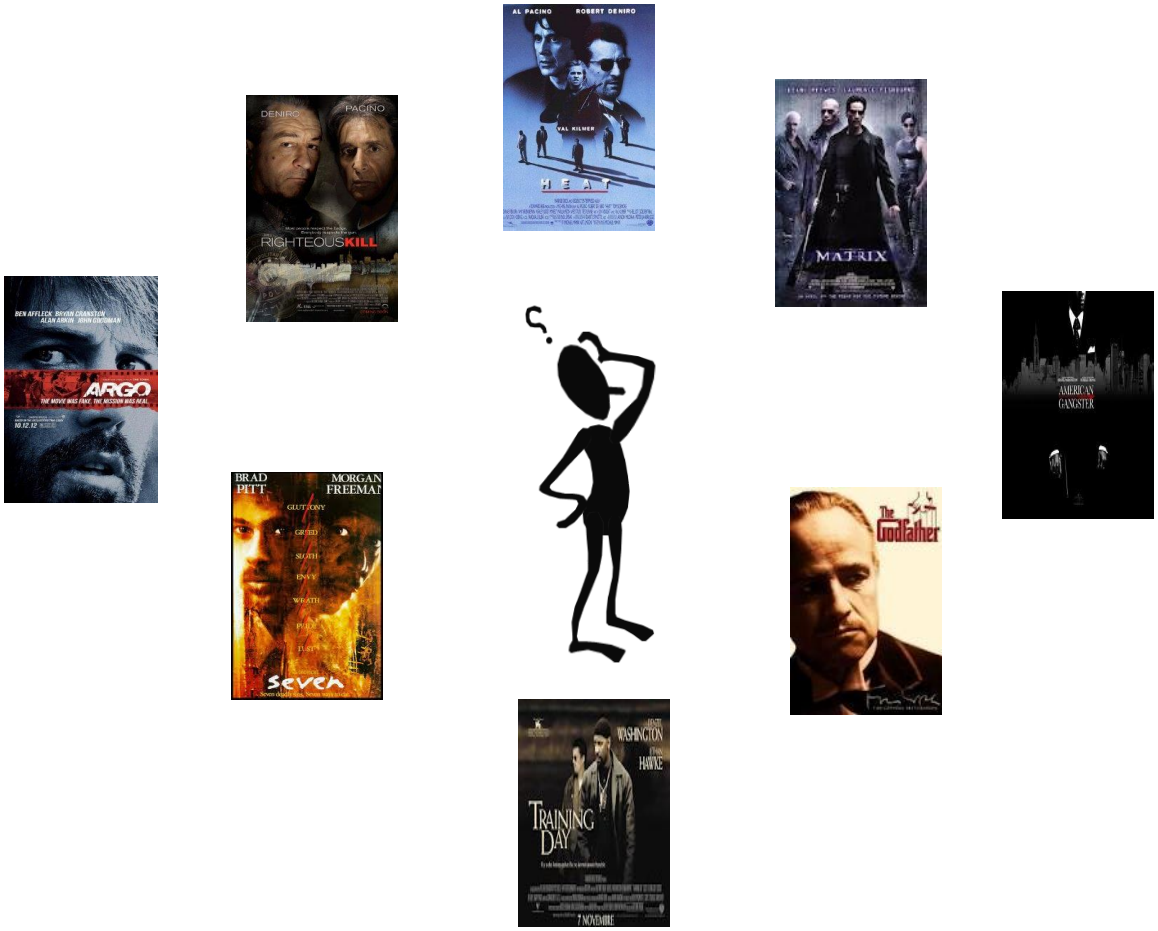
Intelligent Information Access

success stories

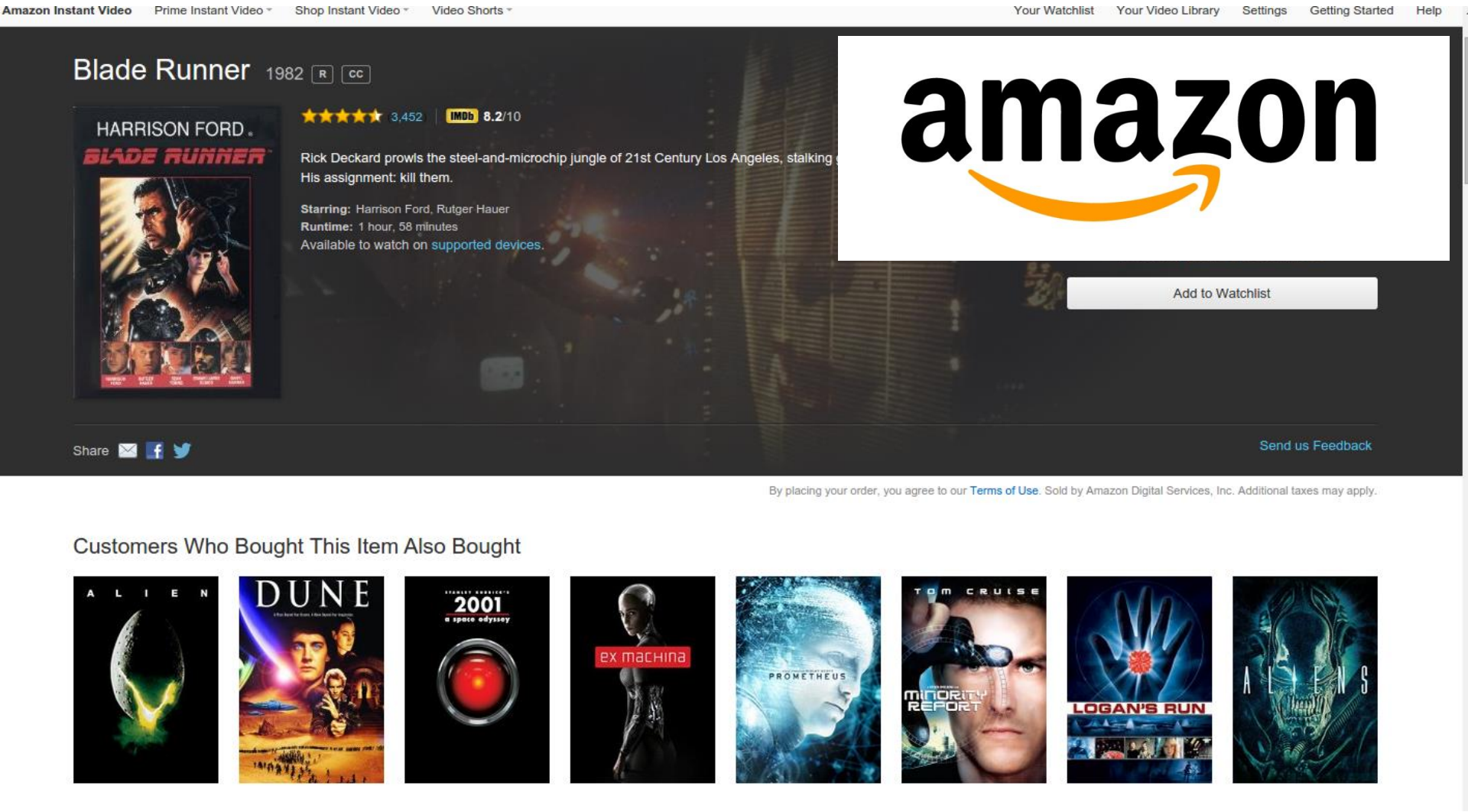
Information Filtering (Recommender Systems)

Recommender Systems

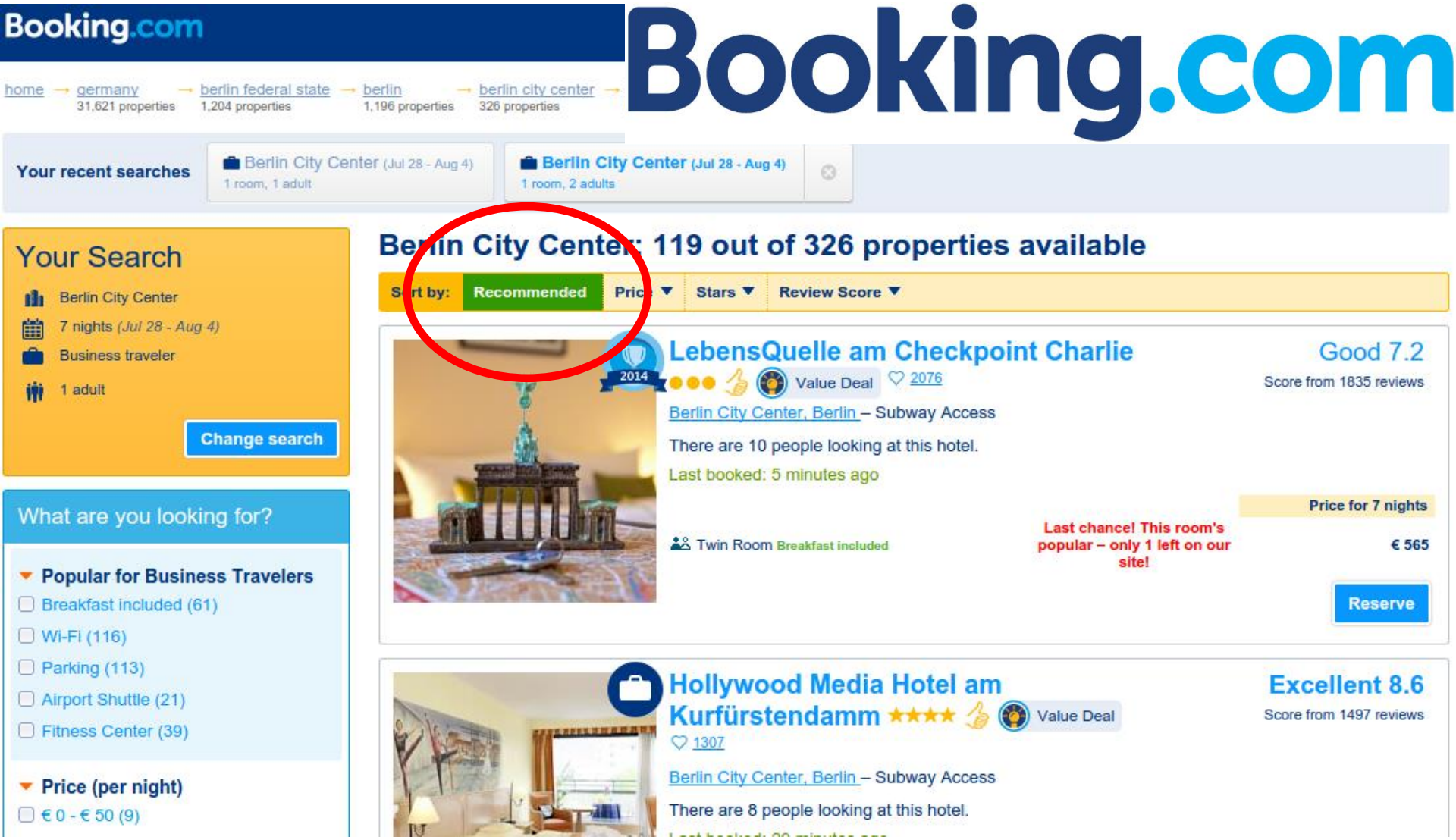
- Help users in dealing with Information/Choice Overload
- Help to match users with items



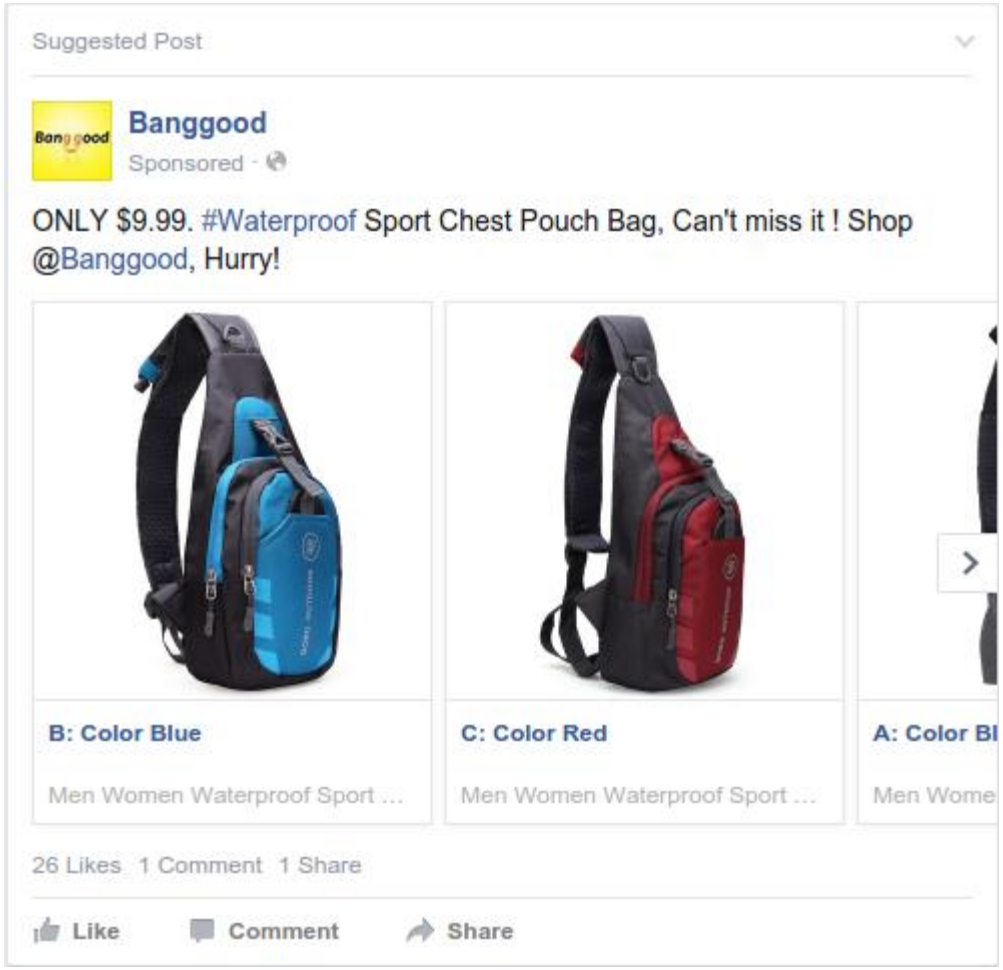
Recommender Systems



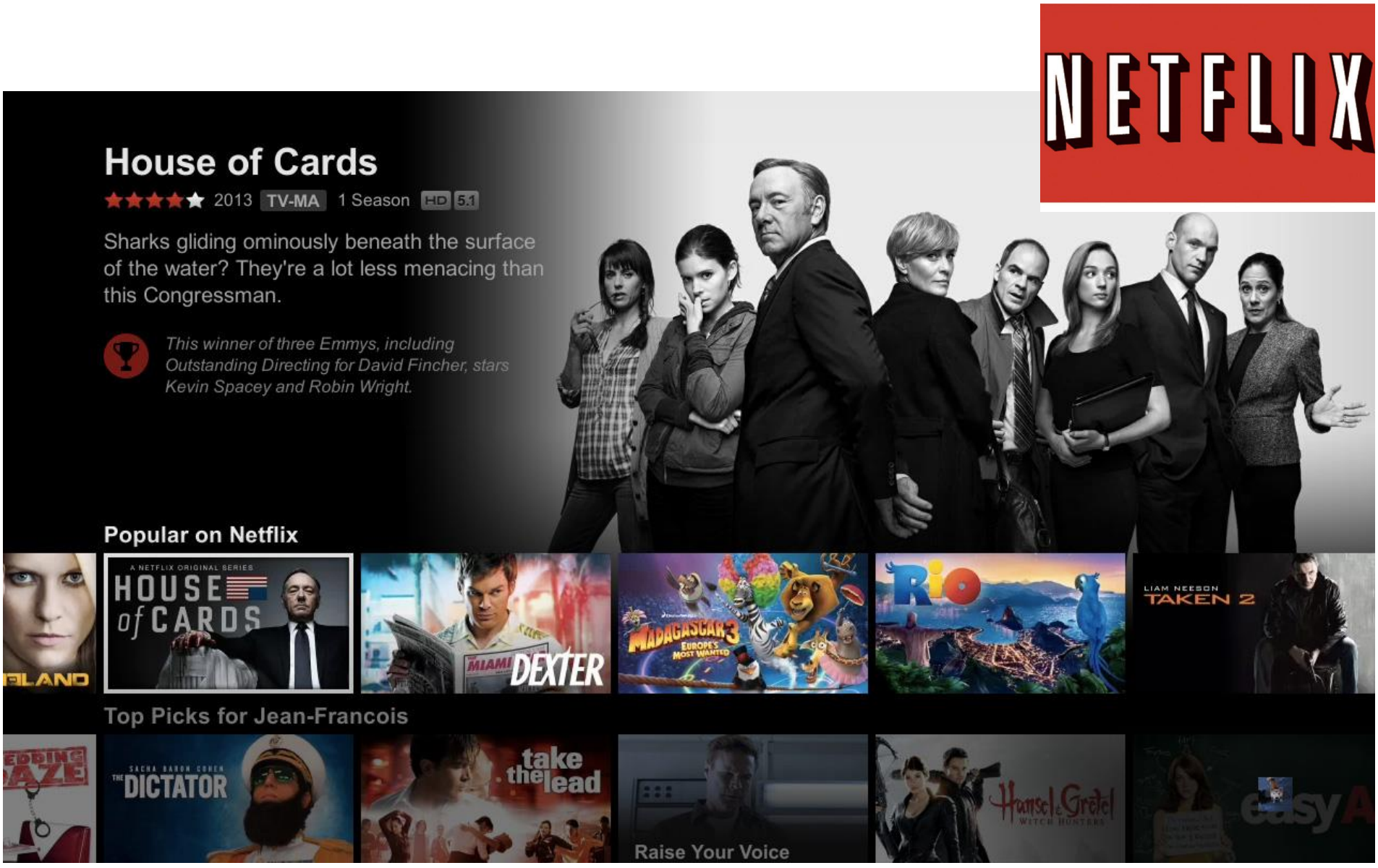
Recommender Systems



Recommender Systems



Recommender Systems



Some definitions

- ✓ In its most common formulation**, the recommendation problem is reduced to the problem of estimating ratings** for the items that have not been seen by a user.

[G. Adomavicius and A. Tuzhilin. **Toward the Next Generation of Recommender Systems: A survey of the State-of-the-Art and Possible Extension**. TKDE, 2005.]

- ✓ 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, 2015.]

The Recommendation Problem

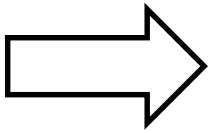
Estimate a utility function to **automatically predict** how much a user will like an item which is unknown to them.

Input

Set of users $U = \{u_1, \dots, u_M\}$

Set of items $X = \{x_1, \dots, x_N\}$

Utility function $f: U \times X \rightarrow R$



Output

$\forall\, u \in U, x'_u = \arg max_{x \in X}\, f(u, x)$

The Rating Matrix

	The Matrix	Titanic	I love shopping	Argo	Love Actually	The hangover
Tommaso	5	1	2	4	3	?
Francesco	2	4	5	3	5	2
Vito	4	3	2	4	1	3
Walter	3	5	1	5	2	4
Cataldo	4	4	5	3	5	2

The Rating Matrix (in real world)

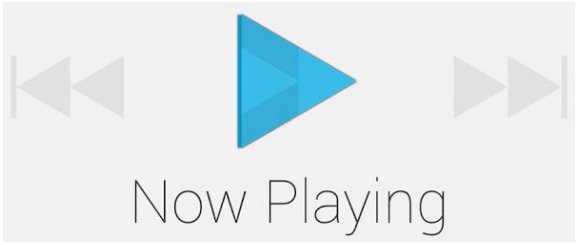
	The Matrix	Titanic	I love shopping	Argo	Love Actually	The hangover
Tommaso	5	?	?	4	3	?
Francesco	2	4	5	?	5	?
Vito	?	3	?	4	?	3
Walter	3	5	?	5	2	?
Cataldo	4	?	5	?	5	2

$$sparsity = 1 - \frac{|R|}{|X| \cdot |U|}$$

Ratings

Explicit

Implicit



Rating Prediction vs Ranking

Best

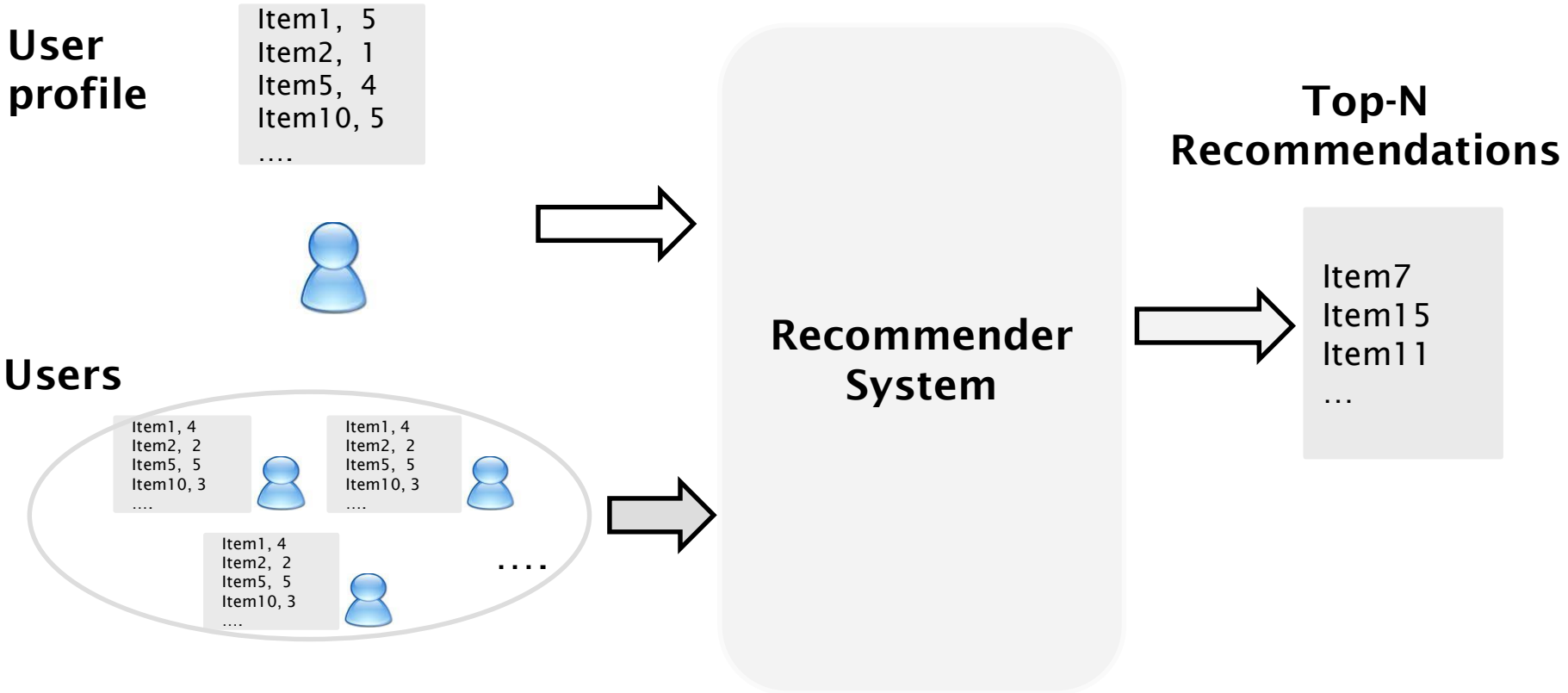
Worst

Recommendation techniques

- Content-based
- Collaborative filtering
- Demographic
- Knowledge-based
- Community-based
- Hybrid recommender systems

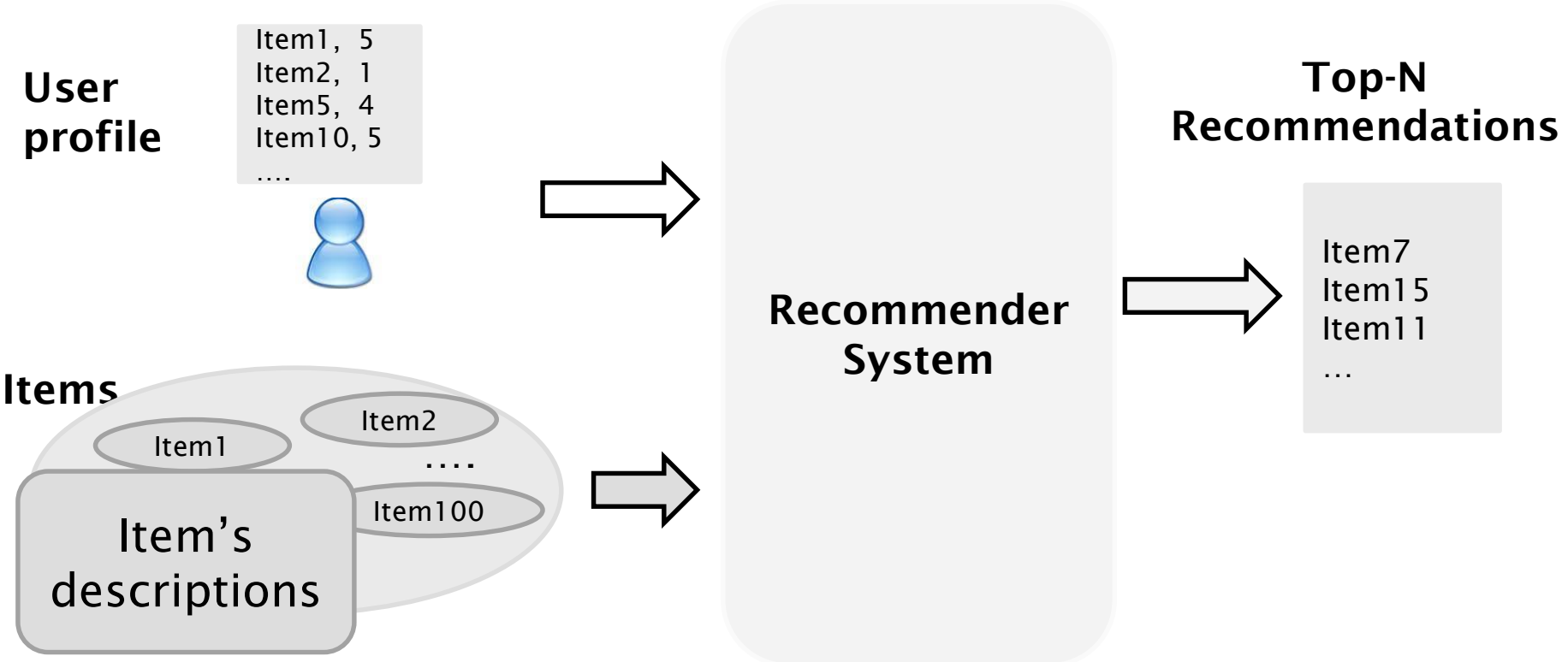
Collaborative RS

Collaborative RSs recommend items to a user by identifying **other users with a similar profile**



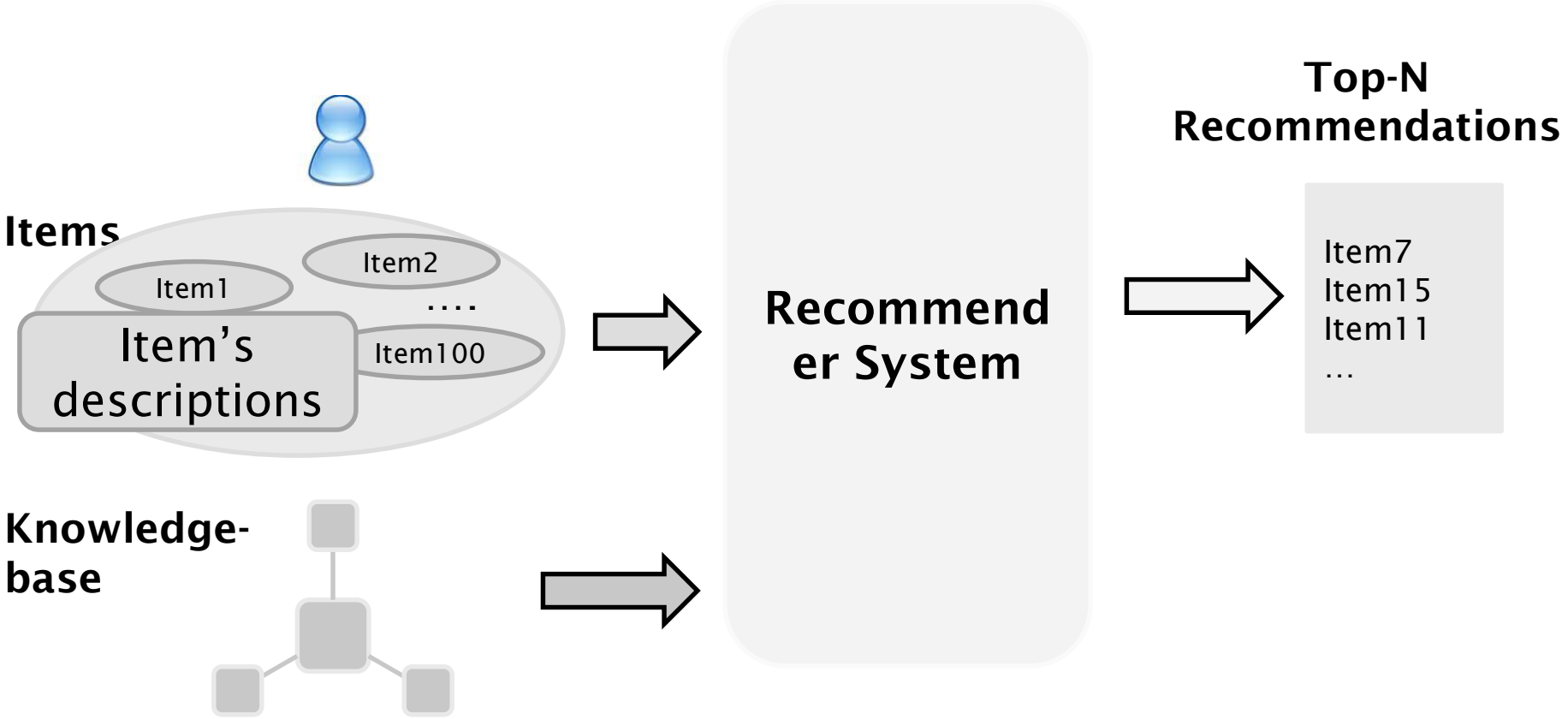
Content-based RS

CB-RSs recommend items to a user **based on their description** and on the profile of the user's interests



Knowledge-based RS

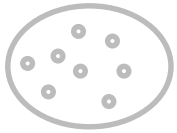
KB-RSs recommend items to a user based on their description and **domain knowledge encoded in a knowledge base**



Collaborative Filtering

- **Memory-based**
  - ✓ Mainly based on k-NN
  - ✓ Does not require any preliminary model building phase
- **Model-based**
  - ✓ Learn a predictive model before computing recommendations

User-based Collaborative Recommendation



$= X$

Pearson's correlation coefficient

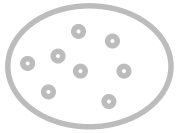
$$sim(u_i, u_j) = \frac{\sum_{x \in X}\left(r_{u_i,x} - \overline{r_{u_i}}\right) * \left(r_{j,x} - \overline{r_{u_j}}\right)}{\sqrt{\sum_{x \in X}\left(r_{u_i,x} - \overline{r_{u_i}}\right)^2} * \sqrt{\sum_{x \in X}\left(r_{u_j,x} - \overline{r_{u_j}}\right)^2}}$$

	The Matrix	Titanic	I love shopping	Argo	Love Actually	The hangover
Tommaso	5	1	2	4	3	?
Francesco	2	4	5	3	5	2
Vito	4	3	2	4	1	3
Walter	3	5	1	5	2	4
Cataldo	4	4	5	3	5	2

Rate prediction

$$\tilde{r}(u_i, x') = \overline{r_{u_i}} + \frac{\sum_{u_j \in N} sim(u_i, u_j) * \left(r_{u_j,x'} - \overline{r_{u_j}}\right)}{\sum_{u_j \in N} sim(u_i, u_j)}$$

User-based Collaborative Recommendation



$= X$

Pearson's correlation coefficient

$$sim(u_i, u_j) = \frac{\sum_{x \in X}(r_{u_i,x} - \overline{r_{u_i}}) * (r_{j,x} - \overline{r_{u_j}})}{\sqrt{\sum_{x \in X}(r_{u_i,x} - \overline{r_{u_i}})^2} * \sqrt{\sum_{x \in X}(r_{u_j,x} - \overline{r_{u_j}})^2}}$$

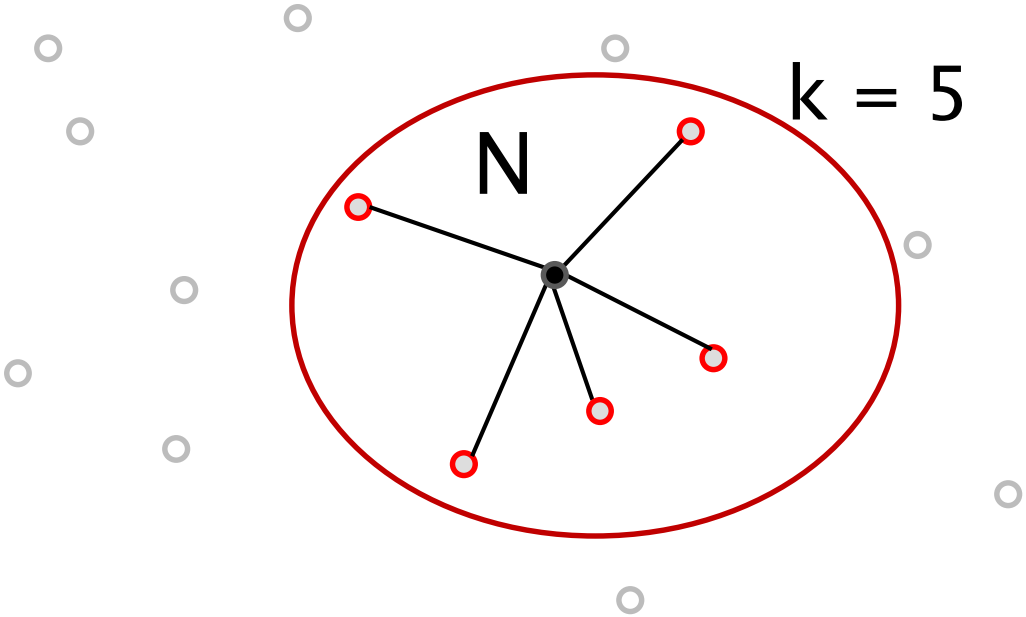
average rating

	The Matrix	Titanic	I love shopping	Argo	Love Actually	The hangover
Tommaso	5	1	2	4	3	?
Francesco	2	4	5	3	5	2
Vito	4	3	2	4	1	3
Walter	3	5	1	5	2	4
Cataldo	4	4	5	3	5	2

Rate prediction

$$\tilde{r}(u_i, x') = \overline{r_{u_i}} + \frac{\sum_{u_j \in N} sim(u_i, u_j) * (r_{u_j,x'} - \overline{r_{u_j}})}{\sum_{u_j \in N} sim(u_i, u_j)}$$

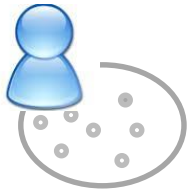
k-Nearest Neighbors



A neighborhood of 20 to 50 neighbors is a reasonable choice

[Herlocker et al. *An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms*, Information Retrieval 5 (2002), no. 4, 287–310.]

Item-based Collaborative Recommendation



$= X_{u_i}$

Cosine Similarity

$$sim(\vec{x_i}, \vec{x_j}) = \frac{\vec{x_i} \cdot \vec{x_j}}{|\vec{x_i}| * |\vec{x_j}|} = \frac{\sum_u r_{u,x_i} * r_{u,x_j}}{\sqrt{\sum_u r_{u,x_i}^2} * \sqrt{\sum_u r_{u,x}^2}}$$

Adjusted Cosine Similarity

$$sim(\vec{x_i}, \vec{x_j}) = \frac{\sum_u (r_{u,x_i} - \bar{r_u}) * (r_{u,x_j} - \bar{r_u})}{\sqrt{\sum_u (r_{u,x_i} - \bar{r_u})^2} * \sqrt{\sum_u (r_{u,x_j} - \bar{r_u})^2}}$$

	The Matrix	Titanic	I love shopping	Argo	Love Actually	The hangover
Tommaso	5	1	2	4	3	?
Francesco	2	4	5	3	5	2
Vito	4	3	2	4	1	3
Walter	3	5	1	5	2	4
Cataldo	4	4	5	3	5	2

Rate prediction

$$\tilde{r}(u_i, x') = \frac{\sum_{x \in X_{u_i}} sim(\vec{x}, \vec{x'}) * r_{x,u_i}}{\sum_{x \in X_{u_i}} sim(\vec{x}, \vec{x'})}$$

CF drawbacks

- **Sparsity / Cold-start**
  - ✓ New user
  - ✓ New item
- **Grey sheep problem**

Content-based RS

- Items are described in terms of **attributes/features**
- **A finite set of values** is associated to each feature
- Item representation **is a (Boolean) vector**

Content-based RS

CB-RSs try to recommend **items similar* to those a given user has liked in the past**

[M. de Gemmis et al. Recommender Systems Handbook. Springer. 2015]

- **Heuristic-based**
  - ✓ Usually adopt techniques borrowed from IR
- **Model-based**
  - ✓ Often we have a model for each user

(*) similar from a content-based perspective

CB drawbacks

- Content overspecialization
- Portfolio effect
- Sparsity / Cold-start
  - ✓ New user

Knowledge-based RS

- Conversational approaches
- Reasoning techniques
  - ✓ Case-based reasoning
  - ✓ Constraint reasoning

Hybrid recommender systems

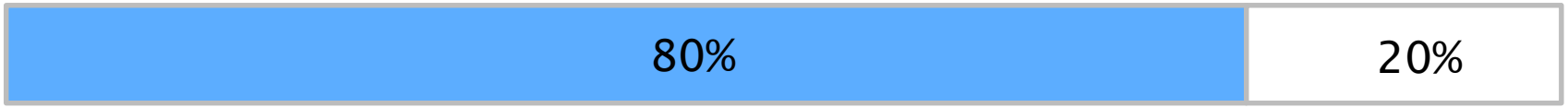
Weighted	**The scores (or votes) of several recommendation techniques are combined together to produce a single recommendation.**
Switching	The system switches between recommendation techniques depending on the current situation.
Mixed	Recommendations from several different recommenders are presented at the same time
Feature combination	Features from different recommendation data sources are thrown together into a single recommendation algorithm.
Cascade	One recommender refines the recommendations given by another.
Feature augmentation	Output from one technique is used as an input feature to another.
Meta-level	The model learned by one recommender is used as input to another.

[Robin D. Burke. Hybrid recommender systems: Survey and experiments. *User Model. User-Adapt. Interact.*, 12(4):331–370, 2002.]

Evaluation: Dataset split

- Training Set
- Test Set (TS)

hold-out



k-fold cross-validation

…

Evaluation Protocols

- **Rated test-items**

- **All unrated items**: compute a score for every item not rated by the user (also items not appearing in the user test set)

Evaluation: metrics for rating prediction

Mean Absolute Error

$$MAE = \frac{1}{|TS|} \cdot \sum_{(u,x_i) \in TS} |\tilde{r}_{u,x_i} - r_{u,x_i}|$$

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{|TS|} \cdot \sum_{(u,x_i) \in TS} (\tilde{r}_{u,x_i} - r_{u,x_i})^2}$$

- **Drawback: not very suitable for top-N recommendation**
  - ✓ Errors in the highest part of the recommendation list are considered in the same way as the ones in the lowest part

Accuracy metrics for top-N recommendation

Precision @ N

$$P_u@N = \frac{|L_u(N) \cap TS_u^+|}{N}$$

$L_u(N)$ is the recommendation list up to the N-th element

TS_u^+ is the set of relevant test items for u

Recall @ N

$$R_u@N = \frac{|L_u(N) \cap TS_u^+|}{|TS_u^+|}$$

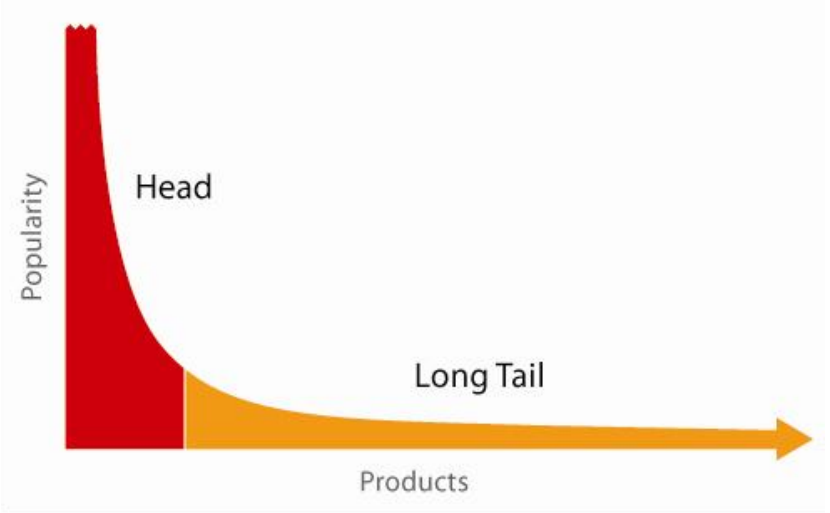
normalized Discount Cumulative Gain @ N

$$nDCG_u@N = \frac{1}{IDCG@N} \sum_{k=1}^{N} \frac{2^{r_{u,k}} - 1}{\log_2(1+k)}$$

IDCG@N indicates the score Obtained by an ideal ranking of $L_u(N)$

Is all about precision?

- **Novelty**
  - ✓ Recommend items in the long tail
- **Diversity**
  - ✓ Avoid to recommend only items in a small subset of the catalog
  - ✓ Suggest diverse items in the recommendation list
- **Serendipity**
  - ✓ Suggest unexpected but interesting items



Is all about precision?

Entropy – Based Novelty

$$EBN_u@N = -\sum_{x \in L_u(N)} p_i \cdot \log_2 p_i$$

$$p_i = \frac{|\{\, u \in U \mid x \text{ is relevant to } u \,\}|}{|U|}$$

Intra – List Diversity

$$ILD_u@N = \frac{1}{2} \cdot \sum_{x_i \in L_u(n)} \sum_{x_j \in L_u(N)} 1 - sim(x_i, x_j)$$

$$ILD@N = \frac{1}{|U|} \cdot \sum_{u \in U} ILD_u@N$$

Aggregate Diversity

$$ADin@N = \frac{|\bigcup_{u \in U} L_u(N)|}{|X|}$$

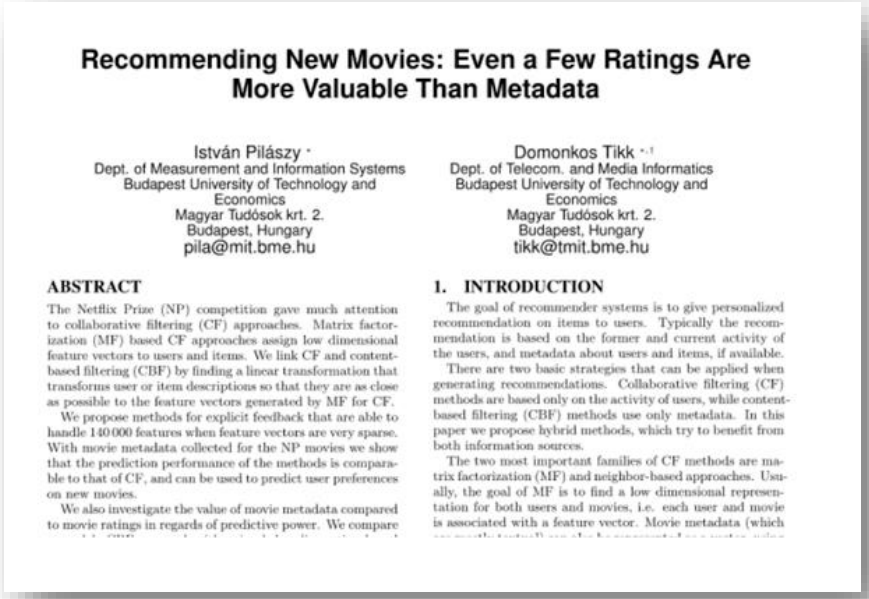
Why?

Why do we need **content**?

Why do we need content?

Several Recommender Systems **perfectly work without using any content!** (e.g.Amazon)

Collaborative Filtering and **Matrix Factorization** are state of the art techniques for implementing Recommender Systems



Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata

István Pilászy
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
pila@mit.bme.hu

Domonkos Tikk
Dept. of Telecom. and Media Informatics
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
tikk@tmit.bme.hu

ABSTRACT

The Netflix Prize (NP) competition gave much attention to collaborative filtering (CF) approaches. Matrix factorization (MF) based CF approaches assign low dimensional feature vectors to users and items. We link CF and content-based filtering (CBF) by finding a linear transformation that transforms user or item descriptions so that they are as close as possible to the feature vectors generated by MF for CF.

We propose methods for explicit feedback that are able to handle 140 000 features when feature vectors are very sparse. With movie metadata collected for the NP movies we show that the prediction performance of the methods is comparable to that of CF, and can be used to predict user preferences on new movies.

We also investigate the value of movie metadata compared to movie ratings in regards of predictive power. We compare

1. INTRODUCTION

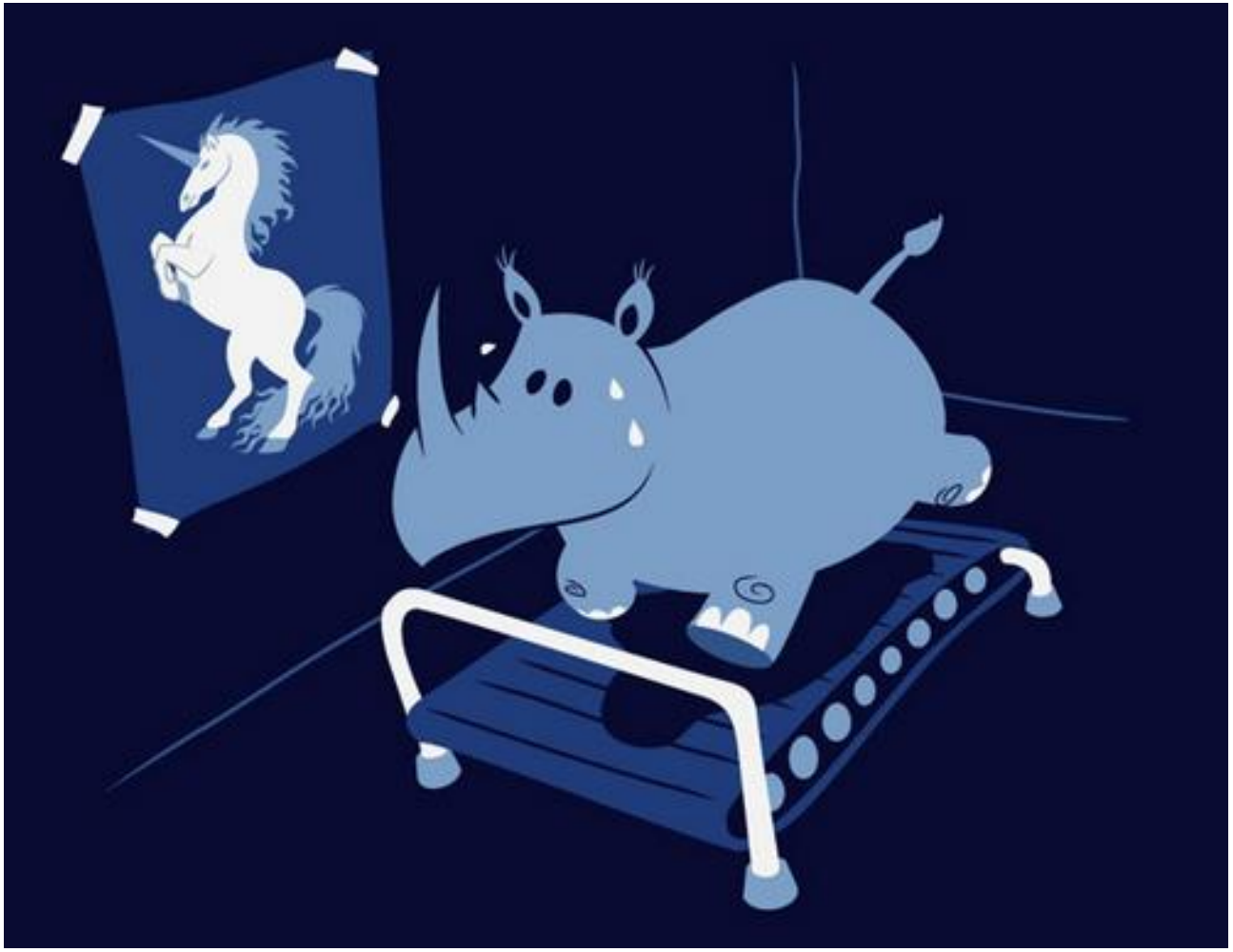
The goal of recommender systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of the users, and metadata about users and items, if available.

There are two basic strategies that can be applied when generating recommendations. Collaborative filtering (CF) methods are based only on the activity of users, while content-based filtering (CBF) methods use only metadata. In this paper we propose hybrid methods, which try to benefit from both information sources.

The two most important families of CF methods are matrix factorization (MF) and neighbor-based approaches. Usually, the goal of MF is to find a low dimensional representation for both users and movies, i.e. each user and movie is associated with a feature vector. Movie metadata (which

(ACM RecSys 2009,
by Neflix Challenge winners)

Why do we need content?



Content **can tackle some issues** of **collaborative filtering**

Why do we need content?

					
				✓	
		✓			
			✓		
					✓
	✓				

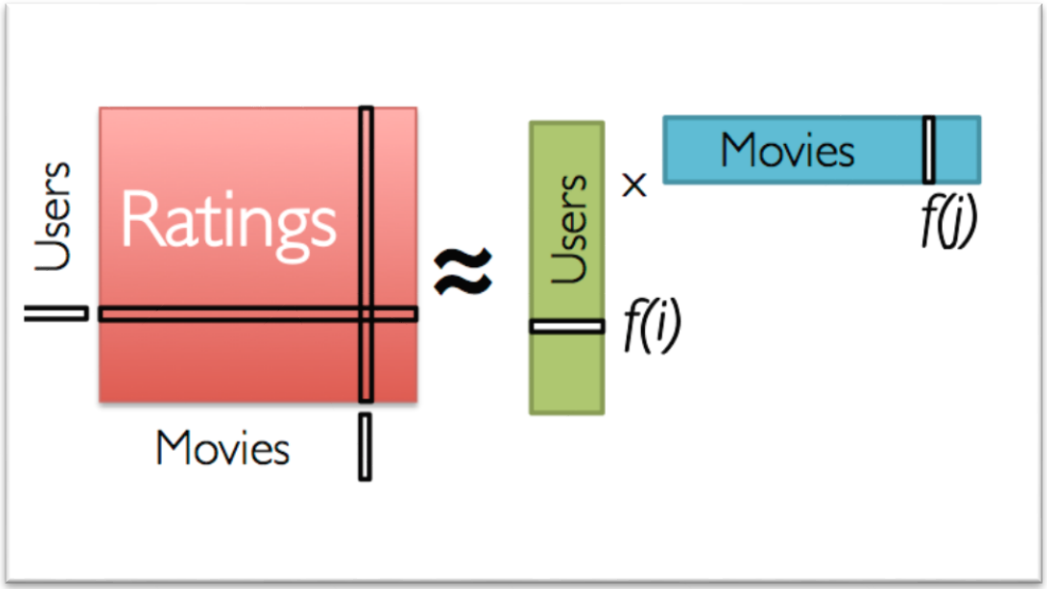
Collaborative Filtering issues: sparsity

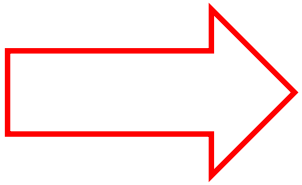
Why do we need content?

					
	✓	✓		✓	
		✓			?
			✓		
				✓	
	✓	✓			

Collaborative Filtering issues: new item problem

Why do we need content?

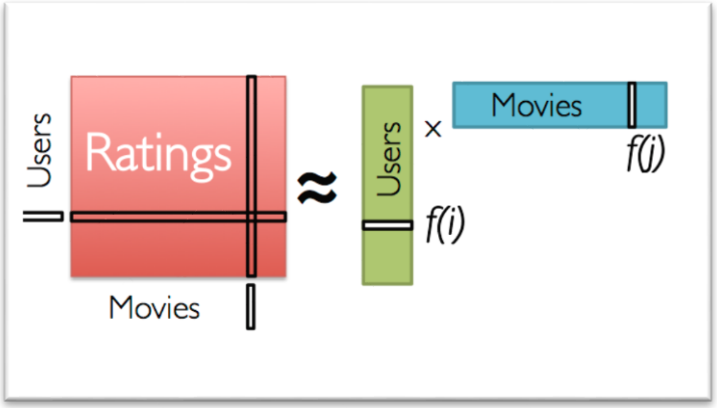




Why?

Collaborative Filtering: lack of transparency!

Why do we need content?



Who knows the «customers who bought…»?

Collaborative Filtering: poor explanations!

Why do we need content?

Because a relevant part of the information spread on social media **is content!**

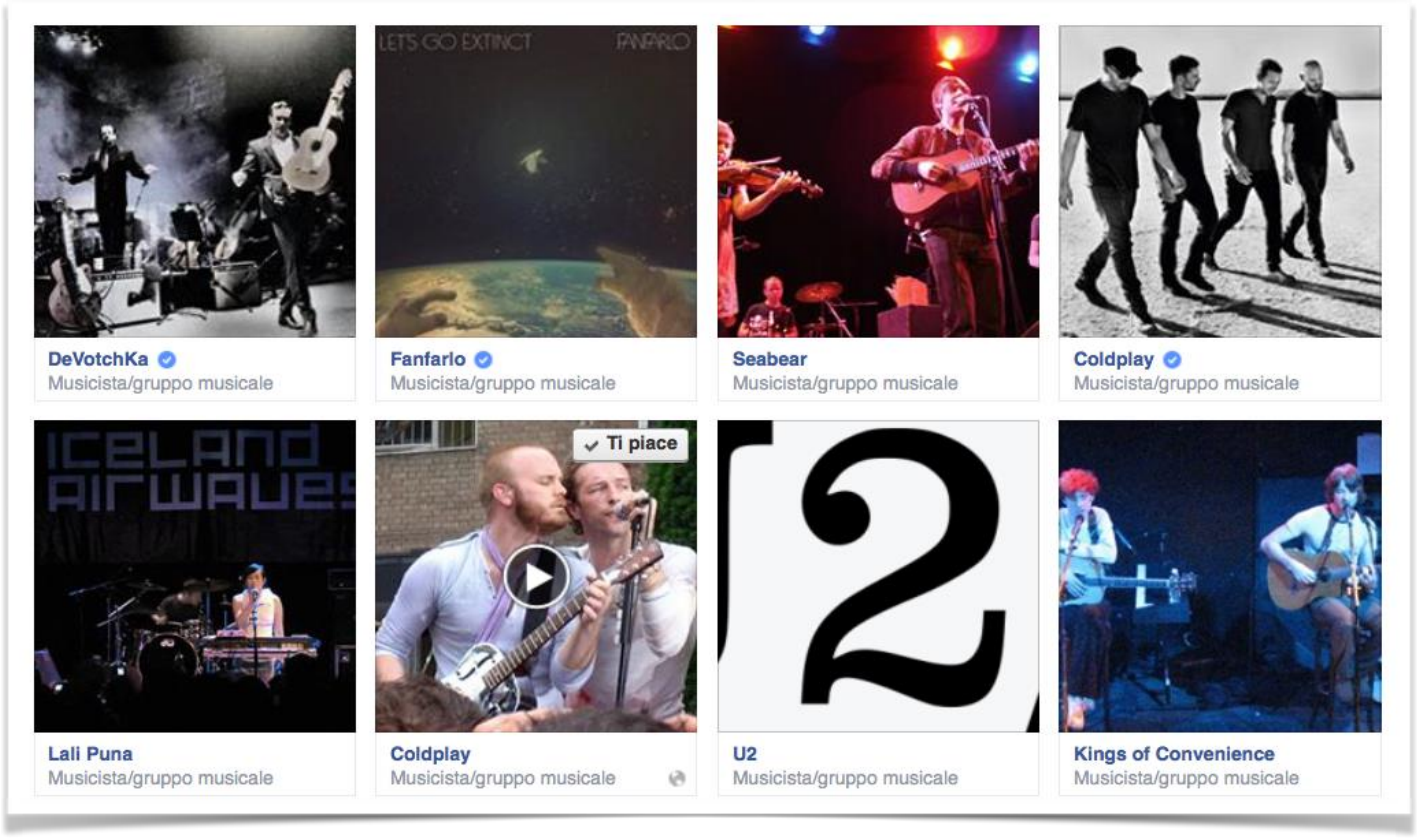
And social media really matter

Because a relevant part of the information spread on social media **is content!**

Social Media

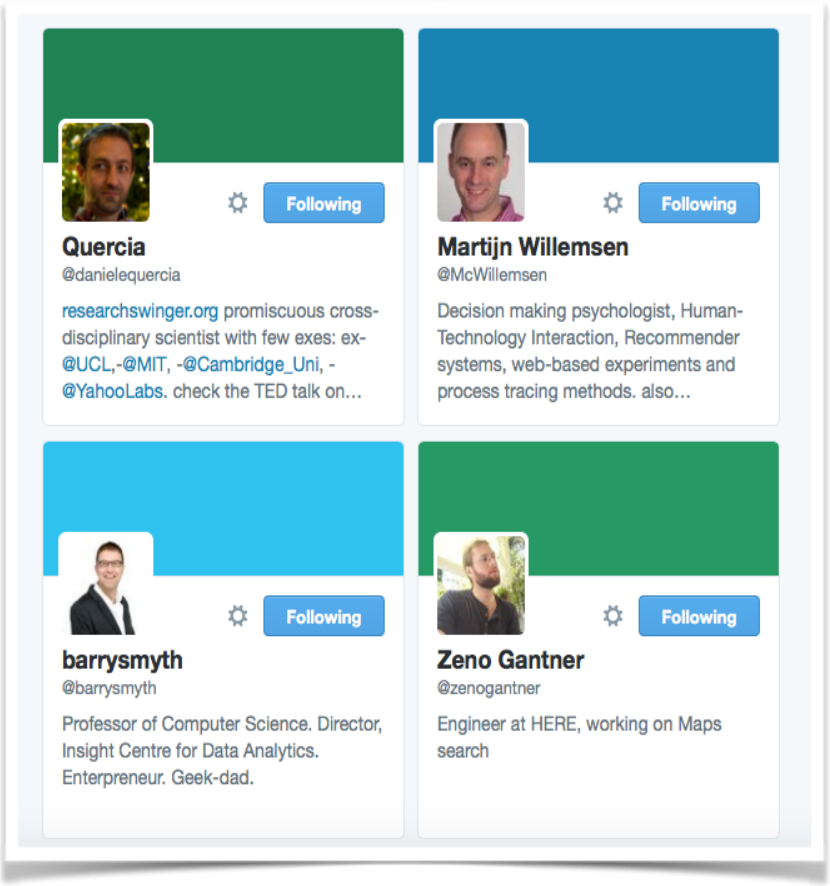
can be considered as novel **data silos**

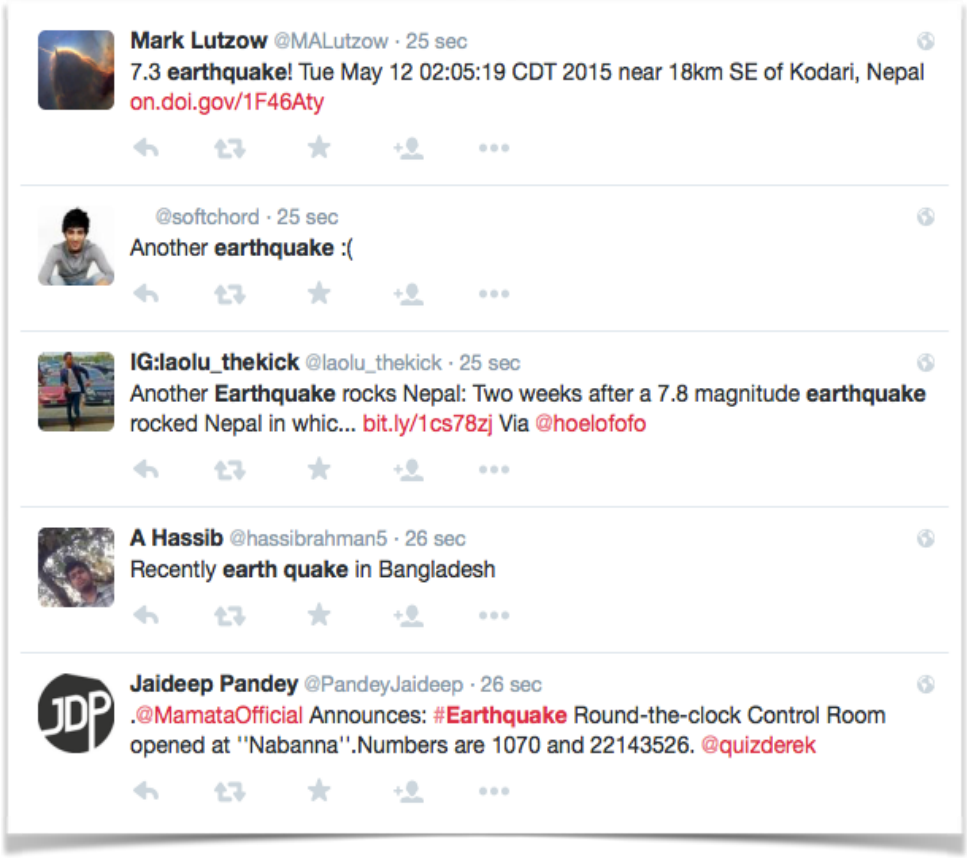
Social Media



information about **preferences**

Social Media

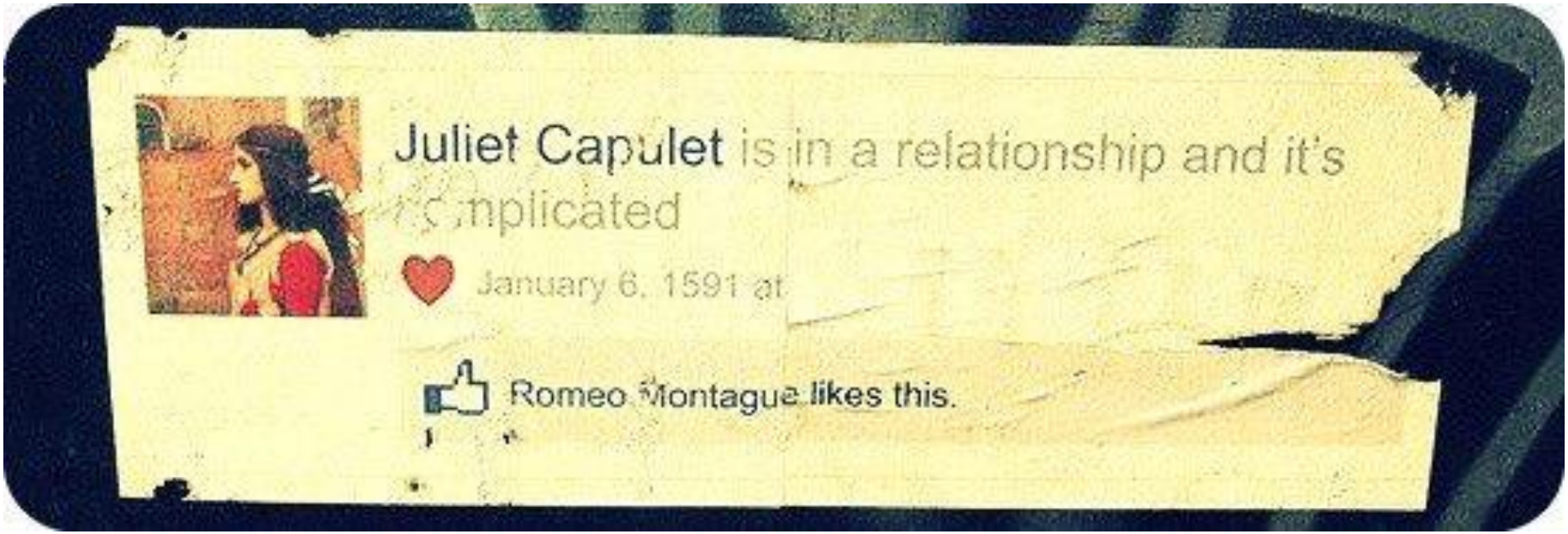




information about

People feelings and connections

Social Media



changed the rule for

user modeling and personalization

Recap #1

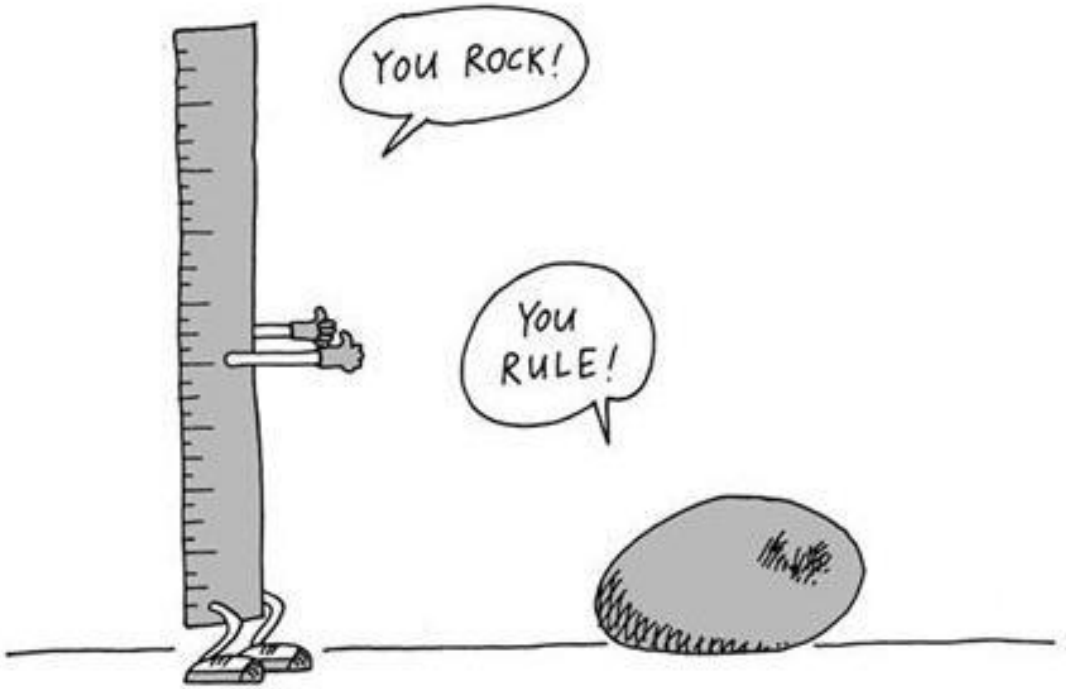
Why do we need content?

- **In general:** to extend and improve user modeling
- To exploit the information **spread on social media**
- To overcome **typical issues of collaborative filtering** and matrix factorization

Why?

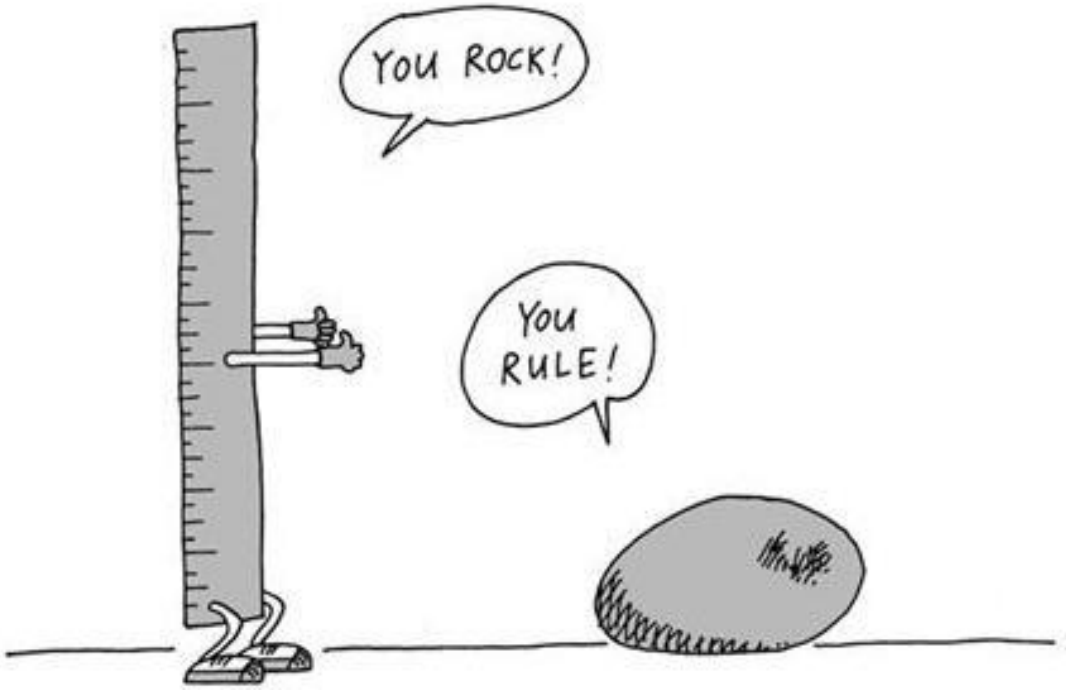
Why do we need **semantics**?

Why do we need semantics?



A **deep comprehension** of the information conveyed by textual content is crucial to **improve the quality of user profiles** and the **effectiveness of intelligent information access platforms.**

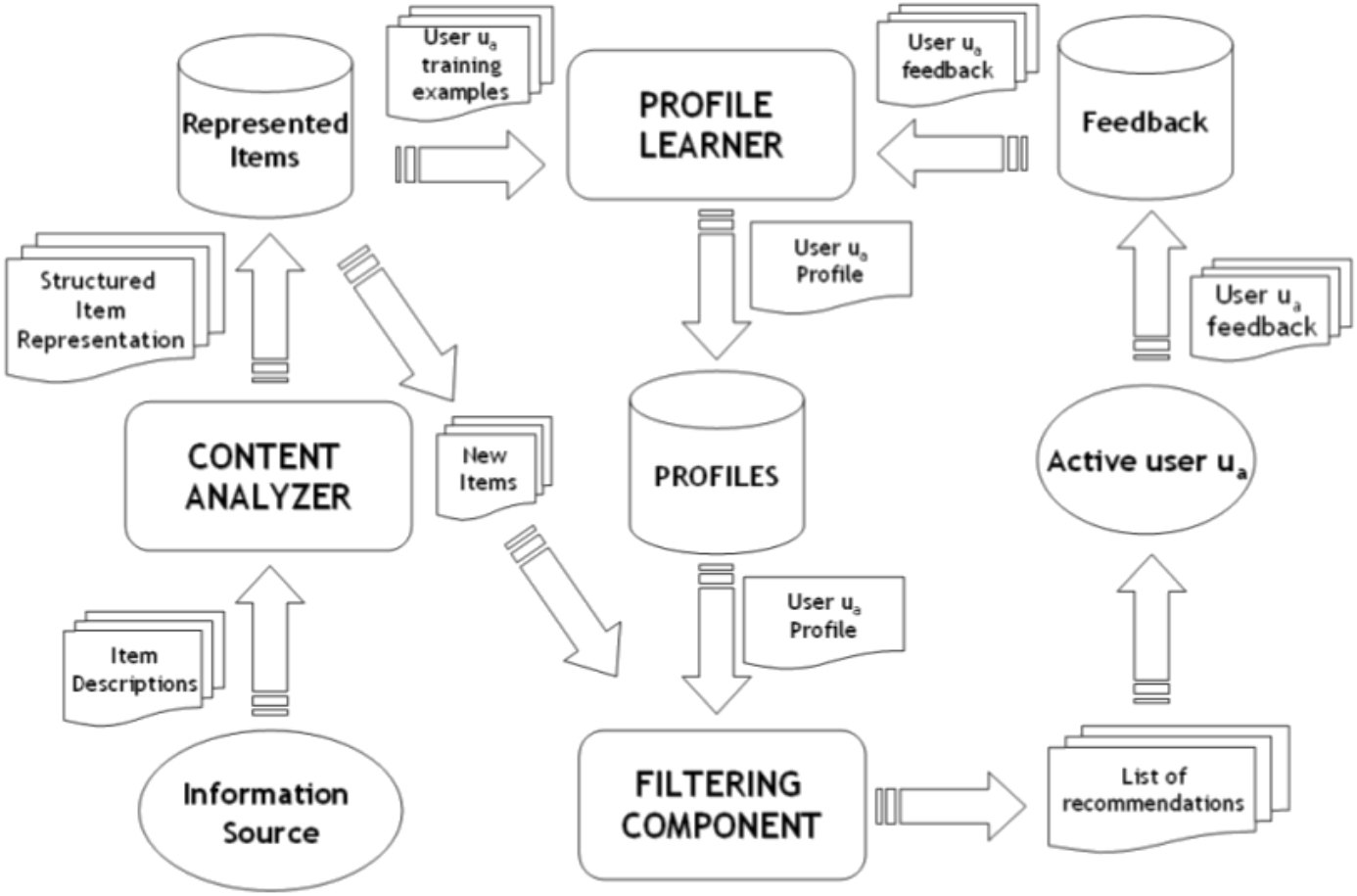
Why do we need semantics?



…some scenarios can be more convincing ☺

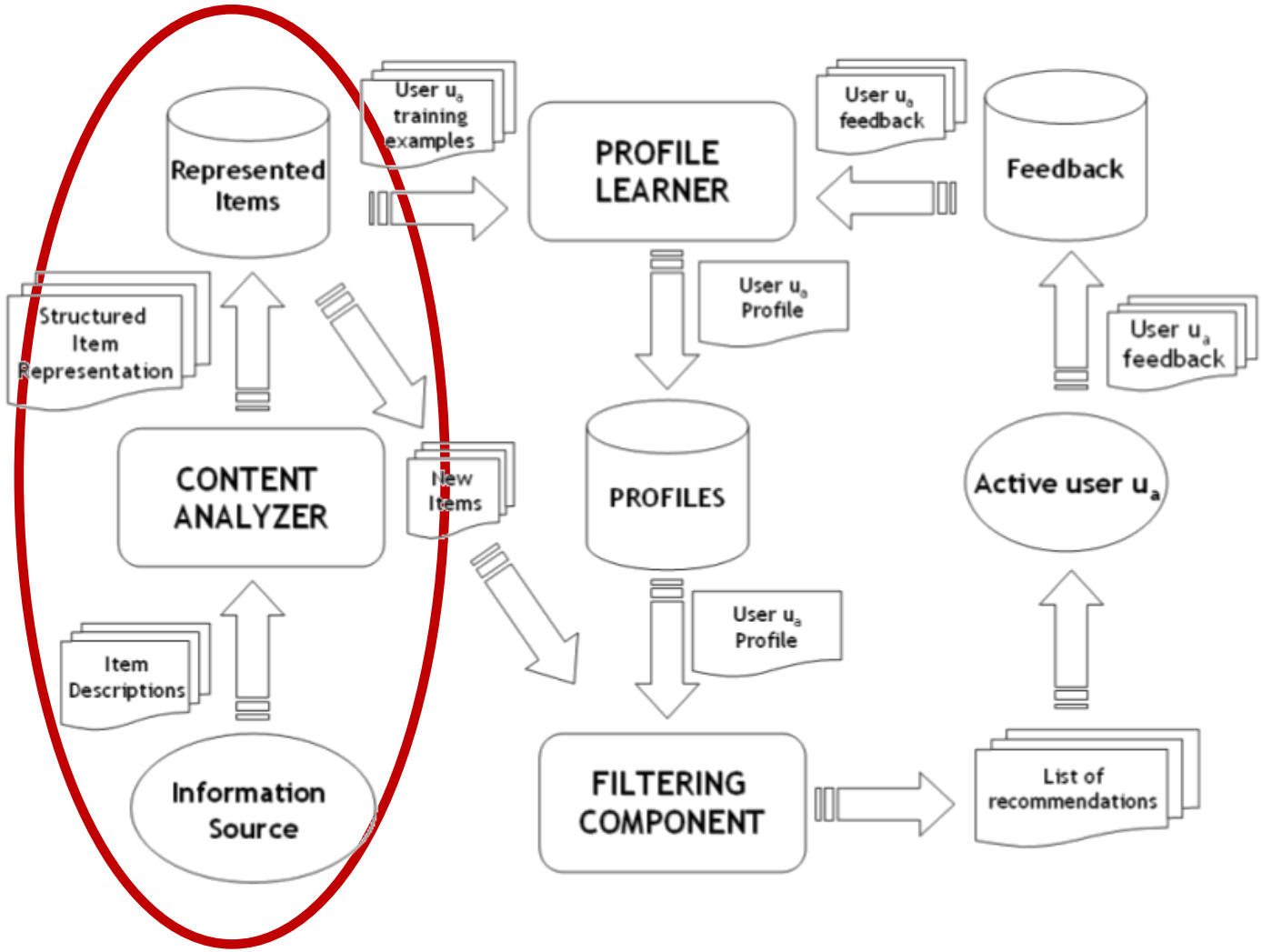
(But we need some basics, before)

Basics: Content-based RecSys (CBRS)



P. Lops, M. de Gemmis, G. Semeraro. Content-based recommender Systems: State of the Art and Trends. In: P. Kantor, F. Ricci, L. Rokach, B. Shapira, editors, Recommender Systems Hankbook: A complete Guide for Research Scientists & Practitioners

Basics: Content-based RecSys (CBRS)



P. Lops, M. de Gemmis, G. Semeraro. Content-based recommender Systems: State of the Art and Trends. In: P. Kantor, F. Ricci, L. Rokach, B. Shapira, editors, Recommender Systems Hankbook: A complete Guide for Research Scientists & Practitioners

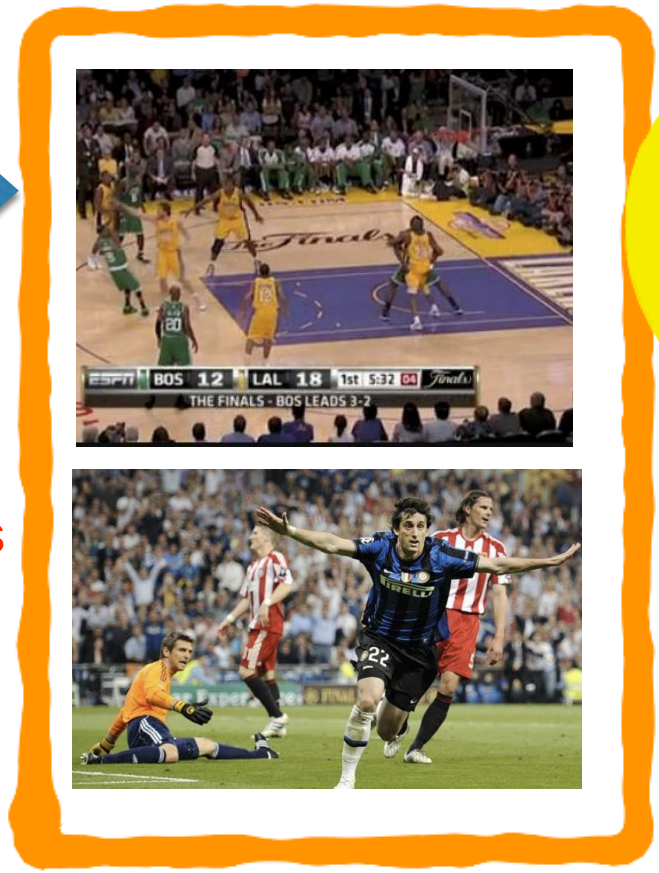
Basics: Content-based RecSys (CBRS)

Recommendation are generated by matching the features stored in the user profile with those describing the items to be recommended.

user profile **items**

Basics: Content-based RecSys (CBRS)

Recommendation are generated by matching the features stored in the user profile with those describing the items to be recommended.

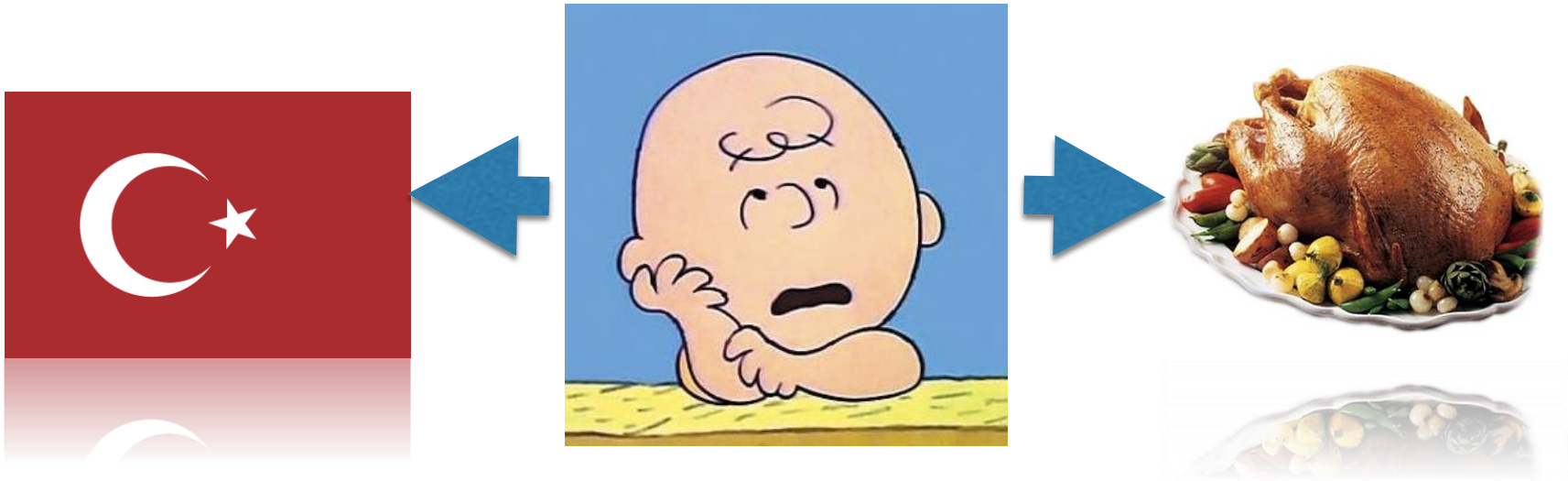


user profile

items

Lack of Semantics in User Models

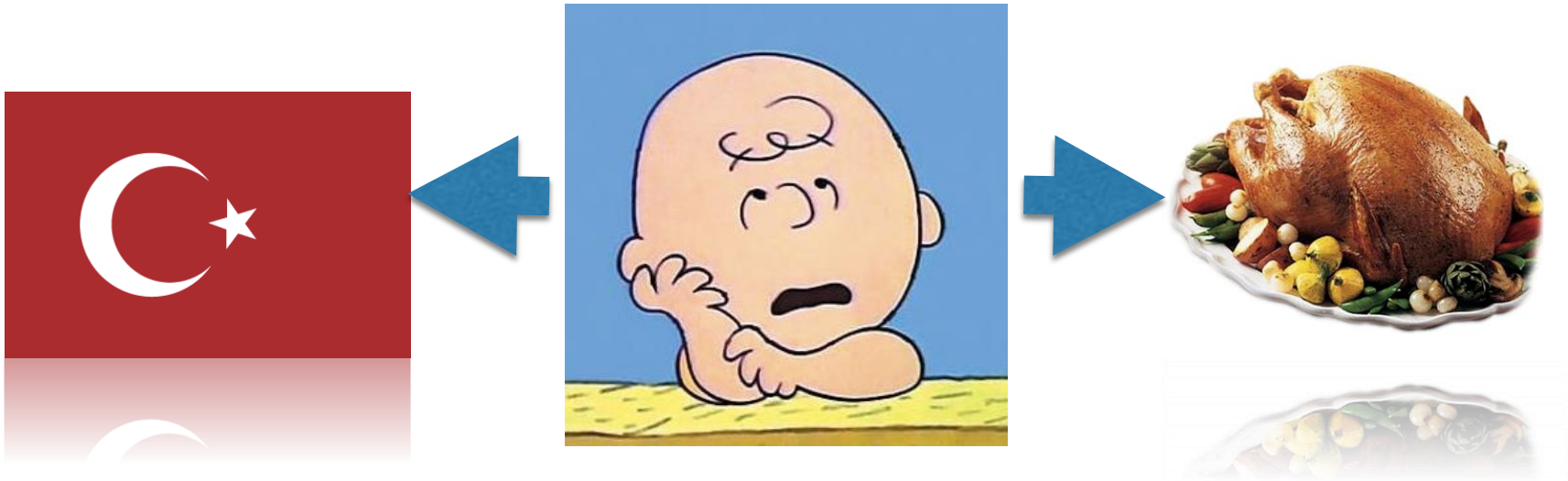
“I love turkey. It’s my choice for these #holidays!



Social Media can be helpful to avoid cold start

Lack of Semantics in User Models

“I love turkey. It’s my choice for these #holidays!

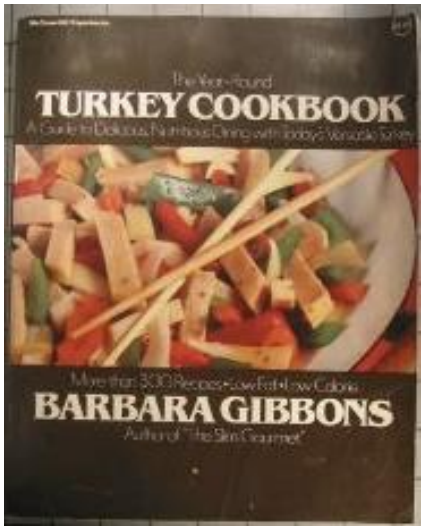


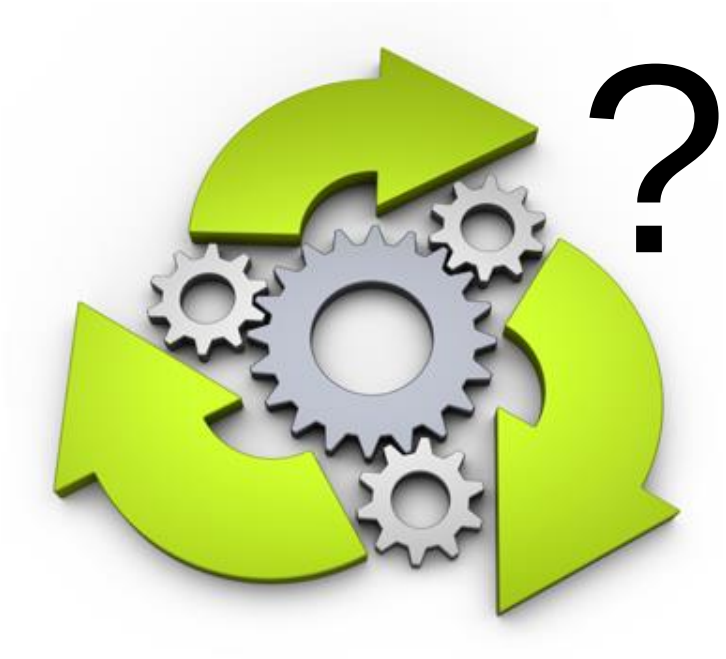
..but pure content-based representations

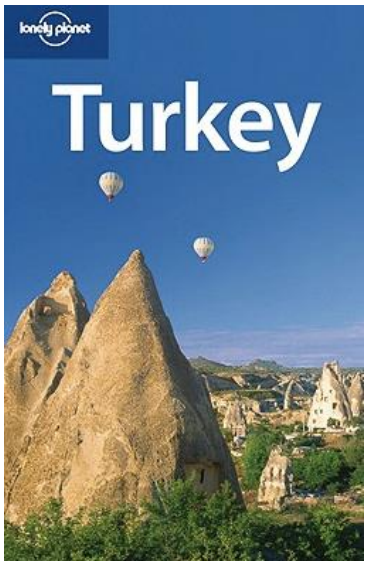
can’t handle polysemy

Lack of Semantics in User Models

“I love turkey. It’s my choice for these #holidays!







Pure Content-based Representation can easily drive a recommender systems towards failures!

Lack of Semantics in User Models

…is not only about **polysemy**

doc1
AI is a branch of computer science

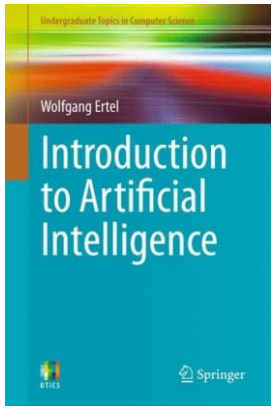
doc2
the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

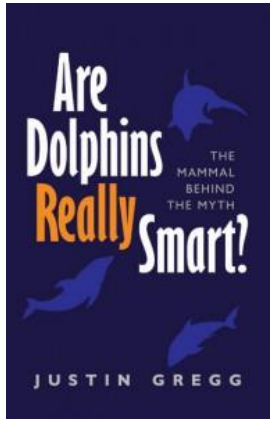
doc3
apple launches a new product…

USER PROFILE

artificial	0.11
intelligence	0.12
apple	0.20
AI	0.18
…	

?





Book recommendation

multi-word concepts

Lack of Semantics in User Models

…is not only about **polysemy**

doc1
AI is a branch of computer science

doc2
the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3
apple launches a new product…

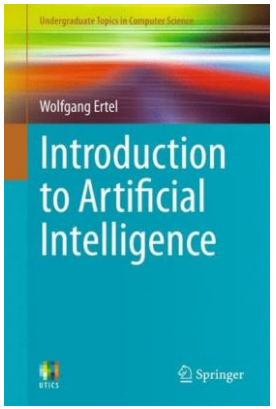
USER PROFILE

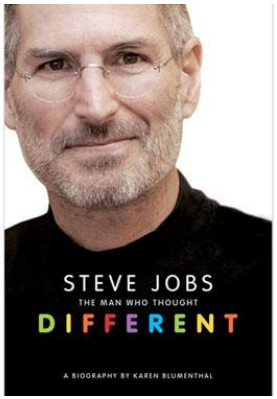
artificial	0.11
intelligence	0.12
apple	0.20
AI	0.18
…	

Most of the preferences regard AI, but due to synonymy «apple» is the most relevant feature in the profile

synonymy

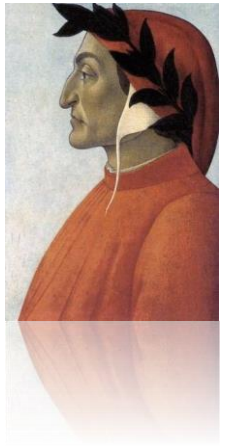
?



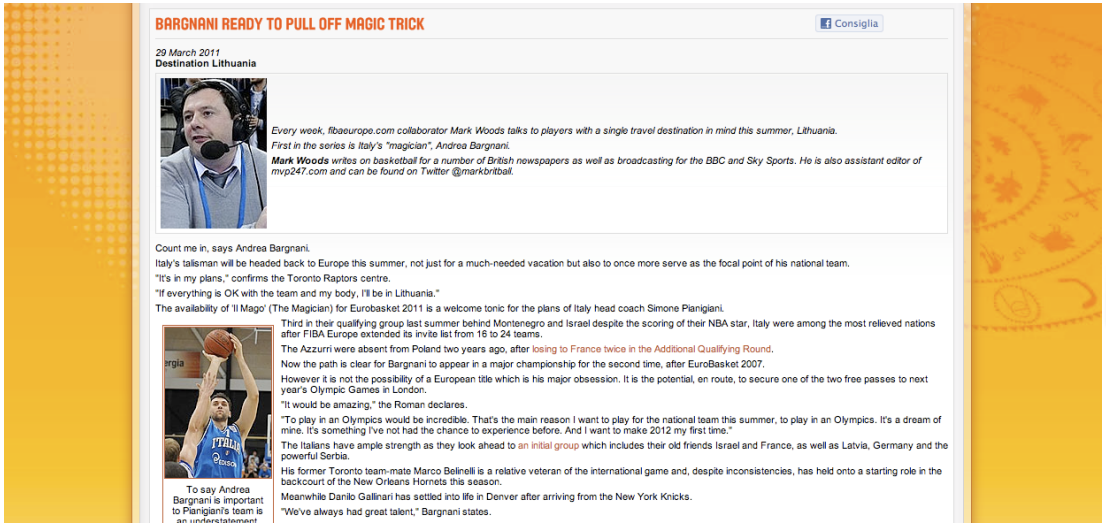


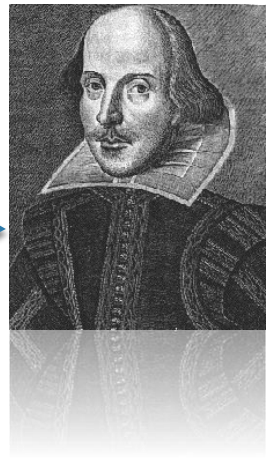
Book recommendation

Lack of Semantics in CBRS



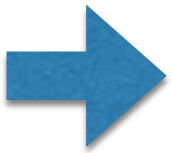
italian



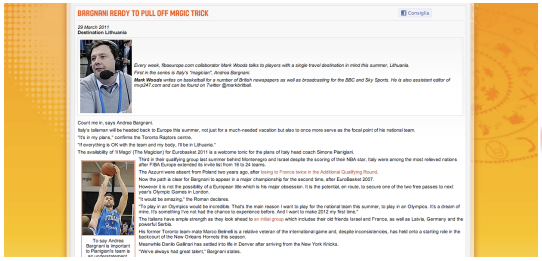


english

Lack of Semantics in CBRS



Italian-language news about basketball



English-language news about basketball

user profile

items

Lack of Semantics in CBRS

It is likely that the algorithm **is not able to suggest a (relevant) english news** since **no overlap between the features occurs!**

Italian-language news about basketball

user profile

English-language news about basketball

items

Lack of Semantics in CBRS

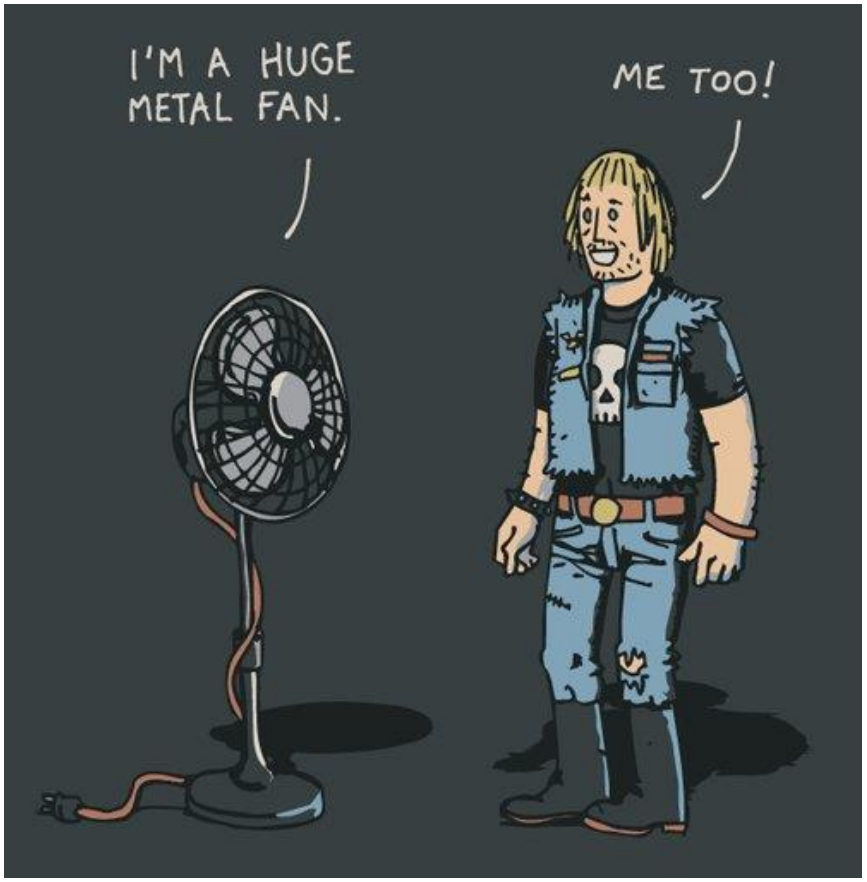
It is
algor
t
(rel
news since **no overlap between the features occurs!**

Content-based recommendations are language-dependent !

user profile **items**

Recap #2

Why do we need semantics?



Becuase language is inherently ambiguous

- **In general: to improve content representation** in intelligent information access platforms
- To avoid typical **issues of natural language representations** (polysemy, synonymy, etc.)
- To **better model user preferences**
- To **better understand** the information spread on social media
- To provide **multilingual recommendations**

How?

How to introduce **semantics**?

Information Retrieval and Filtering

Two sides of the same coin (Belkin&Croft,1992)

Information Retrieval

information need expressed through a **query**

goal: retrieve information which might be **relevant** to a user

Information Filtering

information need expressed through a **user profile**

goal: expose users to only the information that is **relevant** to them, ccording to personal profiles

It's all about searching!

[Belkin&Croft, 1992] Belkin, Nicholas J., and W. Bruce Croft. "Information filtering and information retrieval: Two sides of the same coin?." *Communications of the ACM* 35.12 (1992): 29-38.

Information Retrieval and Filtering

Two sides of the same coin (Belkin&Croft,1992)

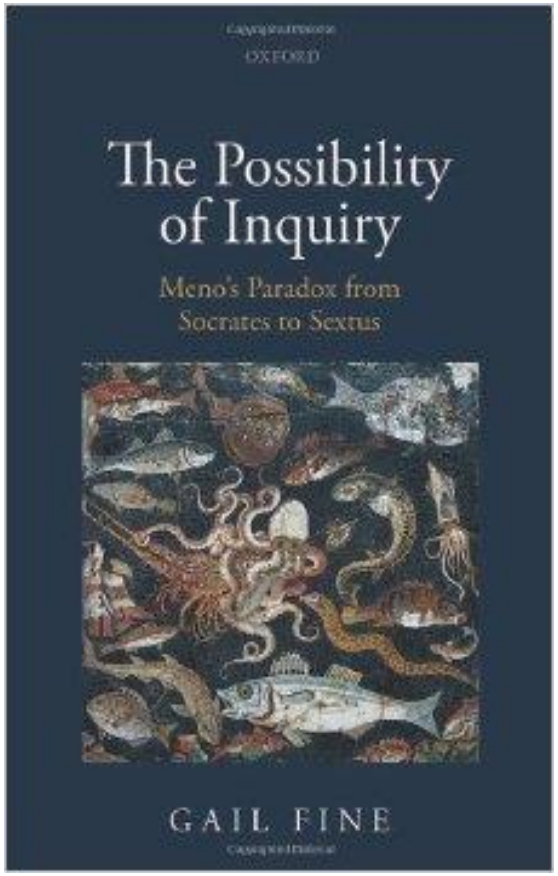
techniques for **semantics-aware content representation** and **semantic user modeling** can be **easily borrowed from Information Retrieval** area *(and viceversa!)*

Search (and Content-based Recommendation) is not so simple as it might seem

Meno's Paradox of Inquiry:

Meno: and how will you enquire, Socrates, into that which you do not know? **What will you put forth as the subject of enquiry**? And if you find what you want, **how will you know that this is the thing you did not know**?

Socrates: I know, Meno, what you mean; but just see what a tiresome dispute you are introducing. **You argue that a man cannot search either for what he knows or for what he does not know**; if he knows it, there is no need to search; and if not, he cannot; he does not know the very subject about which he is to search.



Plato Meno 80d-81a

http://www.gutenberg.org/etext/1643

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

We need **some «intelligent» support (as intelligent information access technologies)**

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

We need **some «intelligent» support (as intelligent information access technologies)**

We need to **better understand and represent the content**

Meno's question at our times:
the "vocabulary mismatch" problem (revisited)

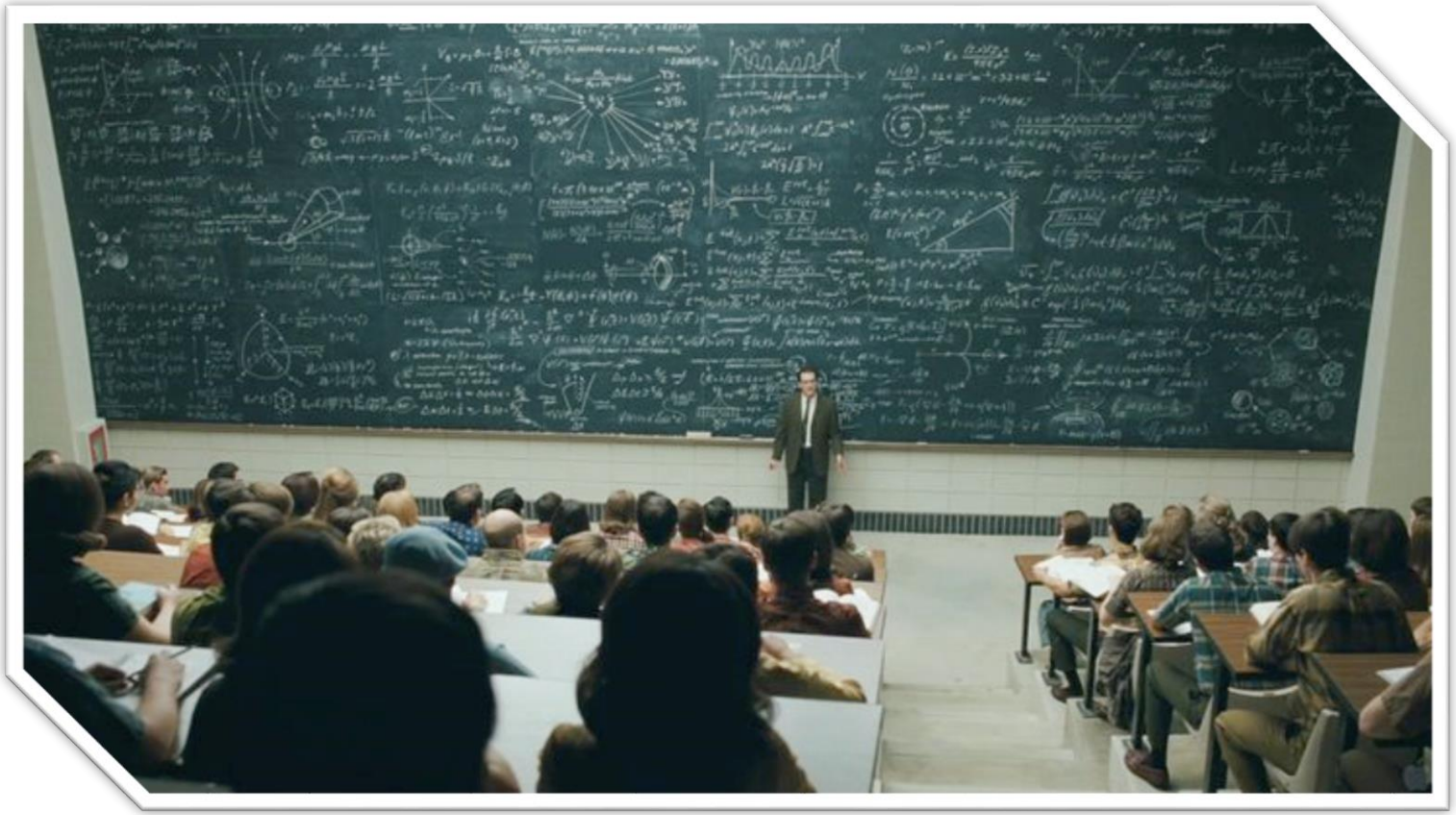
How to discover the **concepts** that connect us to the **the information we are seeking** (**search task**) or **we want to be exposed to** (**recommendation and user modeling tasks**) ?

We need **some «intelligent» support (as intelligent information access technologies)**

We need to **better understand and represent the content**

...before semantics

some **basics**



of **Natural Language Processing (NLP)**

How?

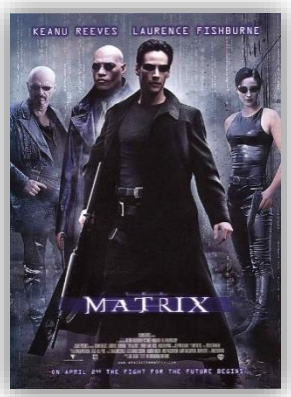
basics of NLP and keyword-based representations

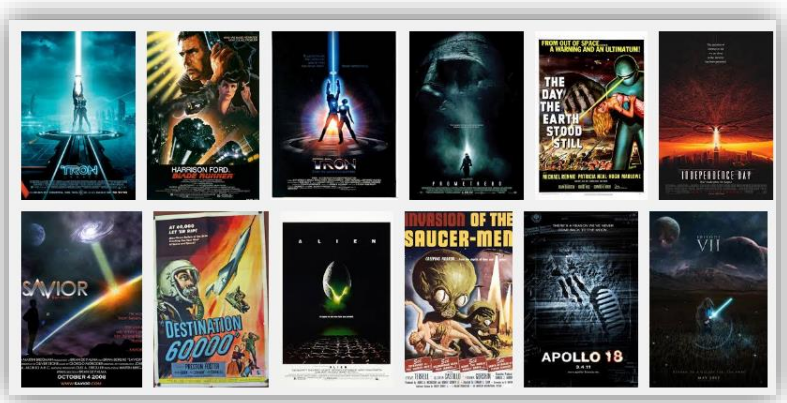
Scenario

Pierpaolo really loves the movie «The Matrix», and he asks a content-based recommender system for some suggestions.

Question

How can we **feed the algorithm with some textual features** related to the **movie to build a (content-based) profile** and provide recommendations?





Recommendation Engine

Scenario

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well-received by critics [10][11] and won four Academy Awards as well as other accolades including BAFTA

The Matrix

Theatrical release poster

the plot can be a **rich source** of **content-based features**

Scenario

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well-received by critics [10][11] and won four Academy Awards as well as other accolades including BAFTA

The Matrix

Theatrical release poster

the plot can be a **rich source** of **content-based features**

…but we need to **properly process it** through a pipeline of **Natural Language Processing** techniques

Basic NLP operations

- **normalization** strip unwanted characters/markup (e.g. HTML/XML tags, punctuation, numbers, etc.)
- **tokenization** break text into tokens
- **stopword removal** exclude common words having little semantic content
- **lemmatization** reduce inflectional/variant forms to base form (lemma in the dictionary), *e.g. am, are, is → be*
- **stemming** reduce terms to their "roots", *e.g. automate(s), automatic, automation all reduced to* ***automat***

Example

The Matrix is a 1999 American-Australian neo-noir science fiction action film written and directed by the Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

Example

The Matrix is a 1999 American-Australian neo-noir science fiction action film written and directed by the Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

normalization

Example

The Matrix is a 1999 American Australian neo noir science fiction action film written and directed by the Wachowskis starring Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix created by sentient machines to subdue the human population while their bodies heat and electrical activity are used as an energy source Computer programmer Neo learns this truth and is drawn into a rebellion against the machines which involves other people who have been freed from the dream world

tokenization

Tokenization issues

compound words

- science-fiction: break up **hyphenated** sequence?
- Keanu Reeves: **one token or two**? How do you decide it is one token?

numbers and dates

- 3/20/91 Mar. 20, 1991 20/3/91
- 55 B.C.
- (800) 234-2333

Tokenization issues

language issues

- German noun compounds not segmented
  Lebensversicherungsgesellschaftsangestellter means
  life insurance company employee
- Chinese and Japanese have no spaces between words (not always guaranteed a unique tokenization)

  莎拉波娃现在居住在美国东南部的佛罗里达

- Arabic (or Hebrew) is basically written right to left, but with certain items like numbers written left to right

  استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

  Algeria achieved its independence in 1962 after 132 years of French occupation

Example

The Matrix is a 1999 American Australian neo noir science fiction action film written and directed by the Wachowskis starring Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix created by sentient machines to subdue the human population while their bodies heat and electrical activity are used as an energy source Computer programmer Neo learns this truth and is drawn into a rebellion against the machines which involves other people who have been freed from the dream world

stopword removal

Example

The Matrix is a 1999 American Australian neo noir science fiction action film written and directed by the Wachowskis starring Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix created by sentient machines to subdue the human population while their bodies heat and electrical activity are used as an energy source Computer programmer Neo learns this truth and is drawn into a rebellion against the machines which involves other people who have been freed from the dream world

stopword removal

Example

The Matrix is a 1999 American Australian neo noir science fiction action film written and directed by the Wachowskis starring Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix created by sentient machines to subdue the human population while their bodyies heat and electrical activity are used as an energy source Computer programmer Neo learns this truth and is drawn into a rebellion against the machines which involves other people who have been freed from the dream world

lemmatization

Example

Matrix 1999 American Australian neo noir science fiction action film write direct Wachowskis star Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving Joe Pantoliano depict dystopian future reality perceived human simulate reality call Matrix create sentient machine subdue human population body heat electrical activity use energy source Computer programmer Neo learn truth draw rebellion against machine involve people free dream world

next step: to give a weight to each feature (e.g. through TF-IDF)

Weighting features: TF-IDF

terms frequency – inverse document frequency best known weighting scheme in information retrieval.

Weight of a term as product of **tf weight** and **idf weight**

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log(N / \mathrm{df}_t)$$

tf **number of times** the term occurs in the document

idf depends on **rarity** of a term in a collection

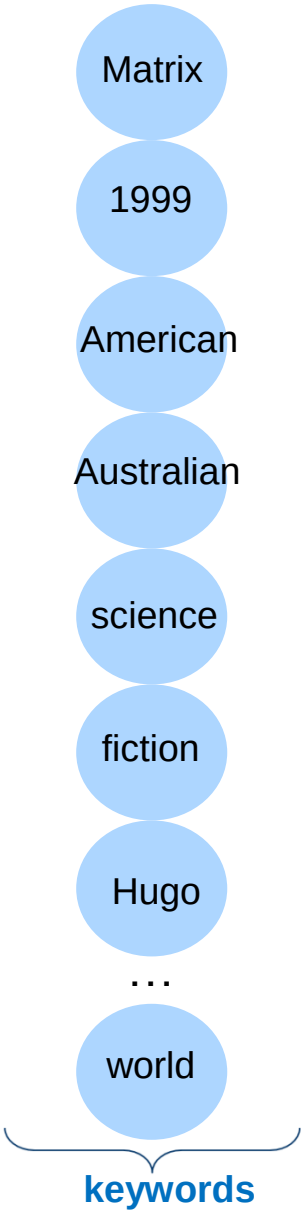
tf-idf increases with the number of occurrences within a document, and with the rarity of the term in the collection.

Example

Matrix 1999 American Australian neo noir science fiction action **film** write direct Wachowskis star Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving Joe Pantoliano depict **dystopian** future reality perceived human simulate reality call Matrix create sentient machine subdue human population body heat electrical activity **use** energy source Computer programmer Neo learn truth draw **rebellion** against machine involve people free dream world

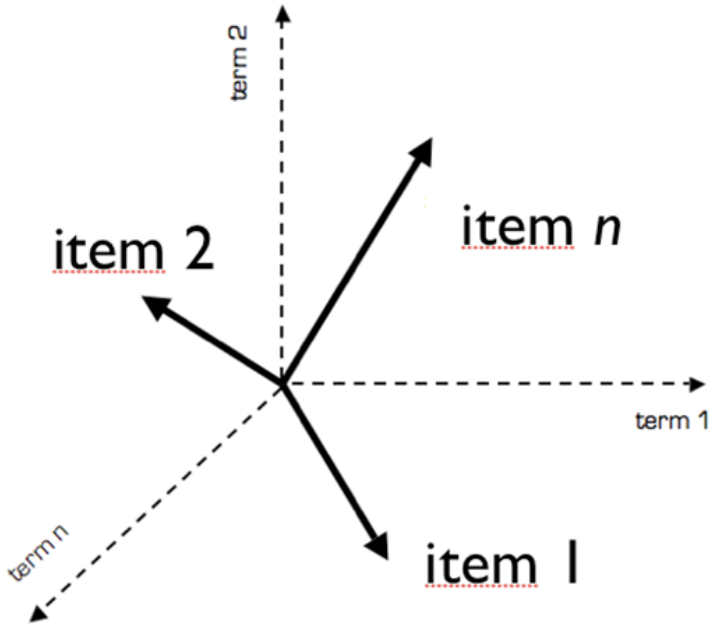
green=high IDF
red=low IDF

The Matrix representation



a portion of Pierpaolo's
content-based profile

Vector Space Model (VSM)



given a set of *n* **features** (vocabulary)

$$f = \{f_1, f_2, \ldots, f_n\}$$

given a set of *M* items, each item *I* represented as a point in a ***n*-dimensional vector space**

$$I = (w_{f1}, \ldots . w_{fn})$$

w_{fi} is the **weight** of feature *i* in the item

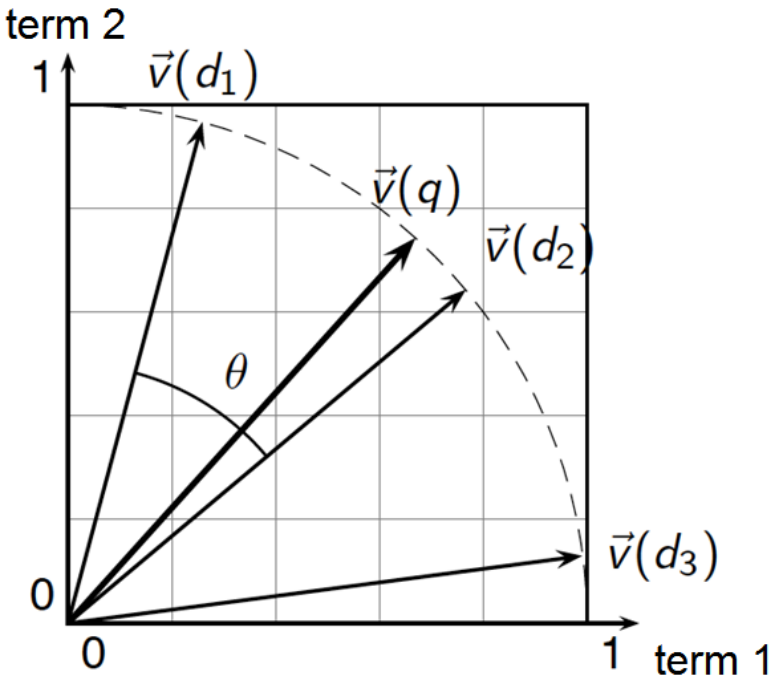
Similarity between vectors

cosine similarity

dot product

unit vectors

$$\cos(\vec{I},\vec{J}) = \frac{\vec{I}\bullet\vec{J}}{\left|\vec{I}\right|\left|\vec{J}\right|} = \frac{\vec{I}}{\left|\vec{I}\right|}\bullet\frac{\vec{J}}{\left|\vec{J}\right|} = \frac{\sum_{i=1}^{|V|} I_i J_i}{\sqrt{\sum_{i=1}^{|V|} I_i^2}\sqrt{\sum_{i=1}^{|V|} J_i^2}}$$



The Matrix representation

Matrix

1999

American

Australian

science

fiction

Hugo

…

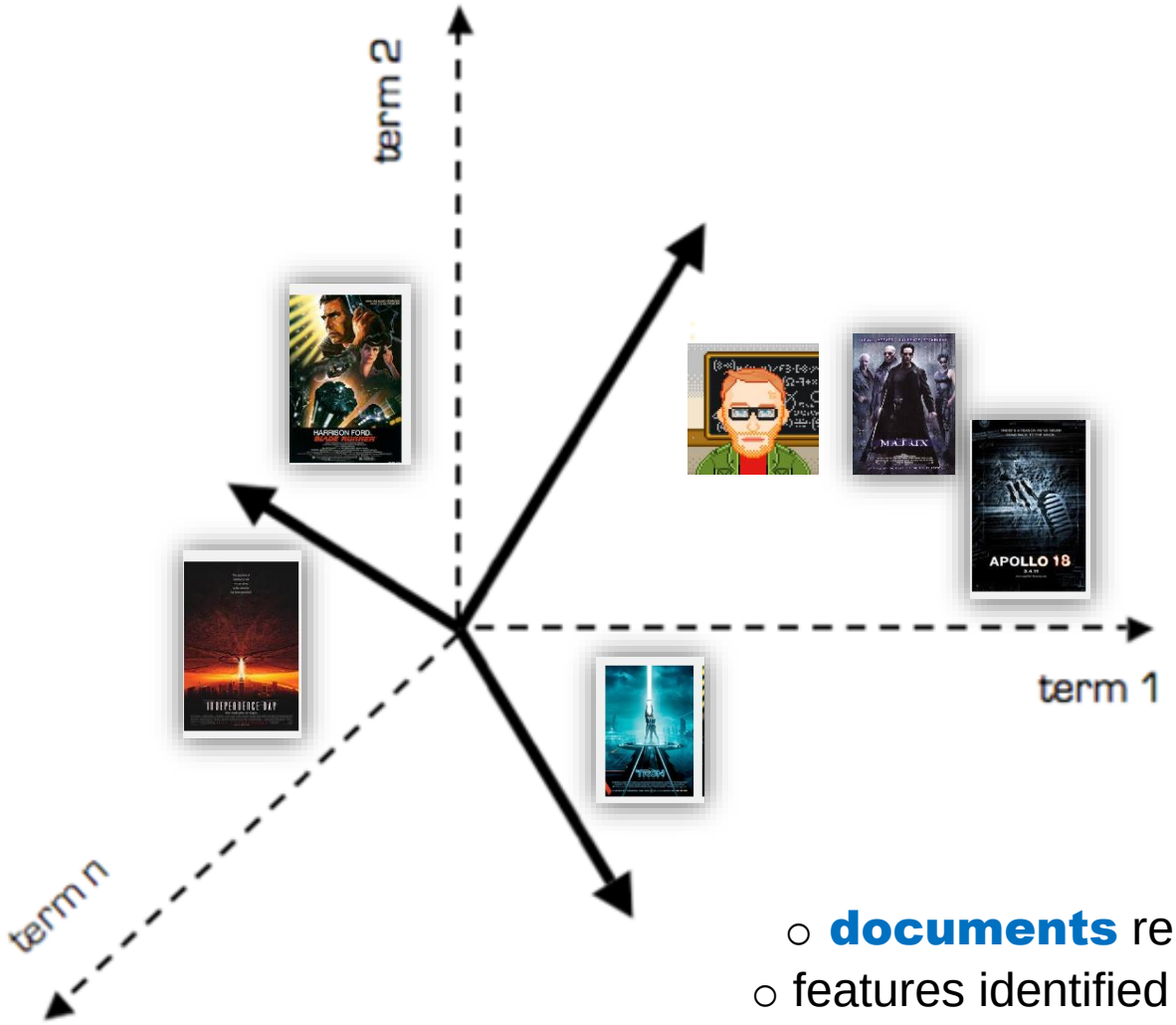
world

keywords

a portion of Pierpaolo's **content-based profile**

given a content-based profile, we can **easily build a basic recommender system** through **Vector Space Model** and **similarity measures**

Basic Content-based Recommendations



- **documents** represented as **vectors**
- features identified through **NLP operations**
- **features** weigthed using **tf-idf**
- **cosine measure** for computing similarity between vectors

Basic Content-based Recommendations
Drawbacks

Matrix

1999

American

Australian

science

fiction

Hugo

…

world

a portion of Pierpaolo's
content-based profile

Recommendation:
Notre Dame de Paris,
by Victor Hugo

Why?
Entities as «Hugo Weaving» were not modeled

Basic Content-based Recommendations
Drawbacks

Matrix

1999

American

Australian

science

fiction

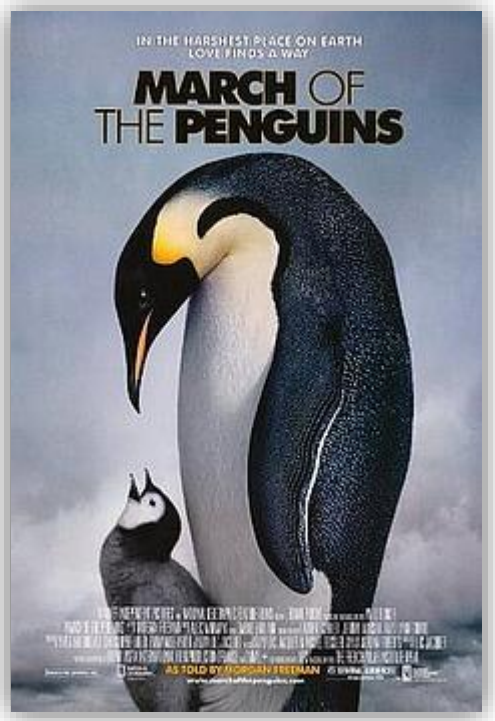
Hugo

…

world

a portion of Pierpaolo's
content-based profile

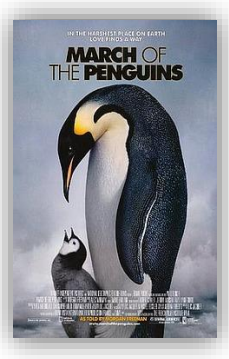
Recommendation:
The March of Penguins



Why?
More complex concepts as «science fiction» were not modeled as single features

Basic Content-based Recommendations

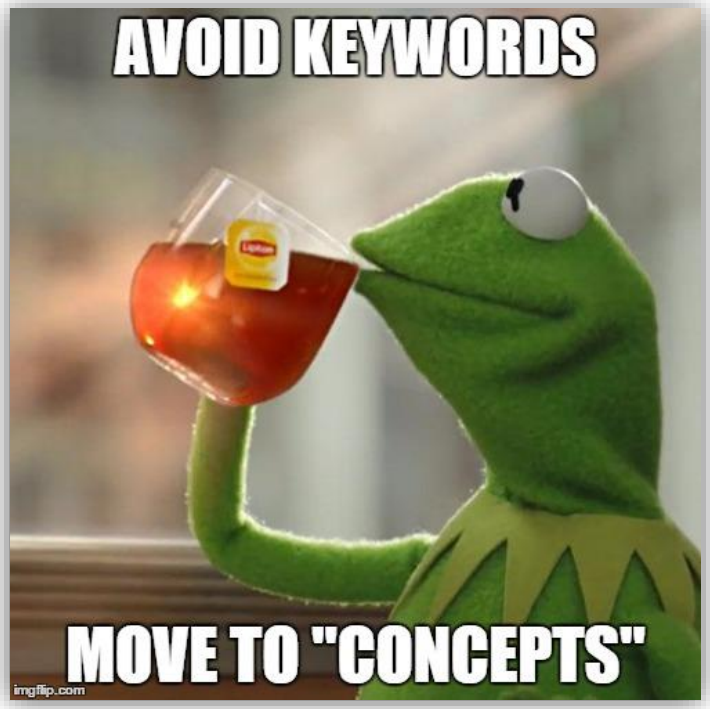
Vision



Basic Content-based Recommendations

Vision

Bad recommendations



Recap #3

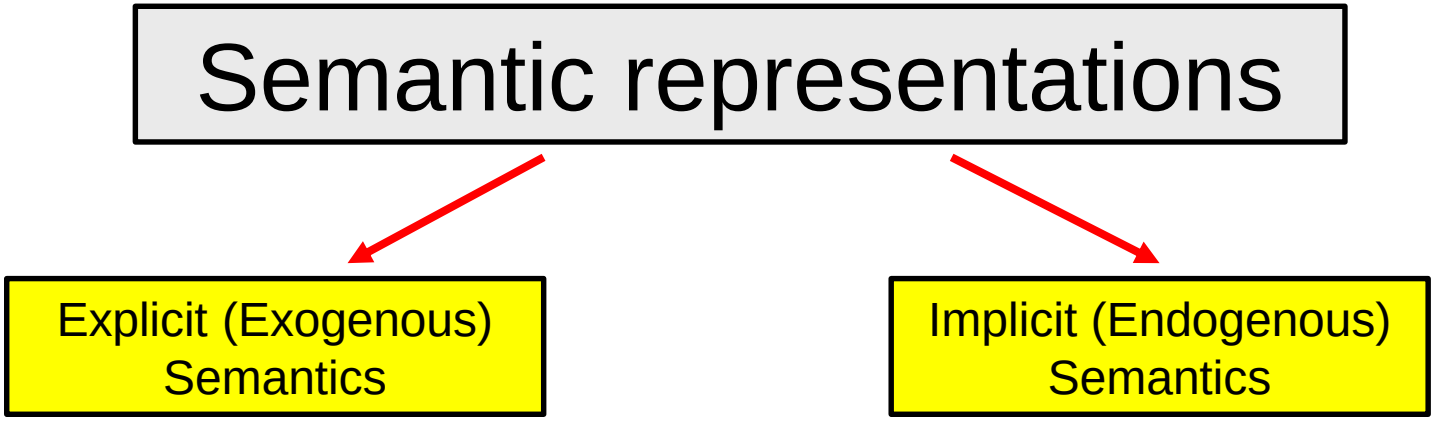
basics of **NLP** and **keyword**-based representation

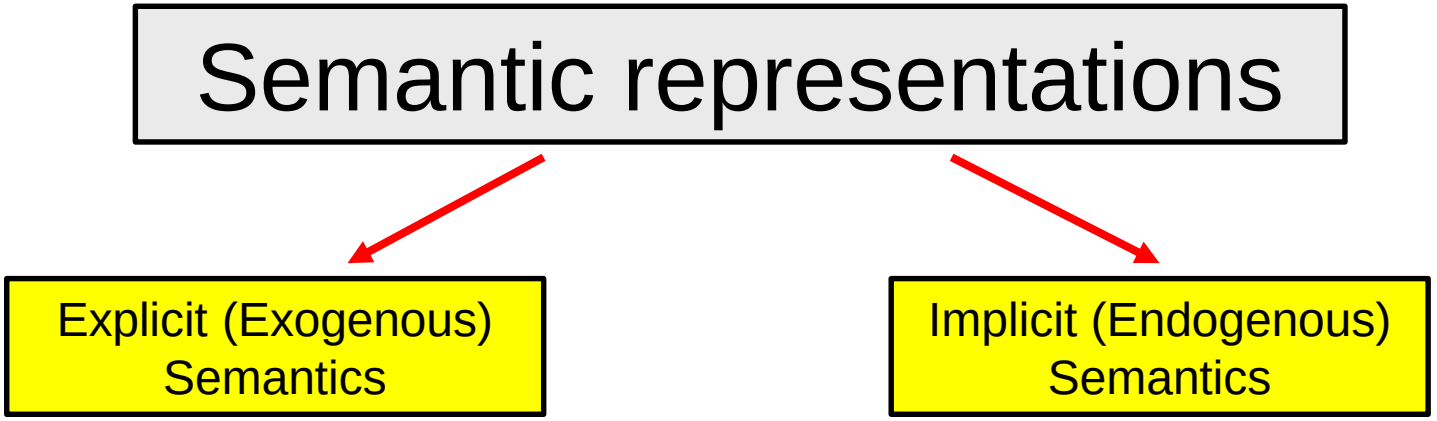
- **Natural Language Processing techniques necessary** to build a content-based profile
- **basic content-based algorithms can be easily built** through TF-IDF
- **keyword-based representation too poor** and can drive to bad modeling of preferences (and bad recommendations)
- **we need to shift from keywords to concepts**

How?

Semantics-aware Content Representation

Semantic representations


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics



top-down

approaches based on the integration of **external knowledge** for representing content. Able to provide the **linguistic, cultural** and **backgroud knowledge** in the **content representation**

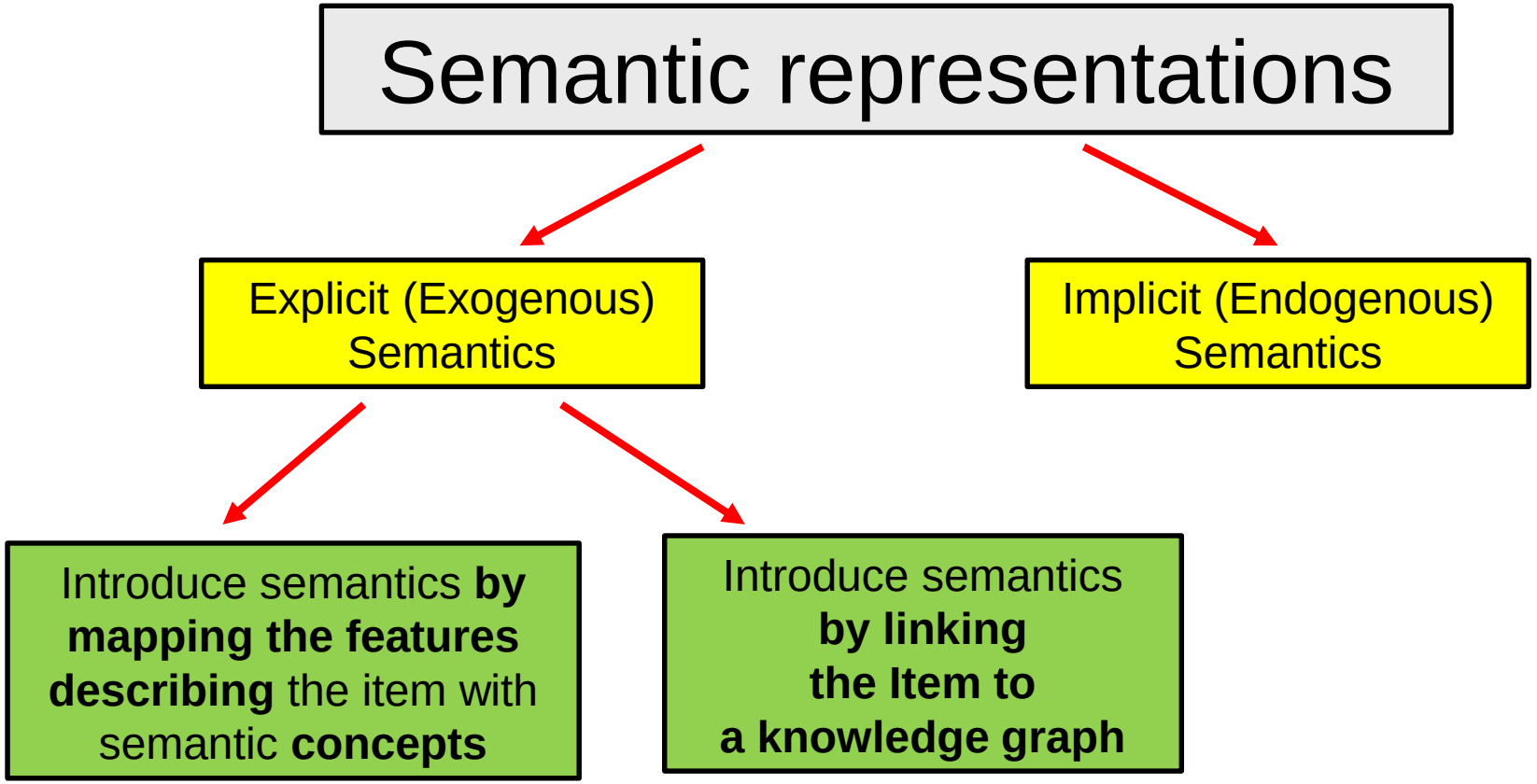
Semantic representations

top-down

approaches based on the integration of **external knowledge** for representing content. Able to provide the **linguistic, cultural** and **backgroud knowledge** in the **content representation**

bottom-up

approaches that determine the **meaning** of a word by analyzing the rules of its **usage** in the context of **ordinary and concrete language behavior**


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking the Item to a knowledge graph**

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

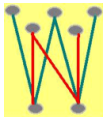
Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the item to a knowledge graph**

Word Sense Disambiguation

Entity Linking

…….



Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

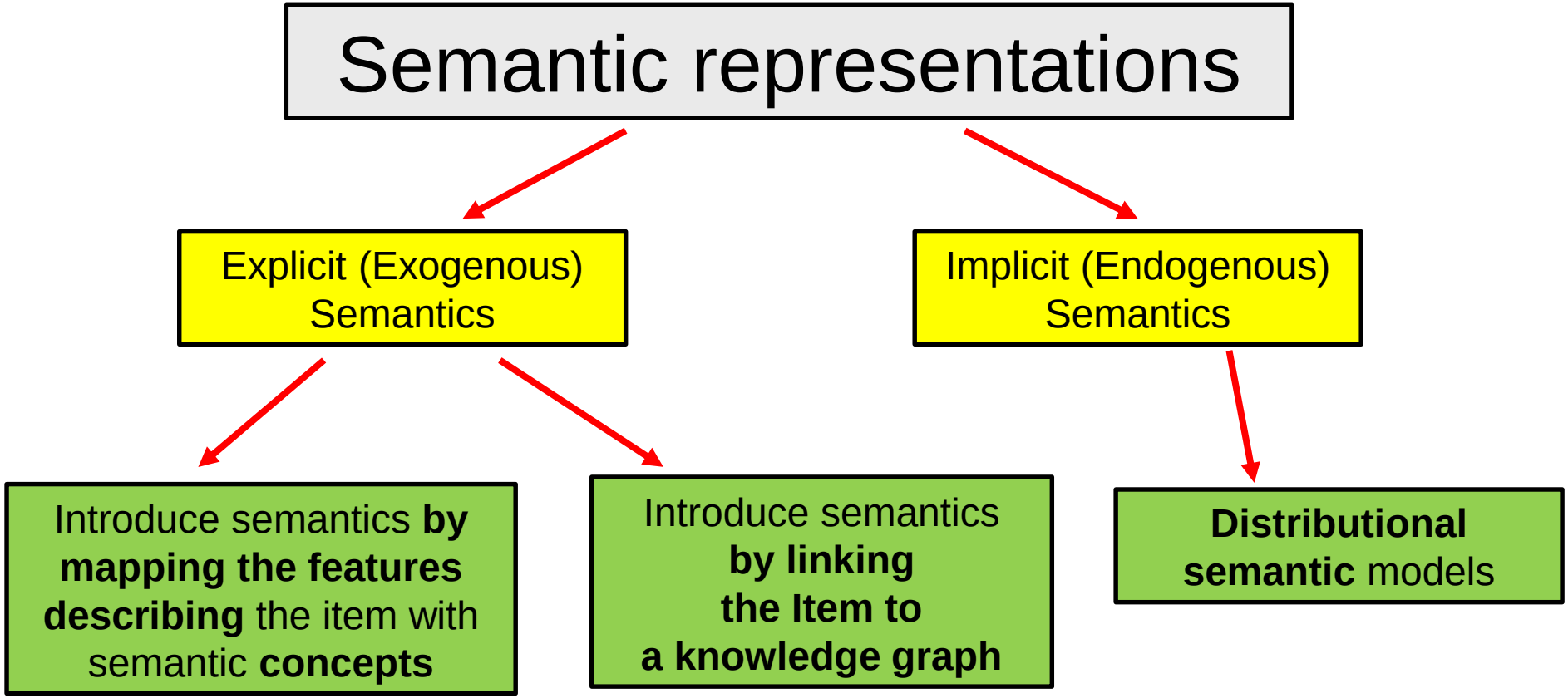
Introduce semantics **by linking the Item to a knowledge graph**

Ontologies

Linked Open Data

DBpedia

.....


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking the Item to a knowledge graph**
Distributional semantic models

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

Random Indexing

Word2Vec

……

How?

Encoding **exogenous semantics**
(top-down approaches)

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

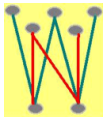
Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Word Sense Disambiguation

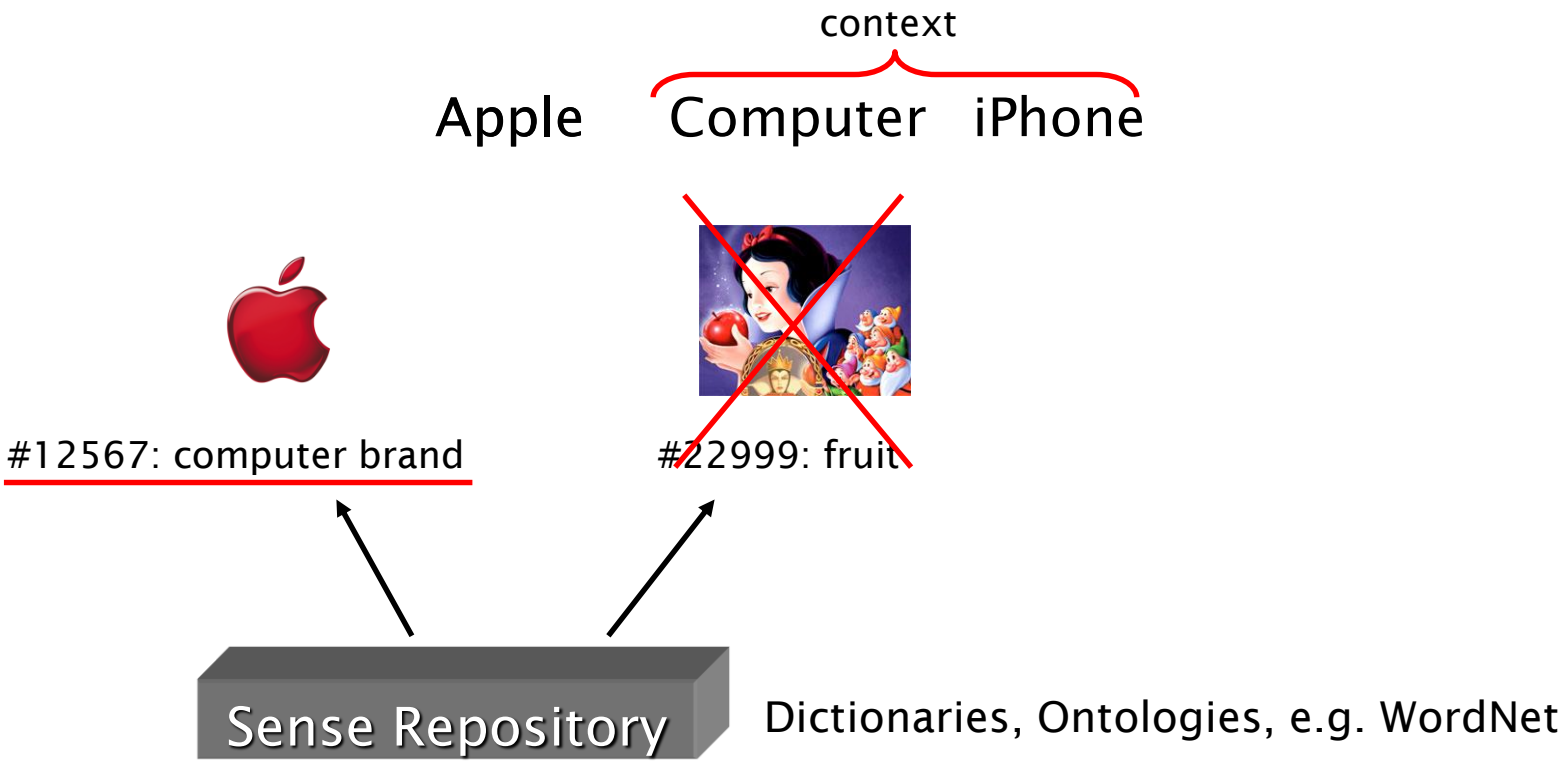
Entity Linking

…….



Word Sense Disambiguation (WSD)
using linguistic knowledge base

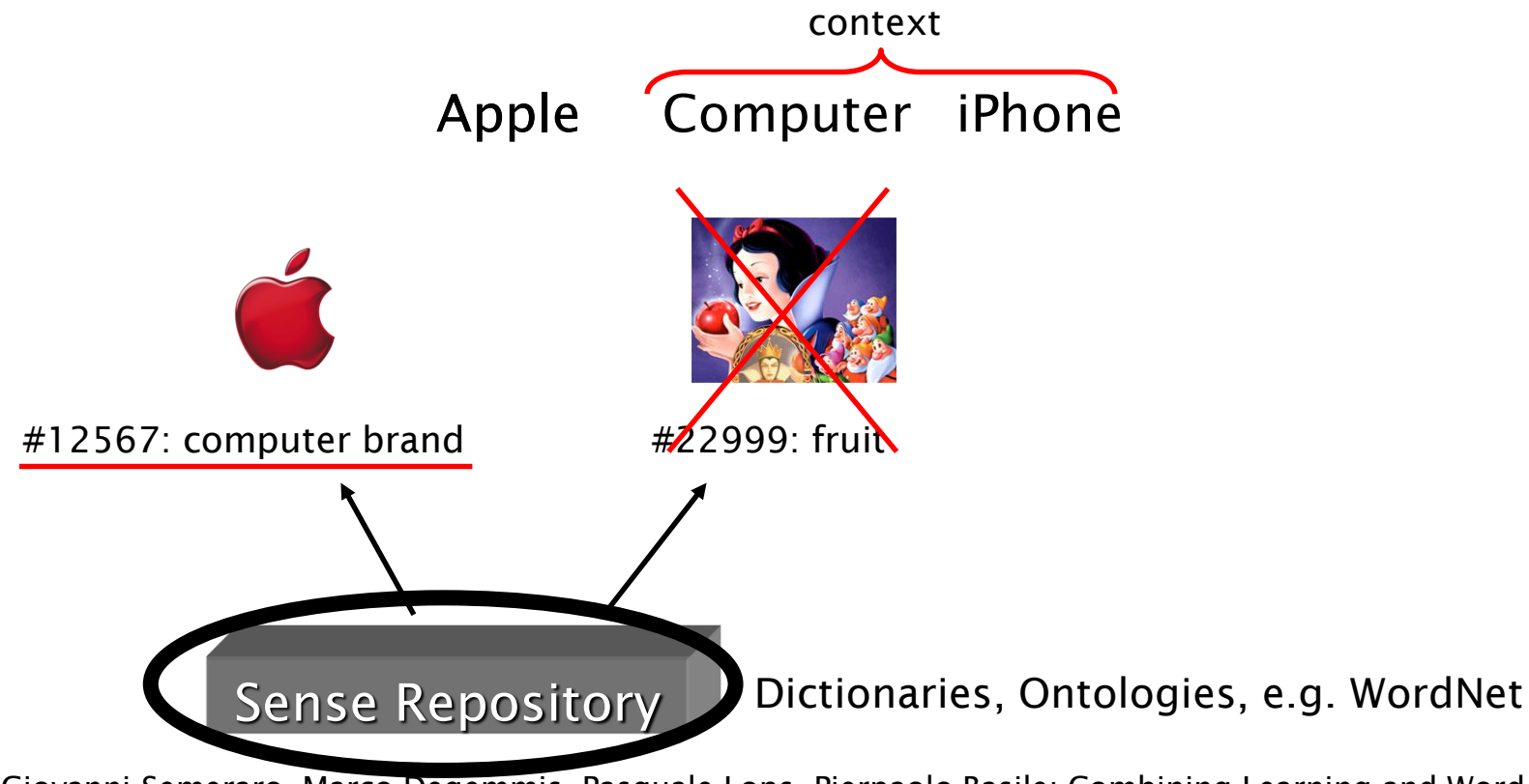
WSD selects the proper meaning, i.e. **sense**, for a word in a text by taking into account the **context** in which it occurs



Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Pierpaolo Basile: Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. IJCAI 2007: 2856-2861

Word Sense Disambiguation (WSD)
using linguistic knowledge base

WSD selects the proper meaning, i.e. **sense**, for a word in a text by taking into account the **context** in which it occurs



Giovanni Semeraro, Marco Degemmis, Pasquale Lops, Pierpaolo Basile: Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. IJCAI 2007: 2856-2861

Sense Repository

WordNet semantic network [*]

https://wordnet.princeton.edu

WordNet groups words into sets of synonyms called **synsets**

It contains **nouns**, **verbs**, **adjectives**, **adverbs**

Word Meanings	Word Forms					
	F_1	F_2	F_3	...	...	F_n
M_1	V(1,1)	V(2,1)				
M_2		V(2,2)	V(3,2)			
M_3						
$M_{...}$						
M_m						V(m,n)

⇦ **Synonym word forms (synset)**

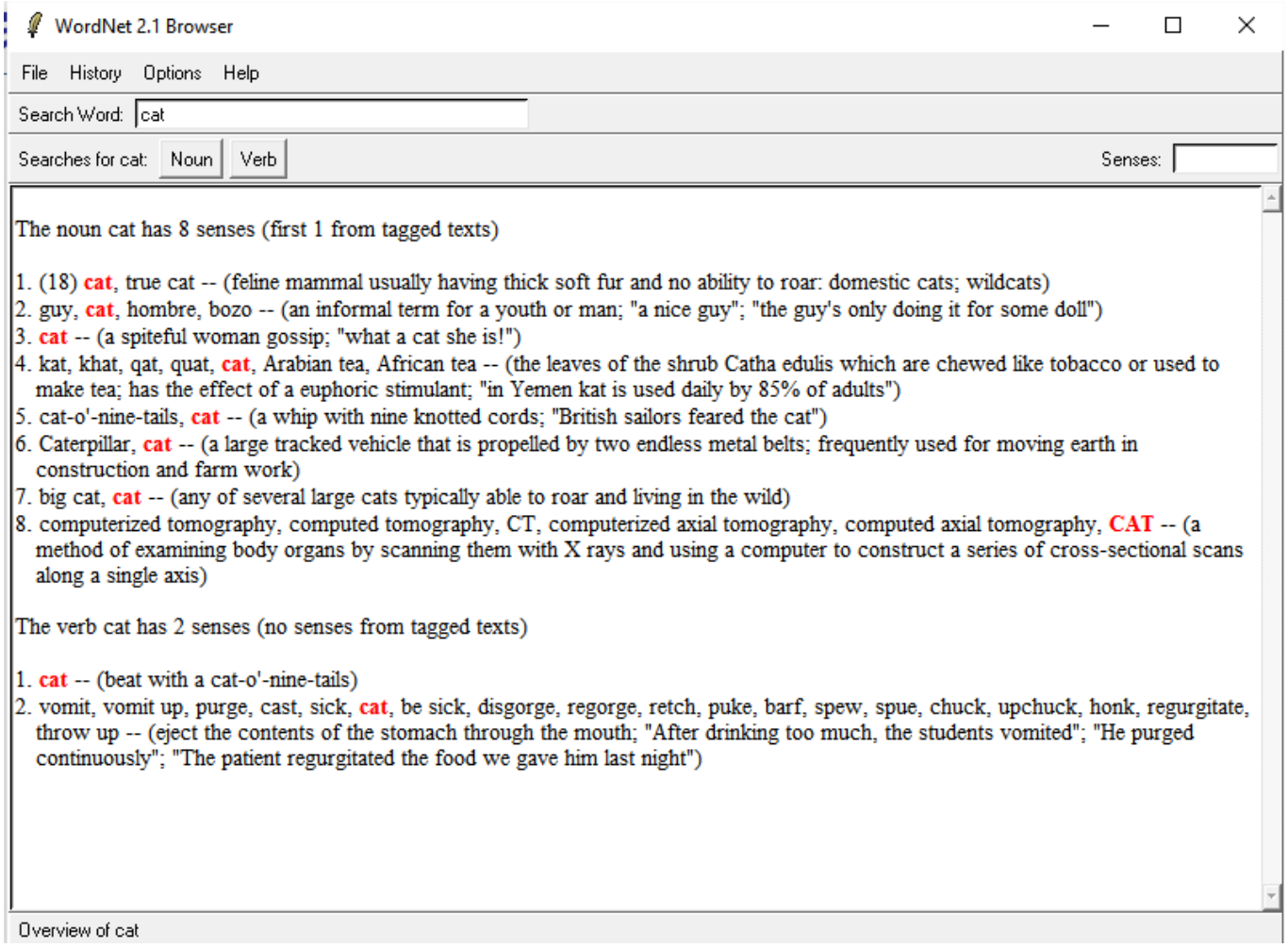
⇧ **polysemous word: disambiguation needed**

[*] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

Sense Repository

WordNet semantic network

https://wordnet.princeton.edu

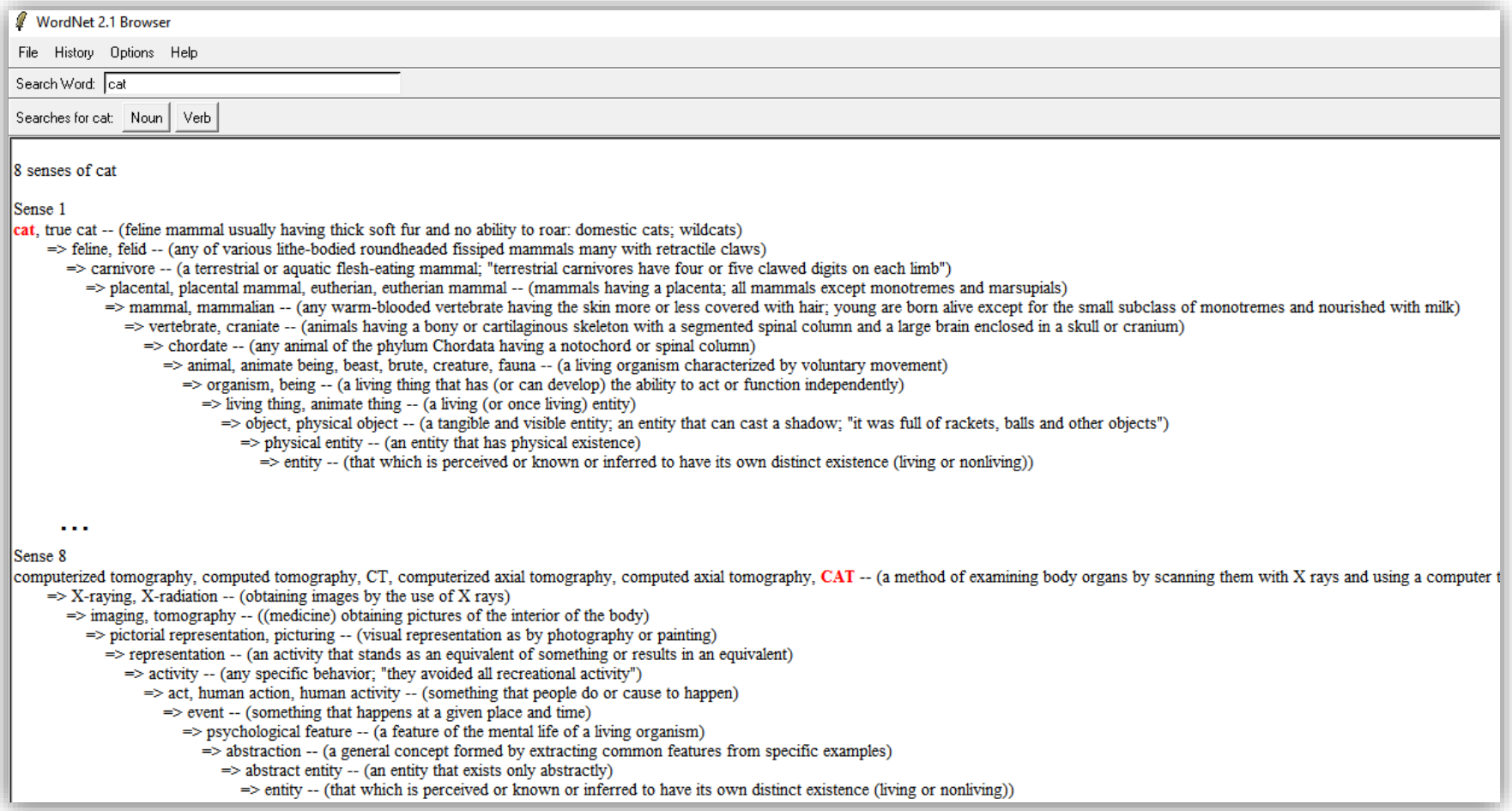


an example of **synset**

Sense Repository

WordNet semantic network

https://wordnet.princeton.edu



WordNet Hierarchies

Word Sense Disambiguation

State of the art: JIGSAW algorithm [*]

Input

- $D = \{w_1, w_2, \ldots, w_h\}$ document

Output

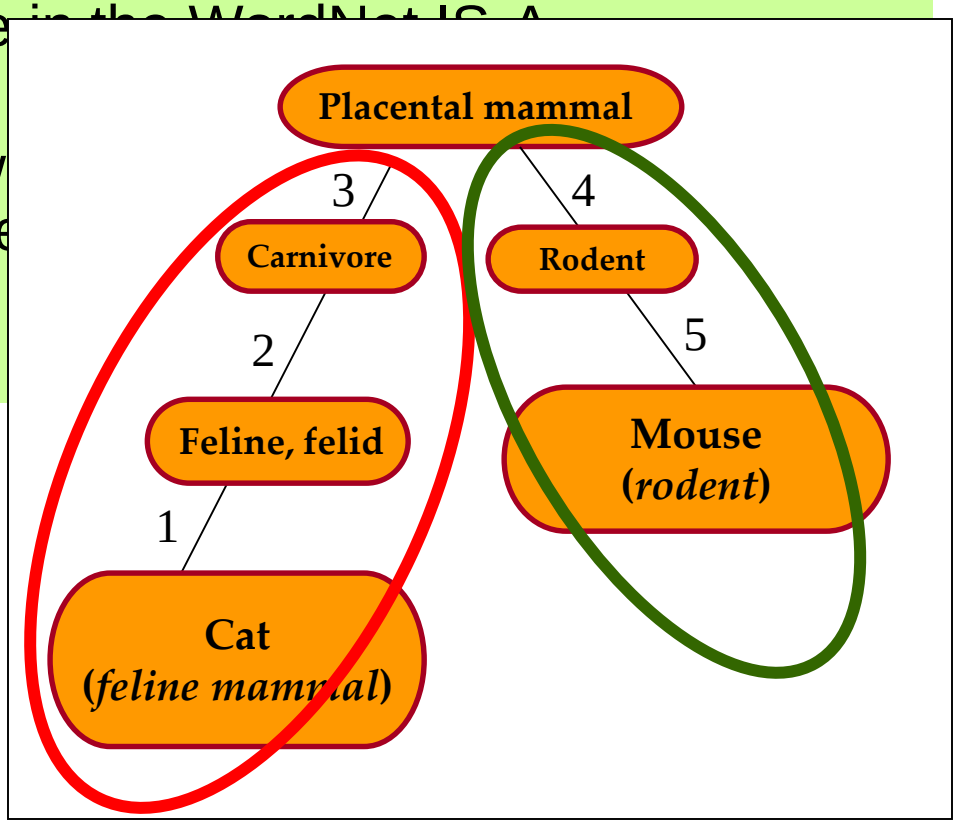
- $X = \{s_1, s_2, \ldots, s_k\}$ $(k \leq h)$
  - Each s_i obtained by disambiguating w_i based on the context of each word
  - Some words not recognized by WordNet
  - Groups of words recognized as a single concept

[*] Basile, P., de Gemmis, M., Gentile, A. L., Lops, P., & Semeraro, G. (2007, June). UNIBA: JIGSAW algorithm for word sense disambiguation. InProceedings of the 4th International Workshop on Semantic Evaluations (pp. 398-401). Association for Computational Linguistics.

JIGSAW WSD algorithm

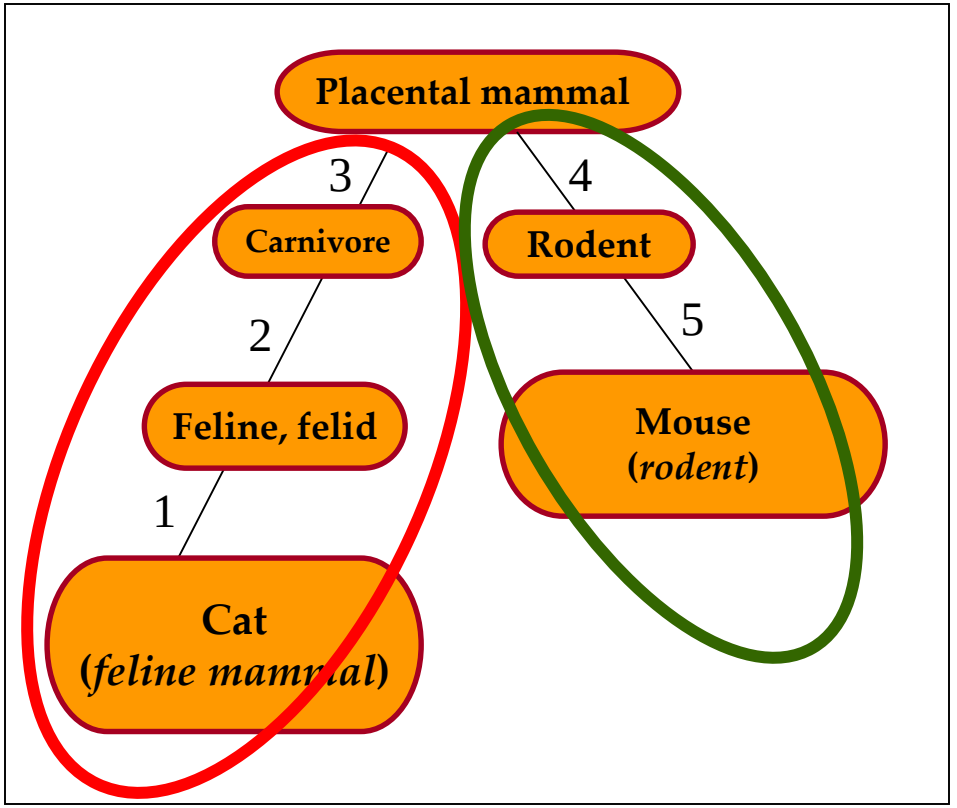
How to use WordNet for WSD?

- **Semantic similarity** between synsets inversely proportional to their distance in the WordNet IS-A hierarchy
- **Path length similarity** betw scores to synsets of a polyse choose the correct sense



Synset semantic similarity

```
24: function SINSIM(a, b)                    ▷ The similarity of the synsets a and b
25:     N_p ← the number of nodes in path p from a to b
26:     D ← maximum depth of the taxonomy    ▷ In WordNet 1.7.1 D = 16
27:     r ← −log(N_p/2D)
28:     return r
29: end function
```

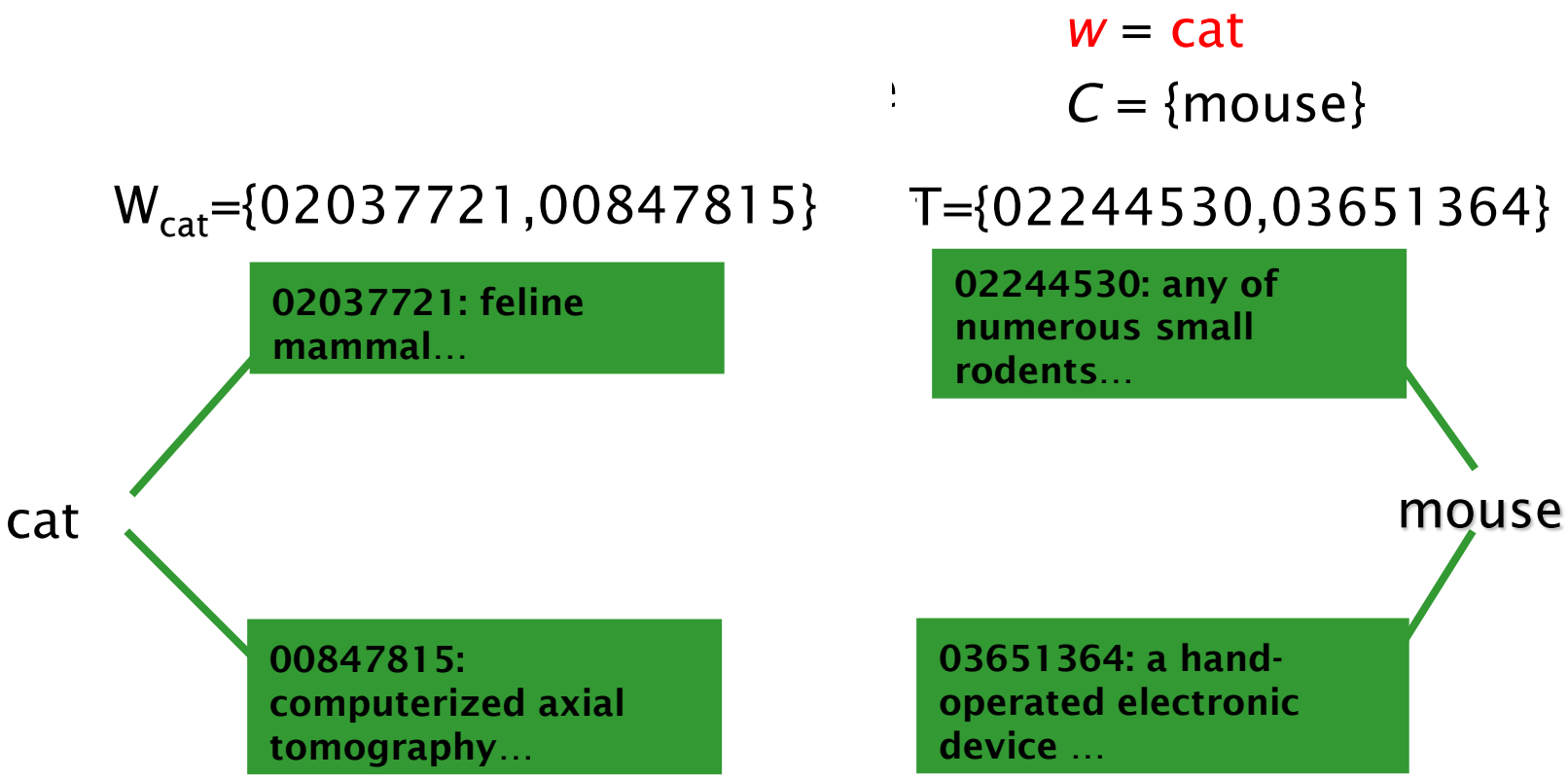


SINSIM(cat,mouse) =
-log(5/32)=0.806

Leacock-Chodorow similarity

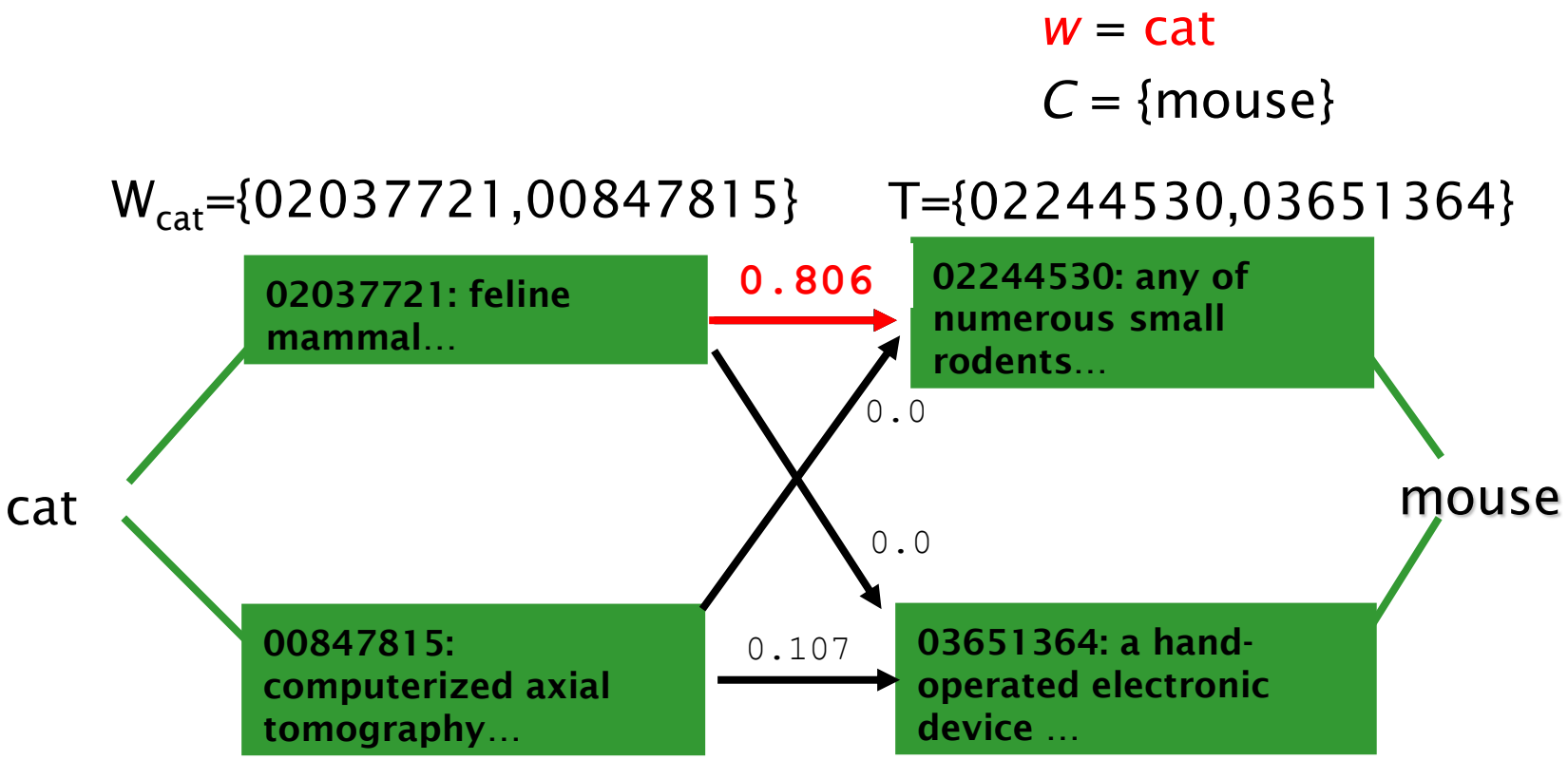
JIGSAW WSD algorithm

"The white cat is hunting the mouse"



JIGSAW WSD algorithm

"The white cat is hunting the mouse"



through WSD can we obtain a

semantics-aware representation

of textual content

Synset-based representation

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well received by critics,[10][11] and won four Academy Awards as well as other accolades including BAFTA

Theatrical release poster

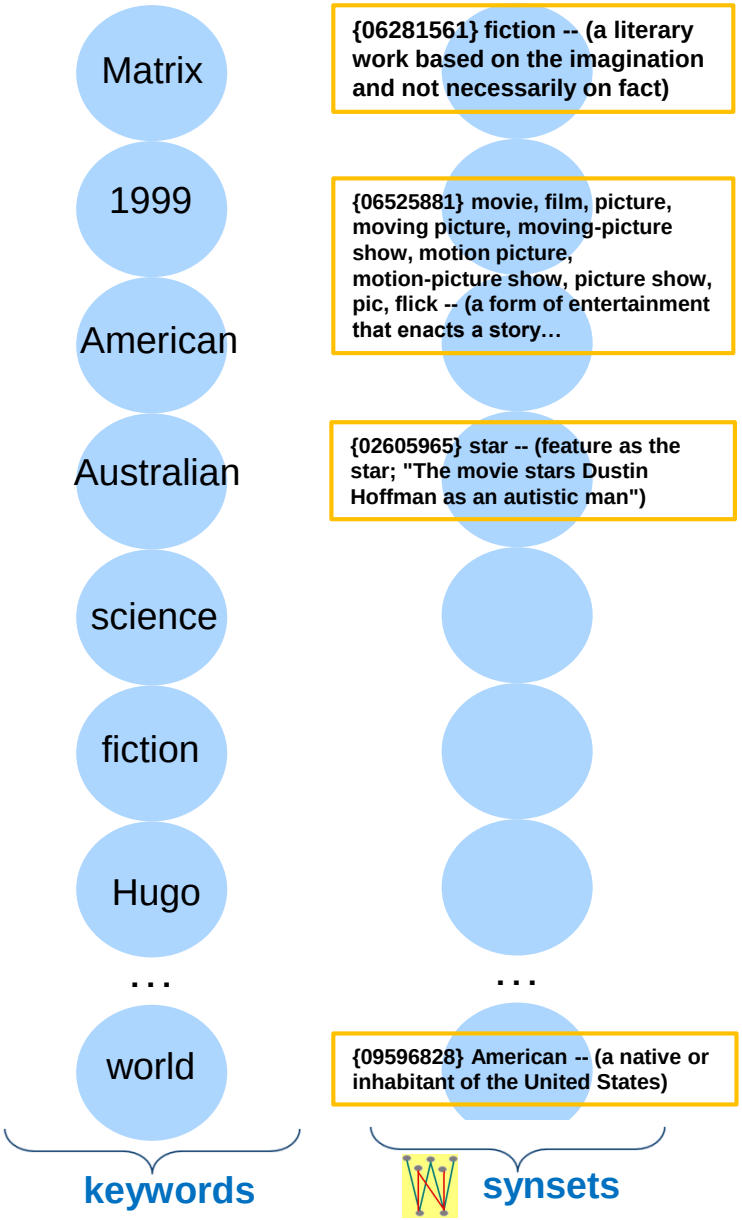
{09596828} American -- (a native or inhabitant of the United States)

{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

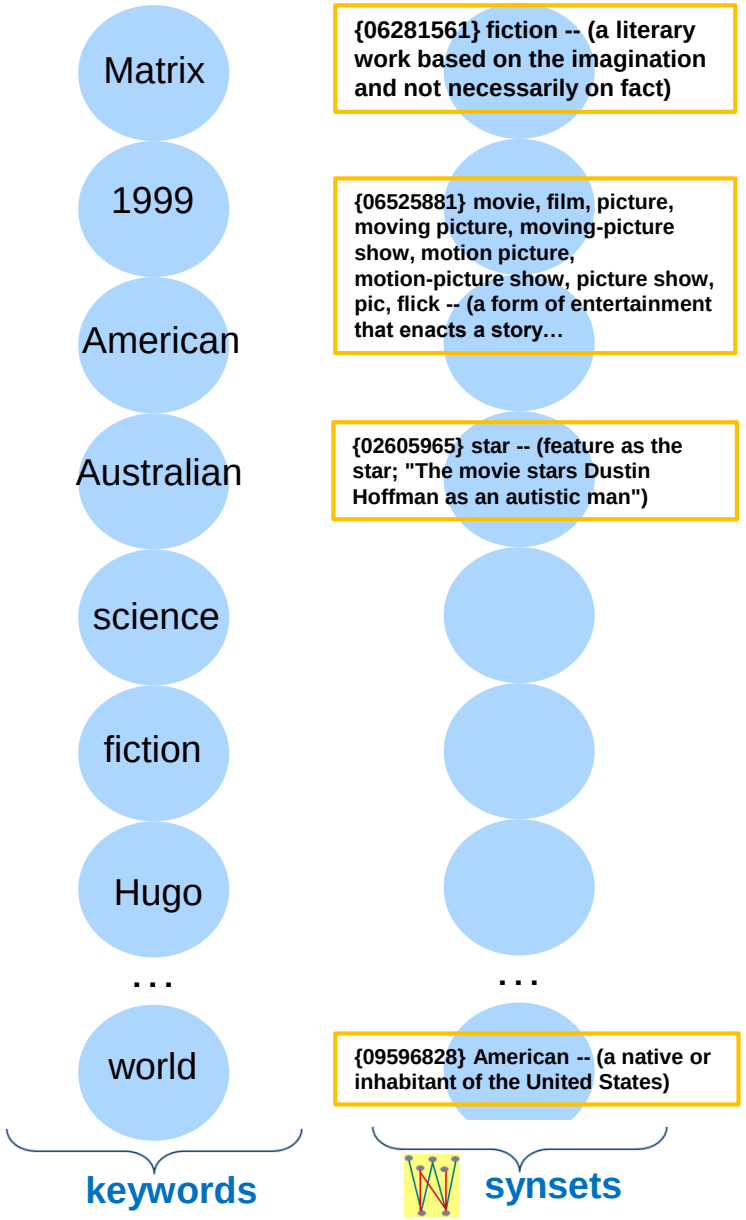
The Matrix representation



through WSD we process the textual description of the item and we obtain a **semantics-aware representation** of the item as output

keyword-based features **replaced with the concepts** (in this case WordNet synsets) they refer to

The Matrix representation



potentially
language-independent representation

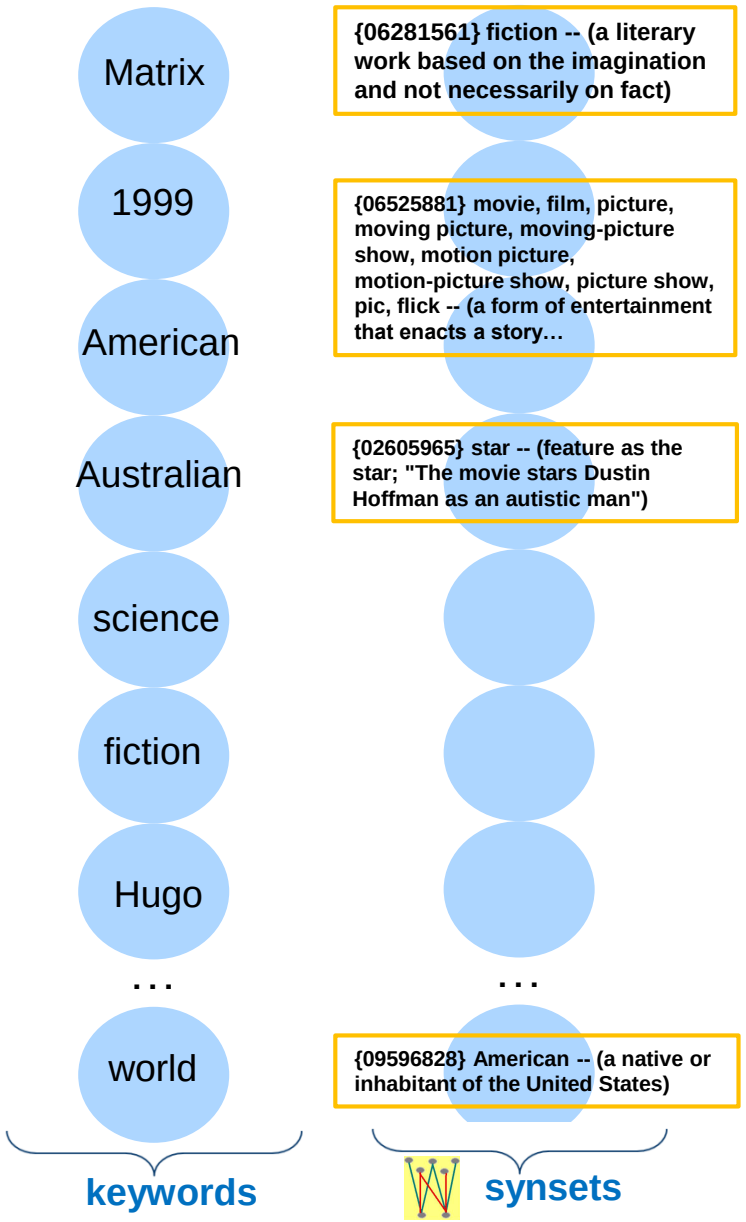
it depends on the WSD algorithm and on the **languages encoded in the lexical resource**

MULTI WordNet

http://multiwordnet.fbk.eu/english/

Multilingual Lexical Resource

The Matrix representation



Word Sense Disambiguation

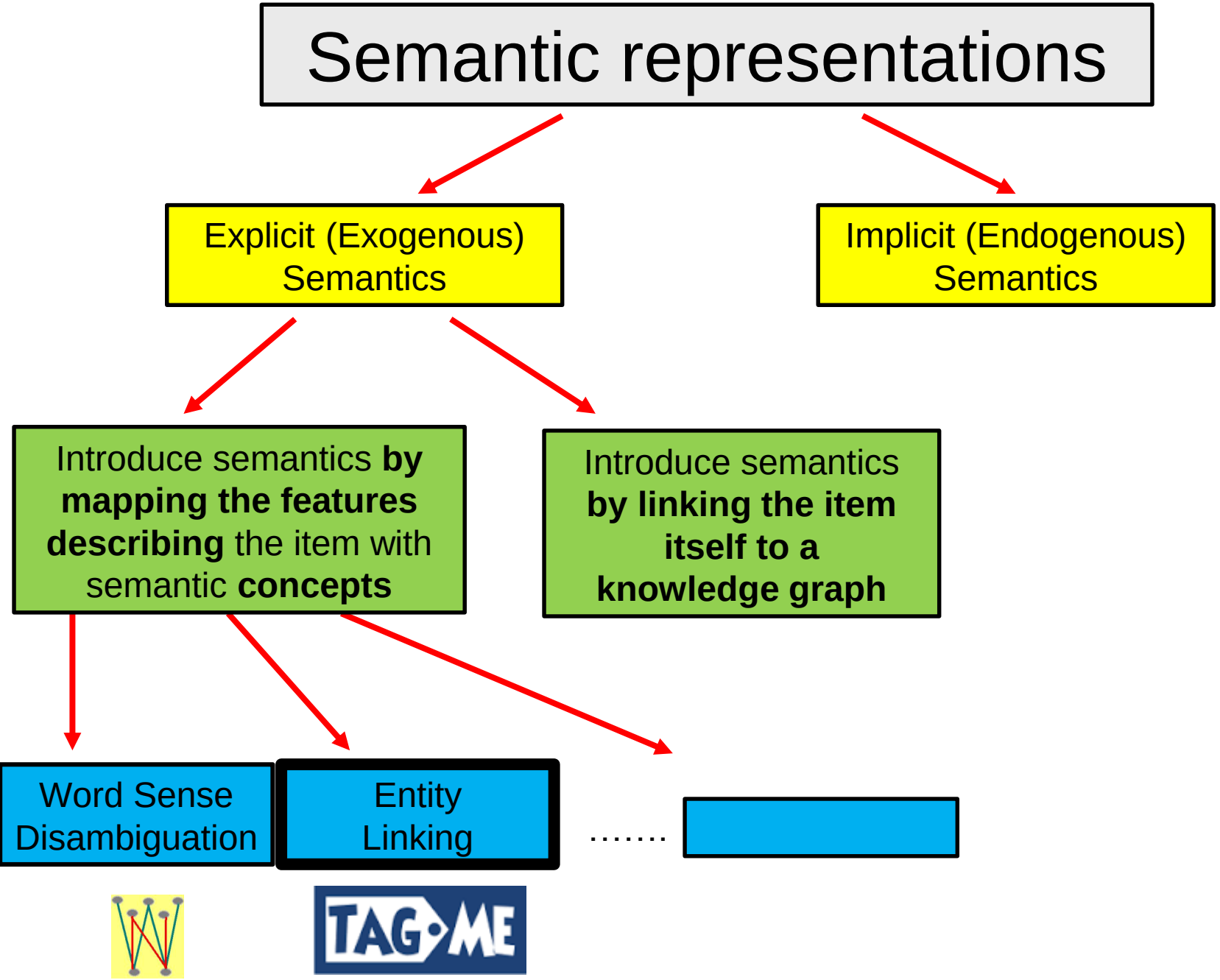
recap

polysemy and **synonymy** effectively handled

classical NLP techniques helpful to remove further noise (e.g. stopwords)

potentially language-independent

entities (e.g. Hugo Weaving) still not recognized


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking the item itself to a knowledge graph**
Word Sense Disambiguation
Entity Linking
.....
TAG·ME

Entity Linking Algorithms

- **Basic Idea**
  - **Input:** free text
    - *e.g. Wikipedia abstract*
  - **Output:** identification of the **entities** mentioned in the text and linking to a knowledge base

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American-Australian science fiction action film written and directed by The Wachowski Brothers, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[4] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[5] Jean Baudrillard's *Simulacra and Simulation*[6] and

The Matrix

Theatrical release poster

The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

Why Entity Linking?

because **we need to identify the entities**
mentioned in the textual description
to better catch user preferences and information needs

Several state-of-the-art implementations are already available

OPEN CALAIS

THOMSON REUTERS

TAG·ME

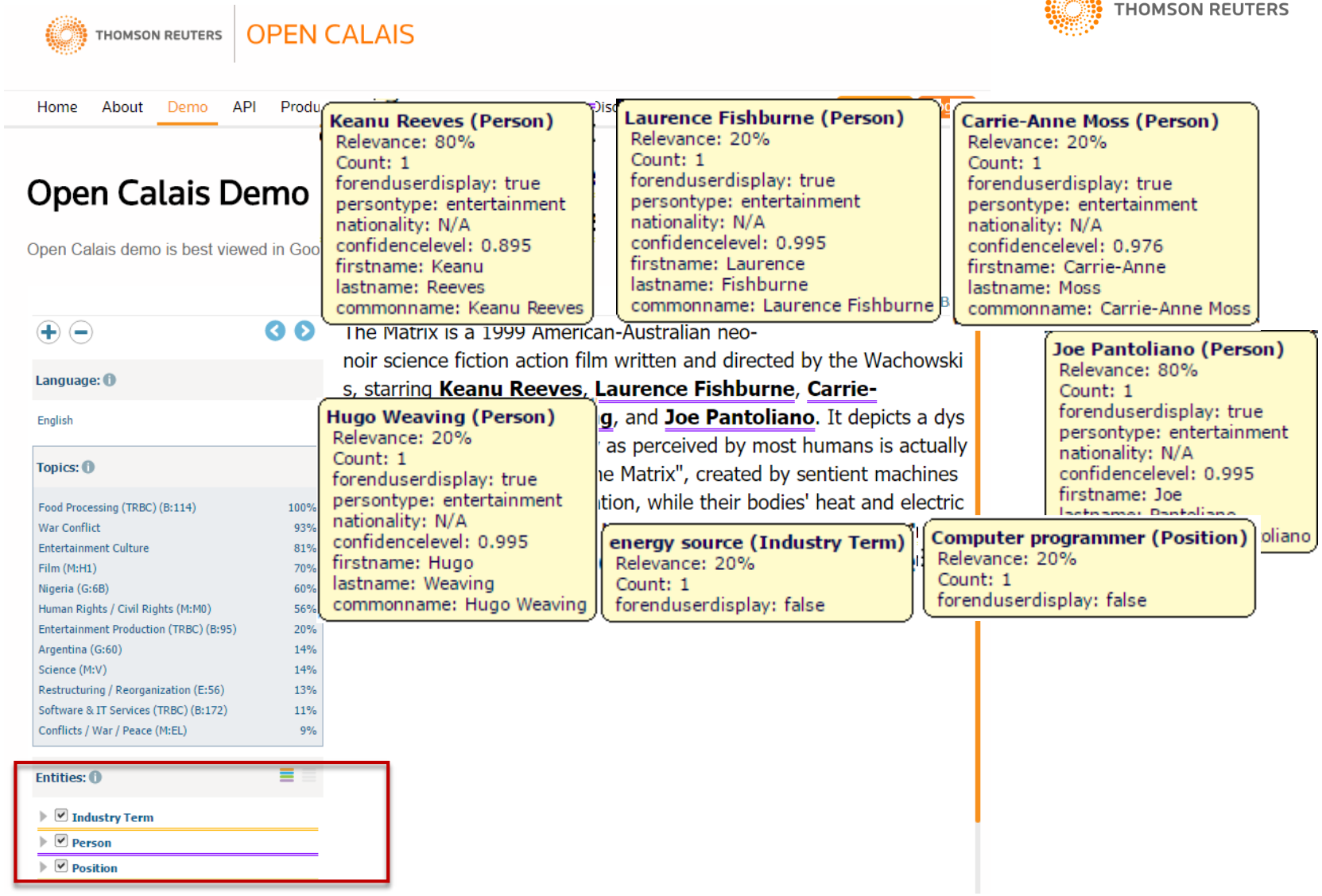
fy Babelfy

… and many more

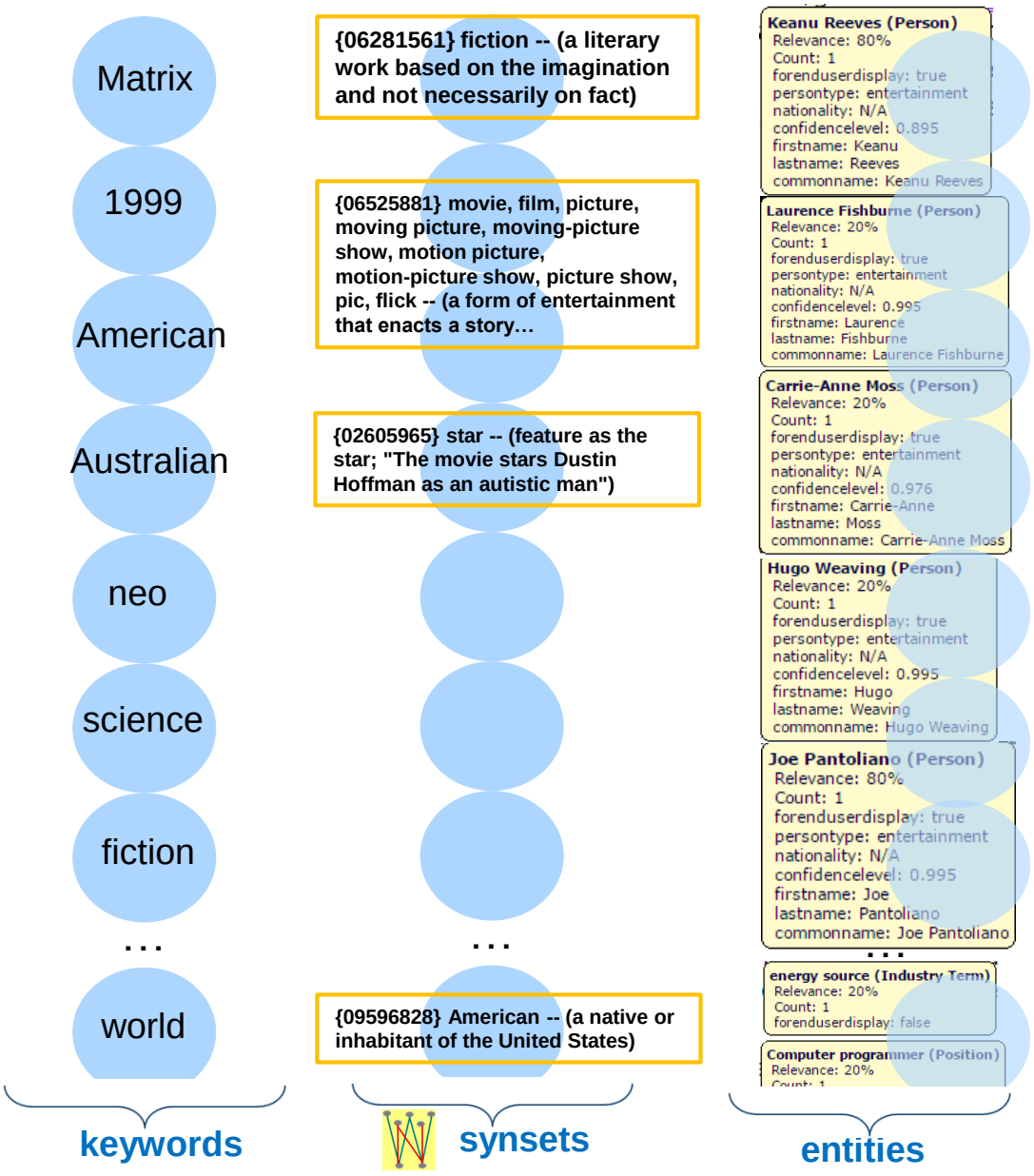
Entity Linking Algorithms

OpenCalais

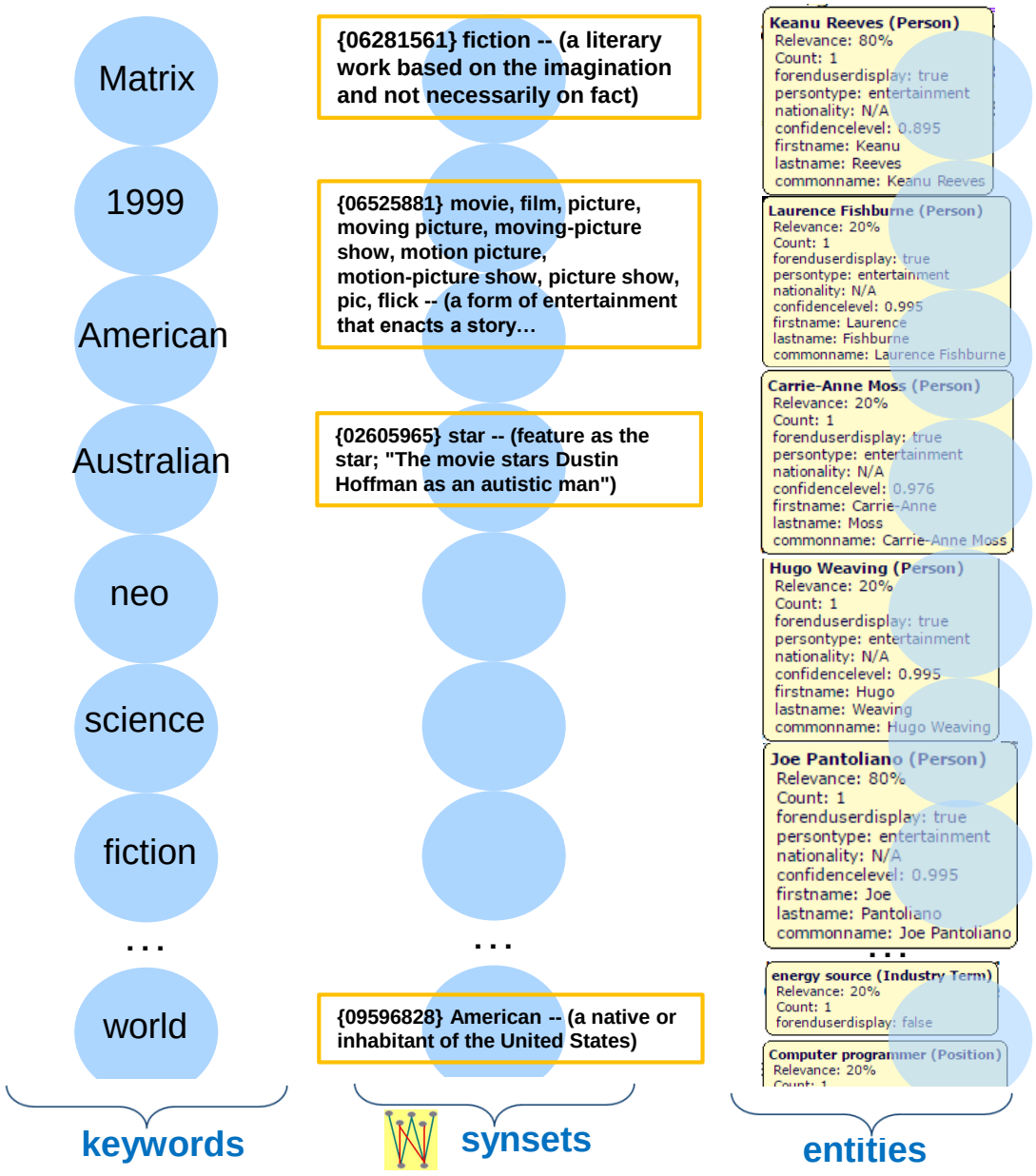
http://www.opencalais.com/opencalais-api/



The Matrix representation



The Matrix representation



OPEN CALAIS

THOMSON REUTERS

entities are correctly recognized and modeled

partially multilingual

(entities are inherently multilingual, but other concepts aren't)

common sense and abstract concepts now ignored

Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/

TAG·ME

Output

The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

very **transparent** and **human-readable** content representation

non-trivial NLP tasks automatically performed

(stopwords removal, n-grams identification, named entities recognition and disambiguation)

Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/

Output

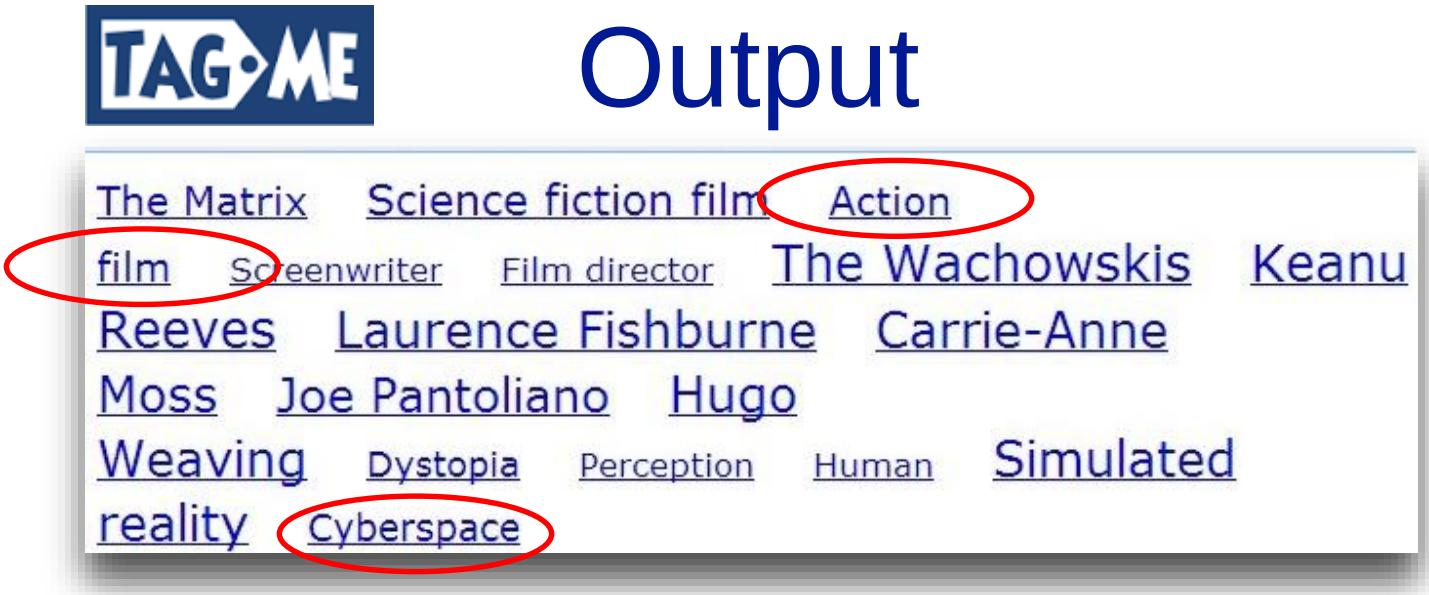
The Matrix Science fiction film Action film Screenwriter Film director The Wachowskis Keanu Reeves Laurence Fishburne Carrie-Anne Moss Joe Pantoliano Hugo Weaving Dystopia Perception Human Simulated reality Cyberspace

each entity identified in the content can be a feature of a **semantics-aware content representation based on entity linking**

Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/



Advantage #1: several common sense concepts are now identified

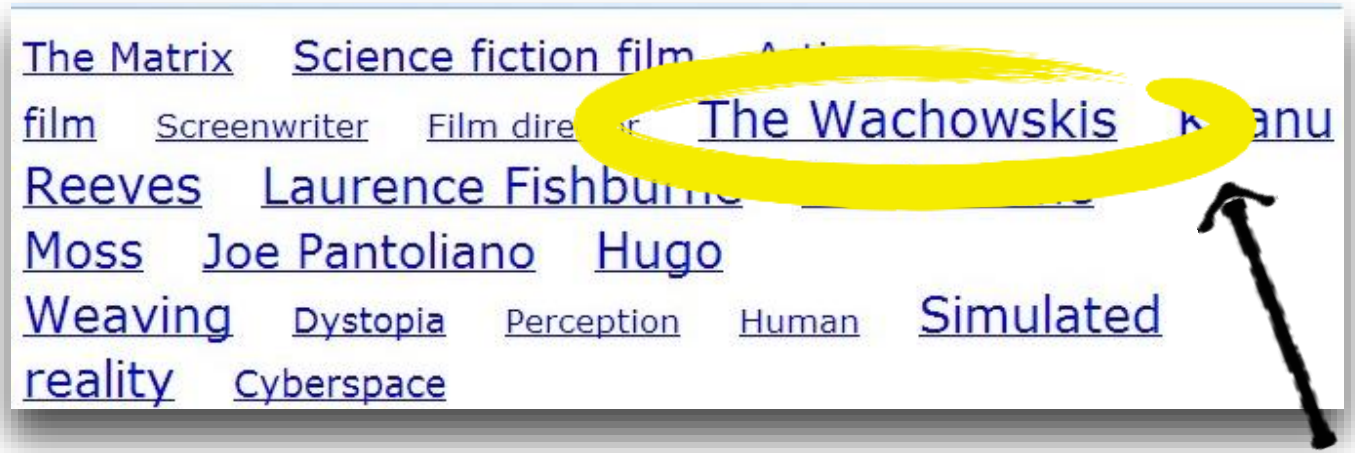
Entity Linking Algorithms

Tag.me

https://tagme.d4science.org/tagme/

TAG·ME

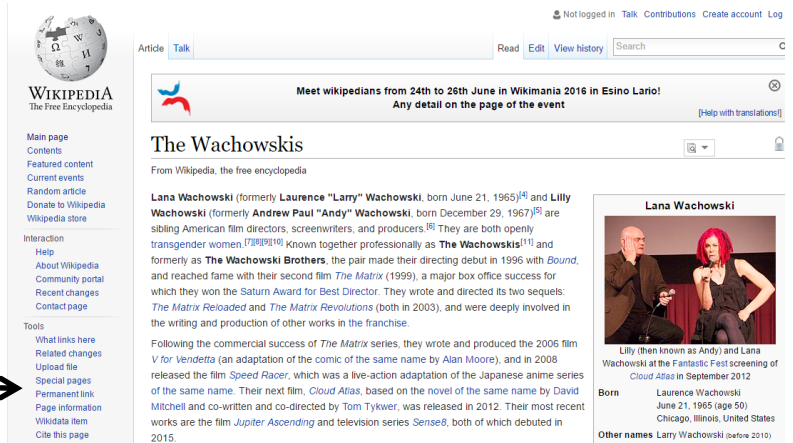
Output



not a simple textual feature!

Advantage #2: each **entity** is a reference to a **Wikipedia page**

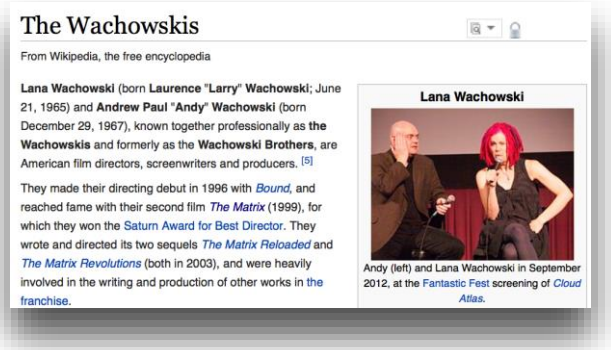
http://en.wikipedia.org/wiki/The_Wachowskis



Entity Linking Algorithms

Tag.me + Wikipedia Categories

https://tagme.d4science.org/tagme/


The Wachowskis

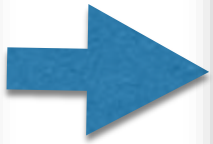
From Wikipedia, the free encyclopedia

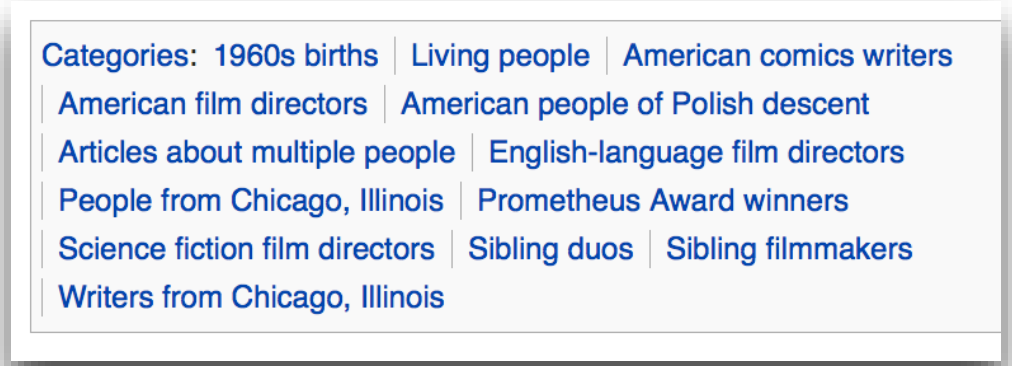
Lana Wachowski (born **Laurence "Larry" Wachowski**; June 21, 1965) and **Andrew Paul "Andy" Wachowski** (born December 29, 1967), known together professionally as **the Wachowskis** and formerly as the **Wachowski Brothers**, are American film directors, screenwriters and producers. [5]

They made their directing debut in 1996 with *Bound*, and reached fame with their second film *The Matrix* (1999), for which they won the Saturn Award for Best Director. They wrote and directed its two sequels *The Matrix Reloaded* and *The Matrix Revolutions* (both in 2003), and were heavily involved in the writing and production of other works in the franchise.

Lana Wachowski

Andy (left) and Lana Wachowski in September 2012, at the Fantastic Fest screening of *Cloud Atlas*.




Categories: 1960s births | Living people | American comics writers | American film directors | American people of Polish descent | Articles about multiple people | English-language film directors | People from Chicago, Illinois | Prometheus Award winners | Science fiction film directors | Sibling duos | Sibling filmmakers | Writers from Chicago, Illinois

We **can enrich** this **entity-based representation**

by exploiting the **Wikipedia categories' tree**

Entity Linking Algorithms

Tag.me + Wikipedia Categories

https://tagme.d4science.org/tagme/

We **can enrich** this **entity-based representation**

by exploiting the **Wikipedia categories' tree**

Entity Linking Algorithms

Tag.me + Wikipedia Categories

https://tagme.d4science.org/tagme/

final representation of items obtained by merging **entities** identified in the text with the **(most relevant) Wikipedia categories** each entity is linked to

The Matrix representation

Matrix

1999

American

Australian

neo

science

fiction

…

world

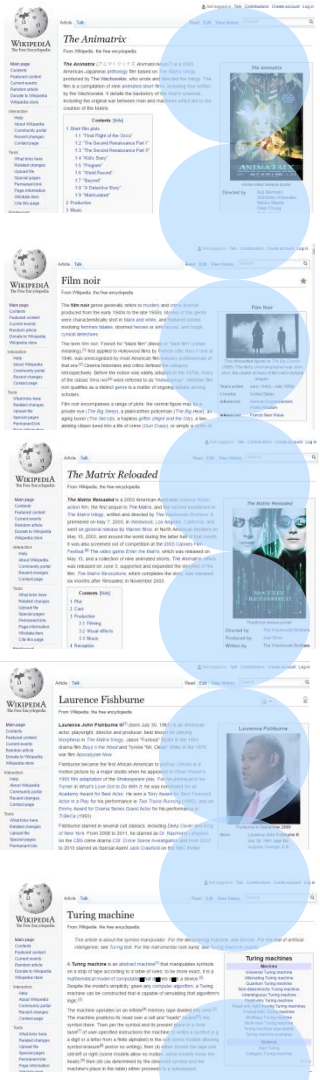
{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

…

{09596828} American -- (a native or inhabitant of the United States)



keywords

synsets

Wikipedia pages

The Matrix representation

Matrix

1999

American

Australian

neo

science

fiction

…

world

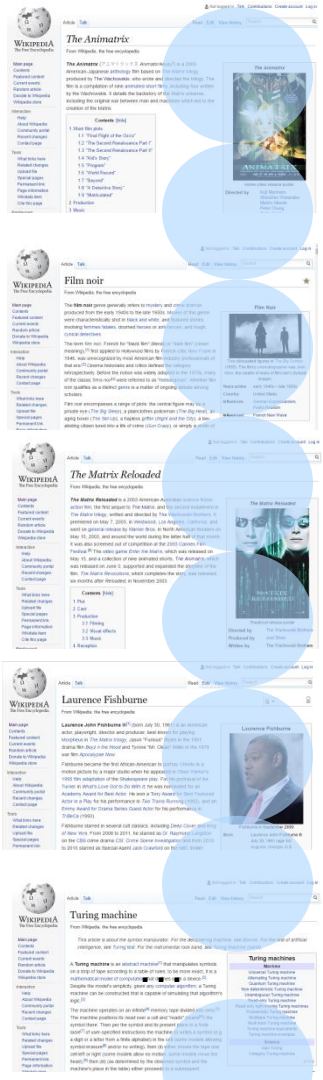
{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

…

{09596828} American -- (a native or inhabitant of the United States)



keywords

synsets

Wikipedia pages

through TAG.ME we are able to **represent an item in terms of a set of Wikipedia pages**

we can further **enrich the representation with new features, as the Wikipedia categories**

The Matrix representation

Matrix

1999

American

Australian

neo

science

fiction

…

world

keywords

{06281561} fiction -- (a literary work based on the imagination and not necessarily on fact)

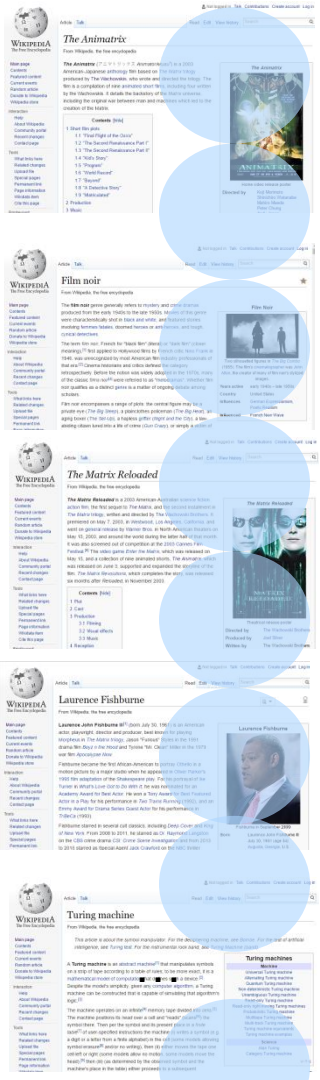
{06525881} movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick -- (a form of entertainment that enacts a story…

{02605965} star -- (feature as the star; "The movie stars Dustin Hoffman as an autistic man")

…

{09596828} American -- (a native or inhabitant of the United States)

synsets



Wikipedia pages

TAG-ME

entities recognized and modeled (as in OpenCalais)

Wikipedia-based representation: some common sense terms included, and new interesting features (e.g. «science-fiction director) can be generated

terms without a Wikipedia mapping are ignored

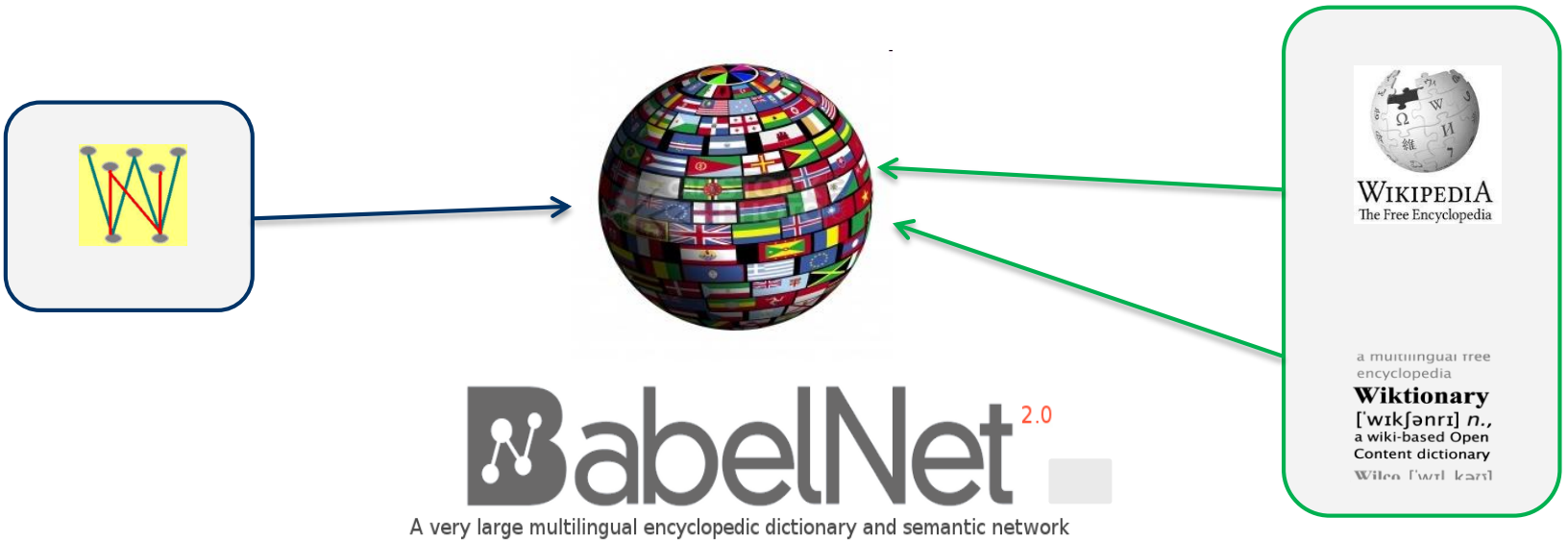
Entity Linking Algorithms

Babelfy

http://babelfy.org/

traditional resources

collaborative resources



- manually curated by experts
- available for a few languages
- difficult to maintain and update

- collaboratively built by the crowd
- highly multilingual
- up-to-date

BabelNet

CERCA, TRADUCI, IMPARA!

Scrivi un termine o un testo... | ITALIAN | TRADUCI IN... | CERCA

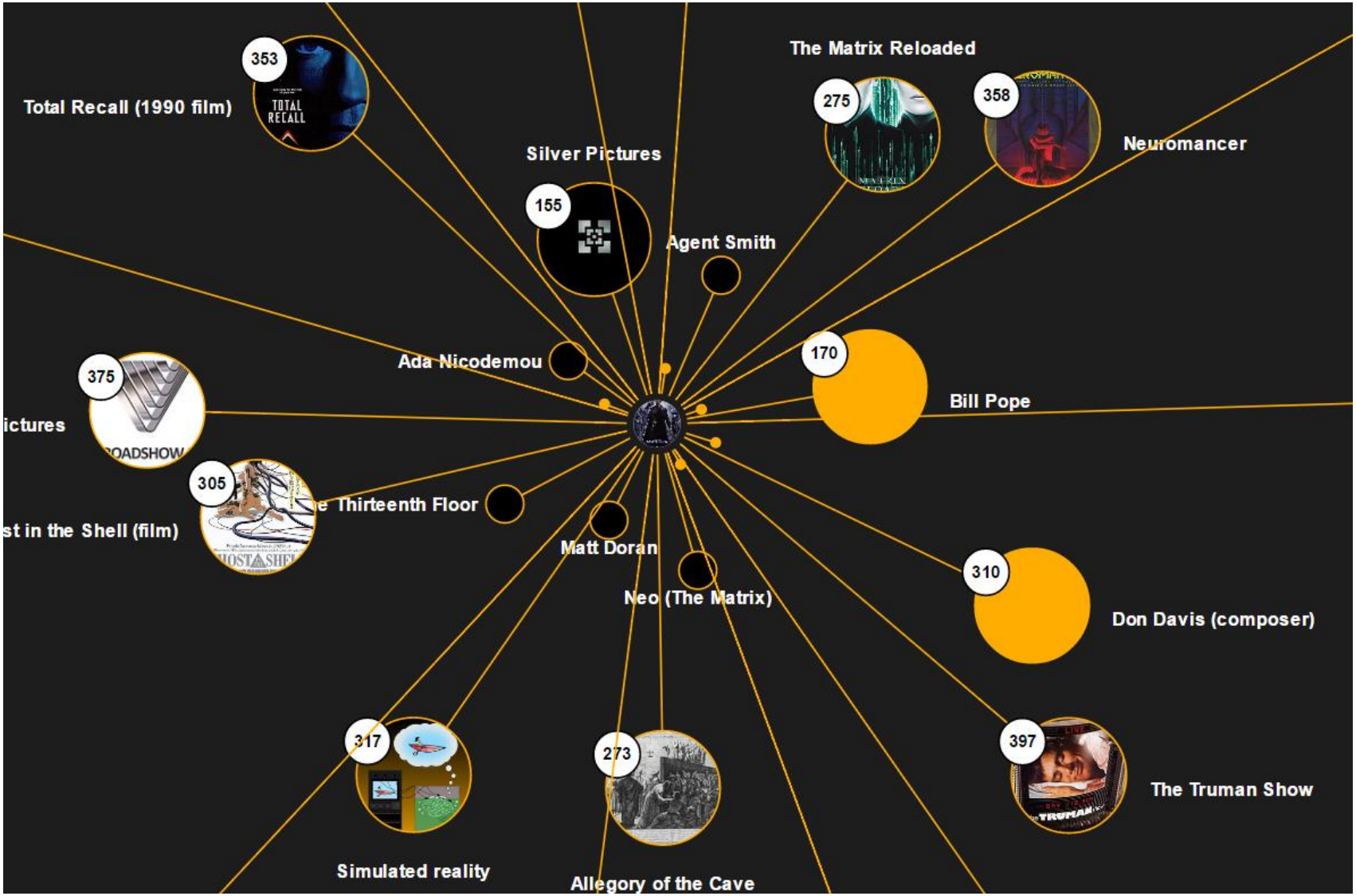
BABELNET IN TIME MAGAZINE!

PREFERENZE

BabelNet 3.6: General statistics

Number of languages:	271
Total number of Babel synsets:	13,801,844
Total number of Babel senses:	745,856,326
Total number of concepts:	6,066,396
Total number of Named Entities:	7,735,448
Total number of lexico-semantic relations:	380,239,084
Total number of glosses (textual definitions):	40,705,588
Total number of images:	10,767,833
Total number of Babel synsets with at least one domain:	1,558,806
Total number of compounds:	743,296
Total number of other forms:	6,393,568
Total number of Babel synsets with at least one picture:	2,948,668
Total number of RDF triples:	1,971,744,856

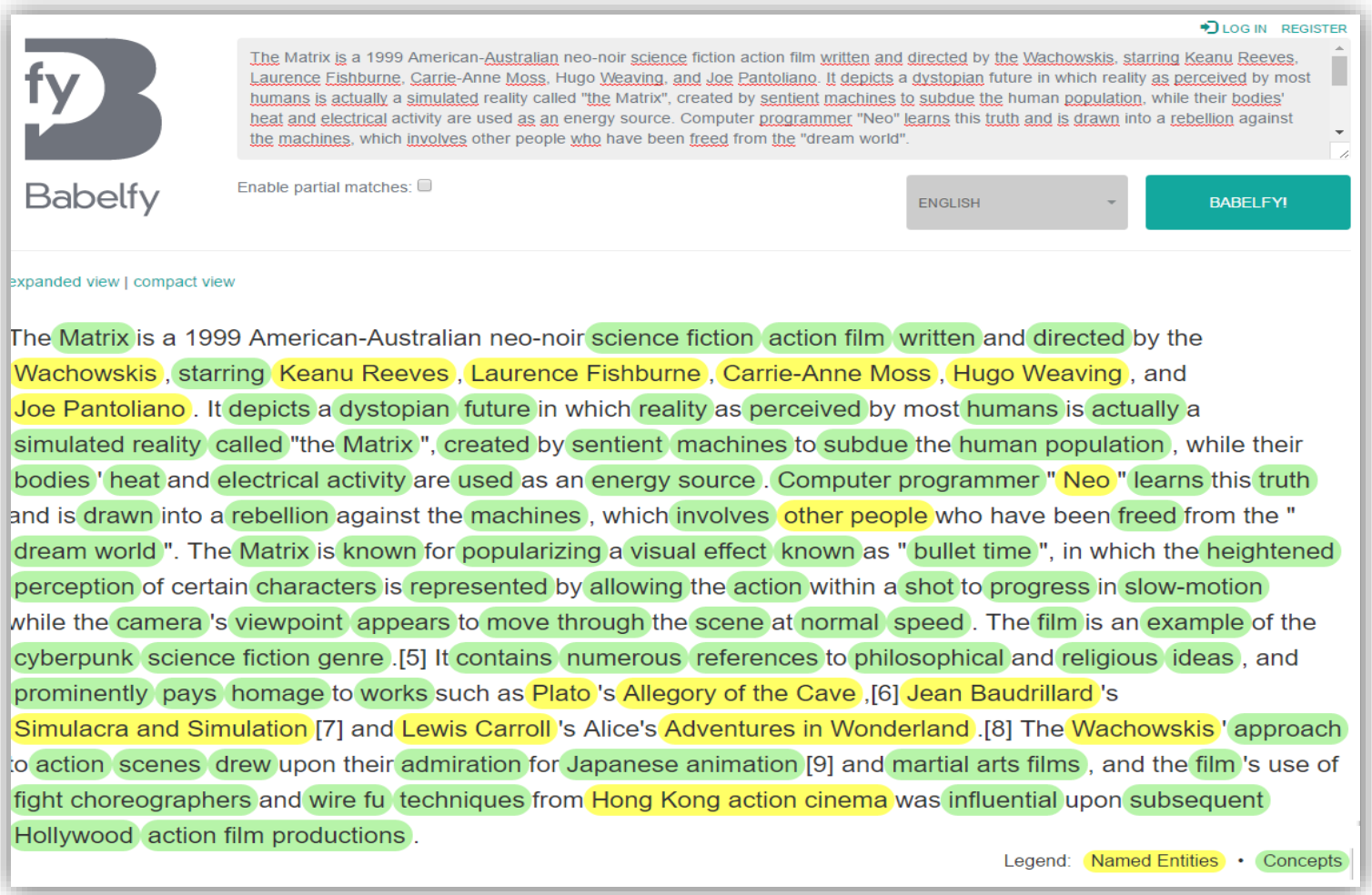
BabelNet: network of concepts



Entity Linking Algorithms

Babelfy

http://babelfy.org/

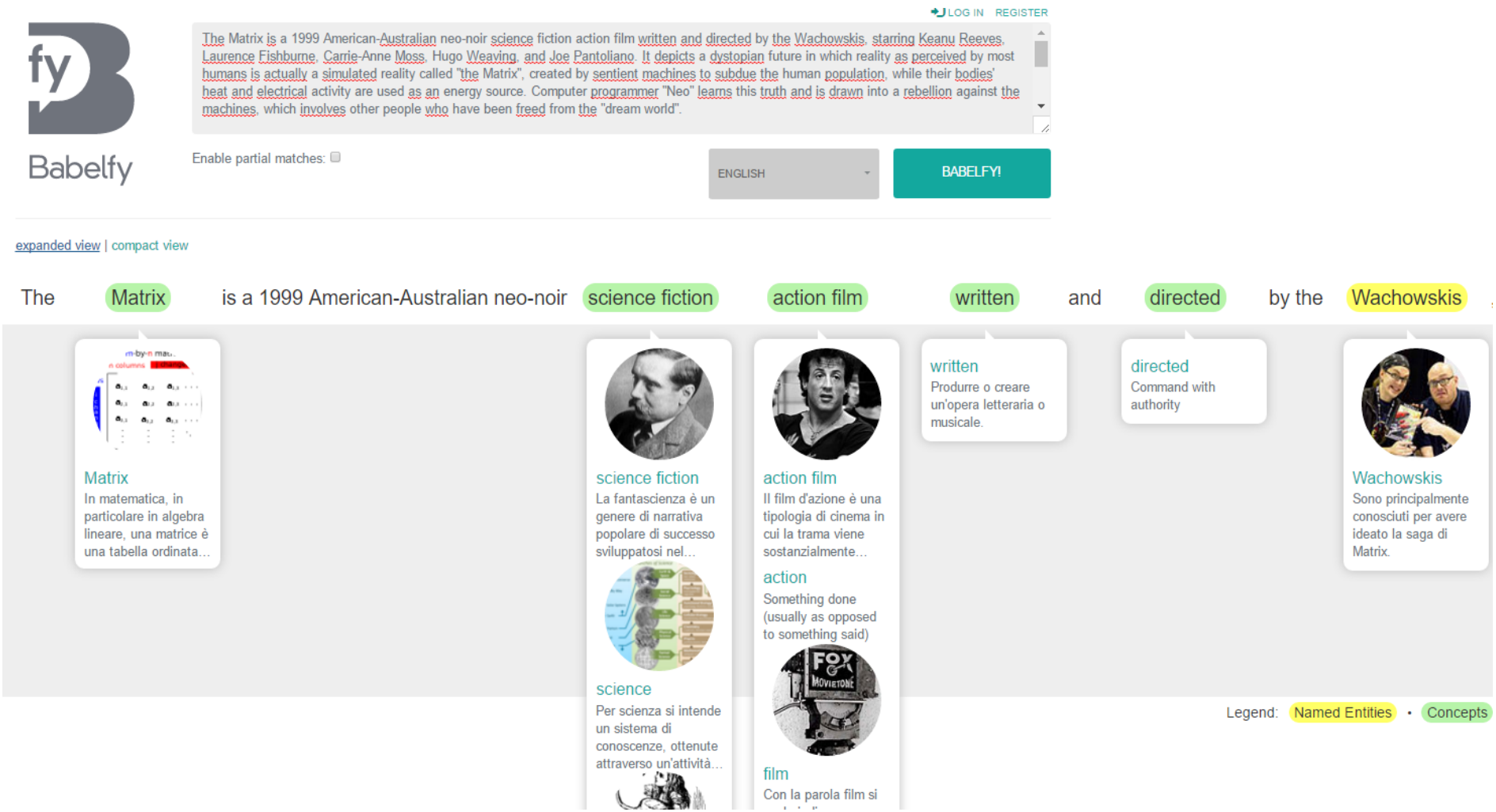


we have both **Named Entities** and **Concepts**!

Entity Linking Algorithms

Babelfy

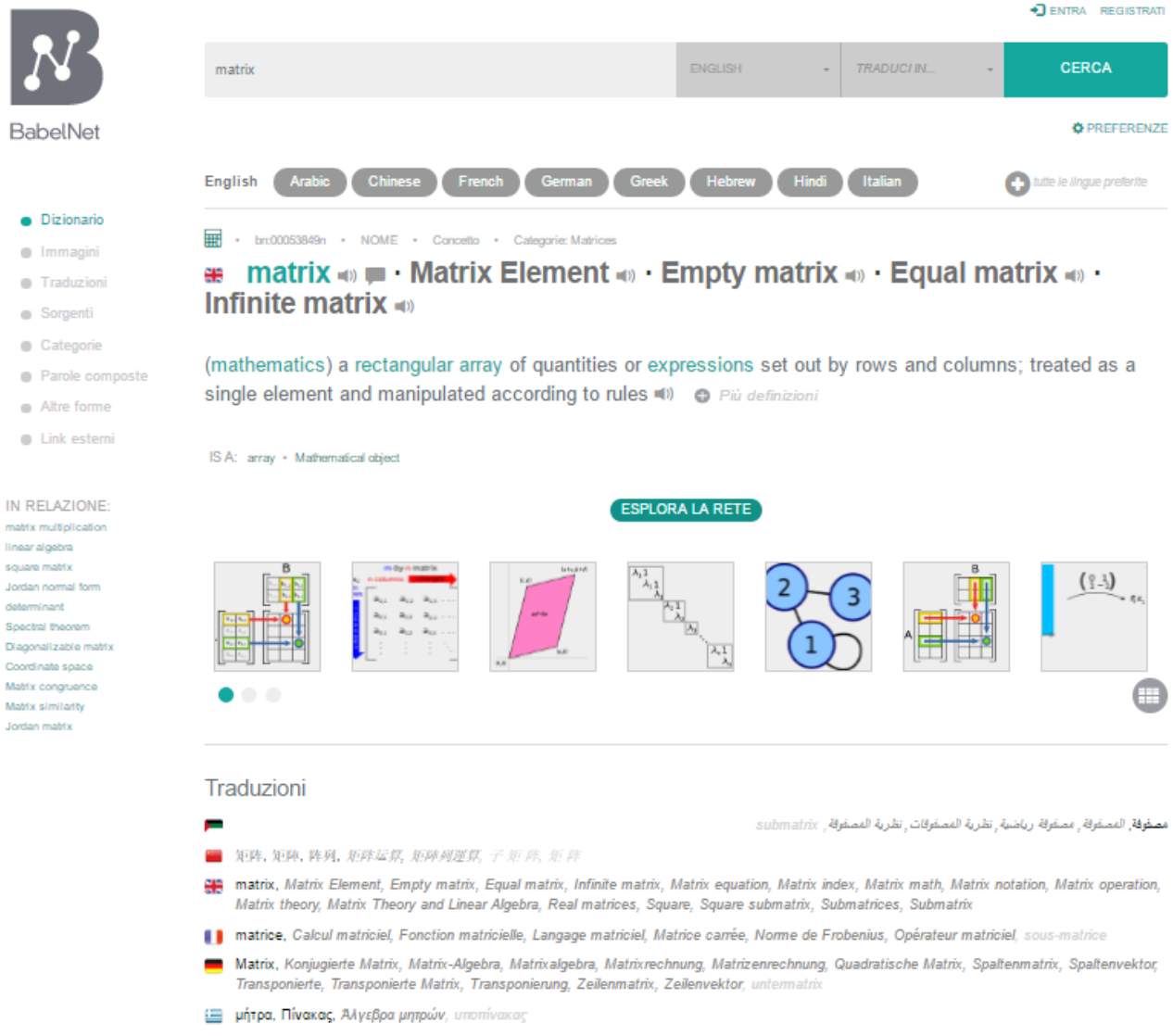
http://babelfy.org/



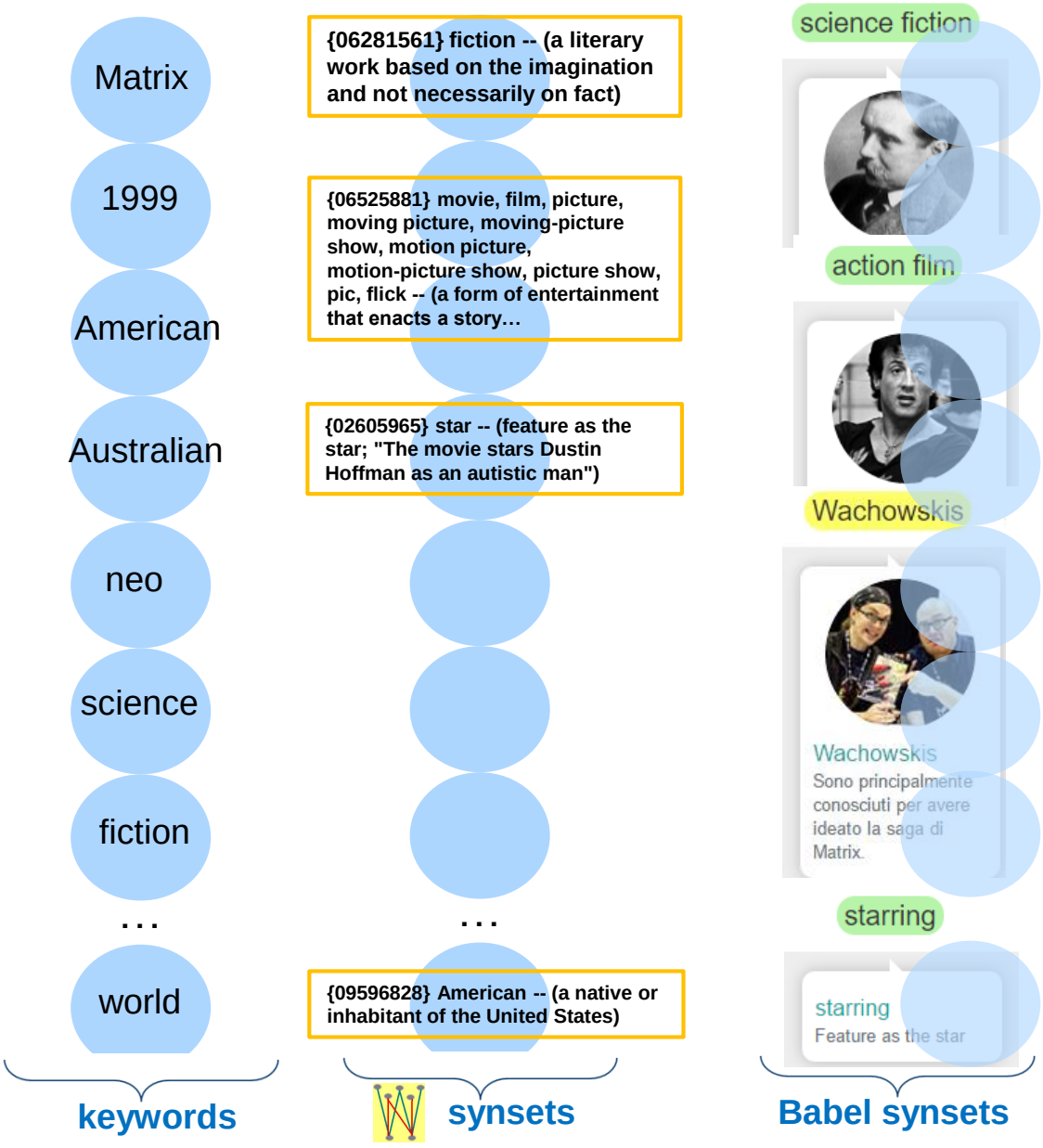
Entity Linking Algorithms

Babelfy

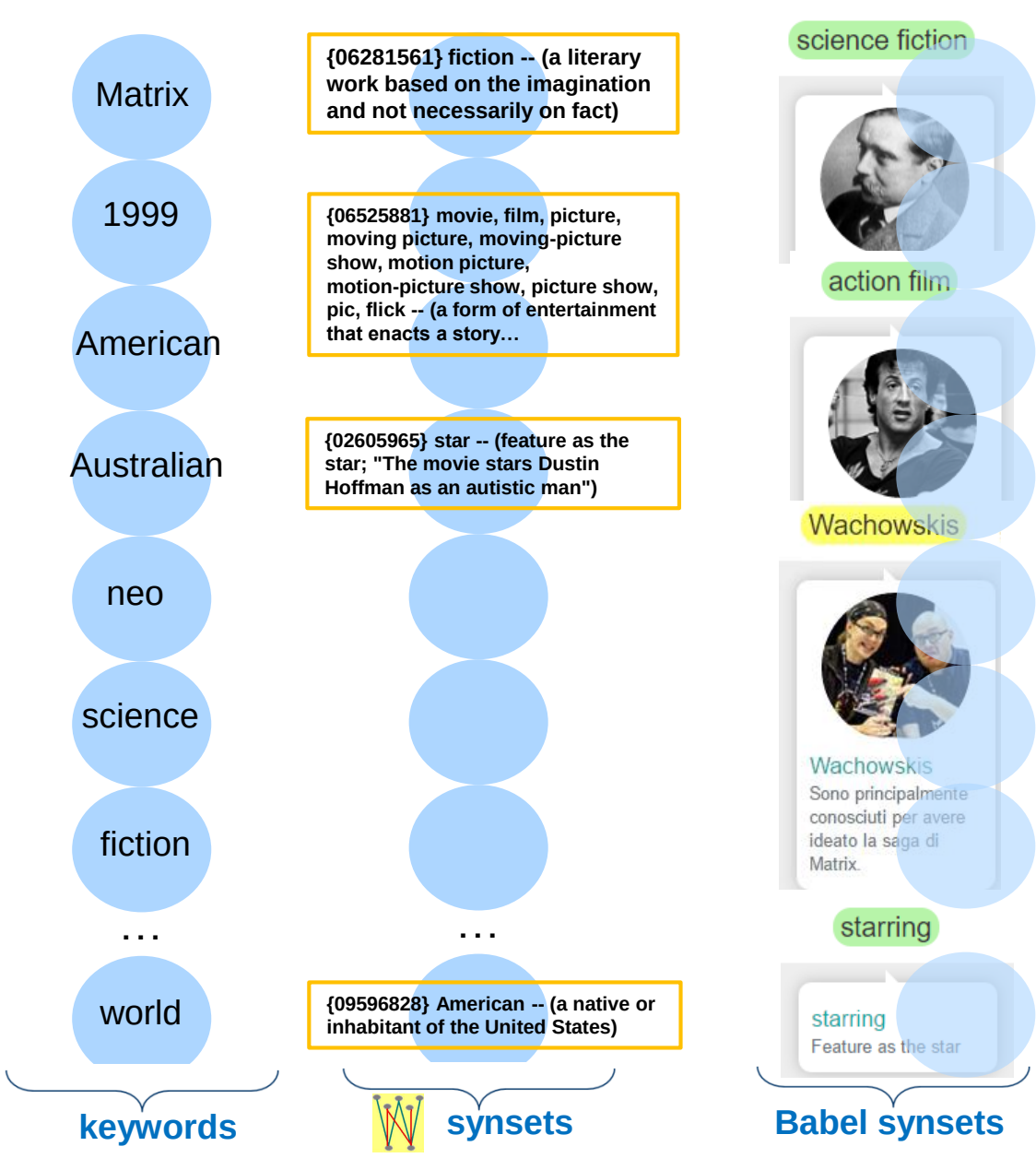
http://babelfy.org/



The Matrix representation



The Matrix representation



BabelNet 2.5

A very large multilingual encyclopedic dictionary and semantic network

entities recognized and modeled (as in OpenCalais and Tag.me)

Wikipedia-based representation: some common sense terms included, and new interesting features (e.g. «science-fiction director) can be generated

includes linguistic knowledge and is able to disambiguate terms

also multilingual!

Recap #4

encoding **exogenous semantics** by processing textual descriptions

- **«Exogenous» techniques use external knowledge sources to inject semantics**
- **Word Sense Disambiguation** algorithms process the textual description and replace keywords with semantic concepts (as synsets)
- **Entity Linking algorithms** focus on the identification of the entities. Some recent approaches also able to identify common sense terms
- **Combination of both approaches is potentially the best strategy**

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

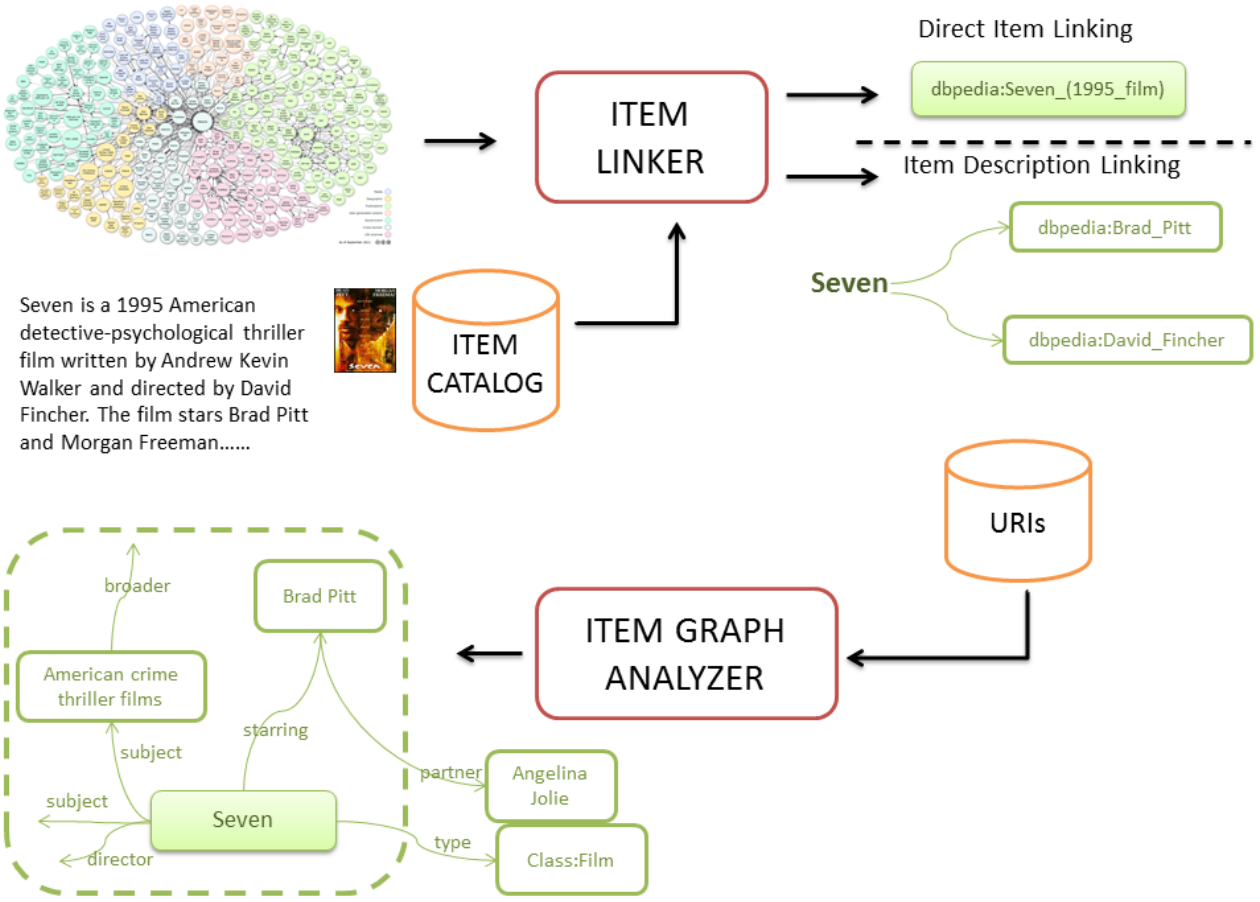
Ontologies

Linked Open Data

DBpedia

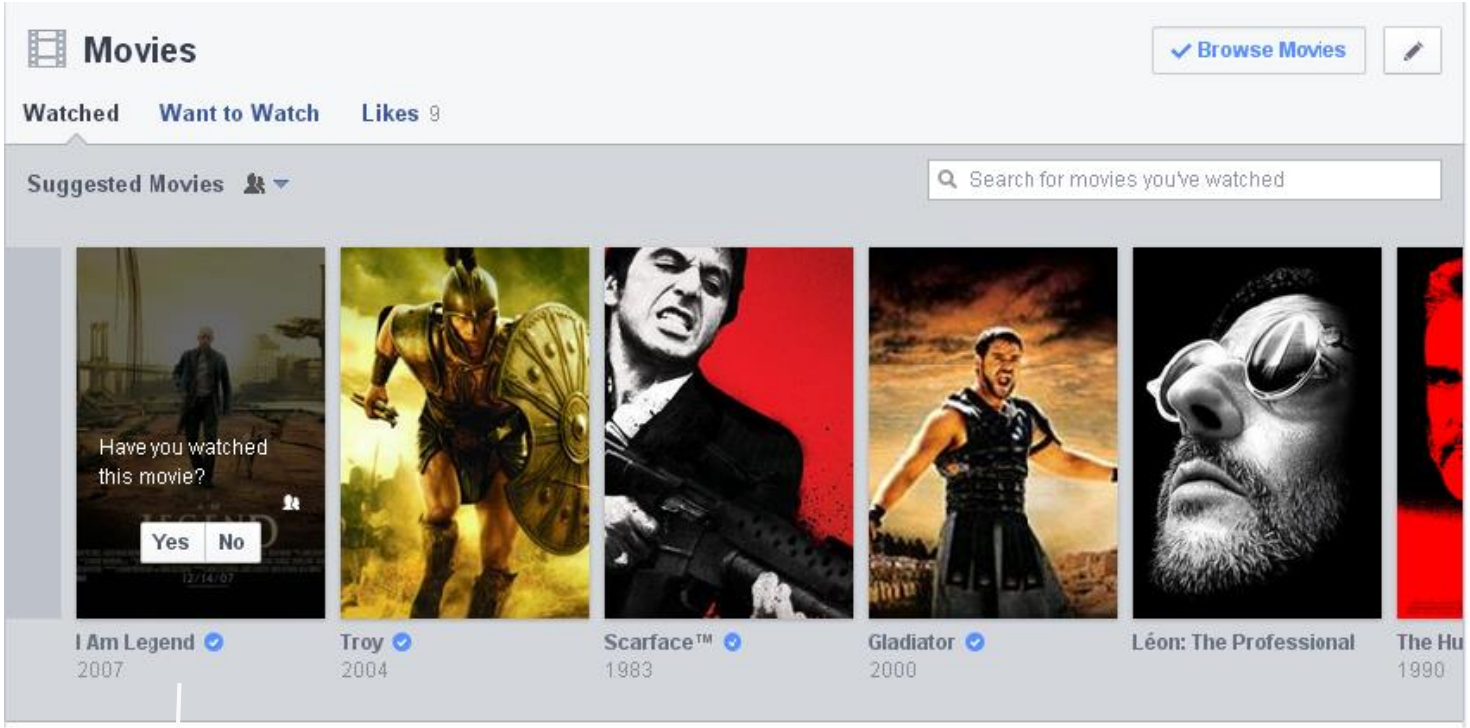
.....

Direct Item Linking



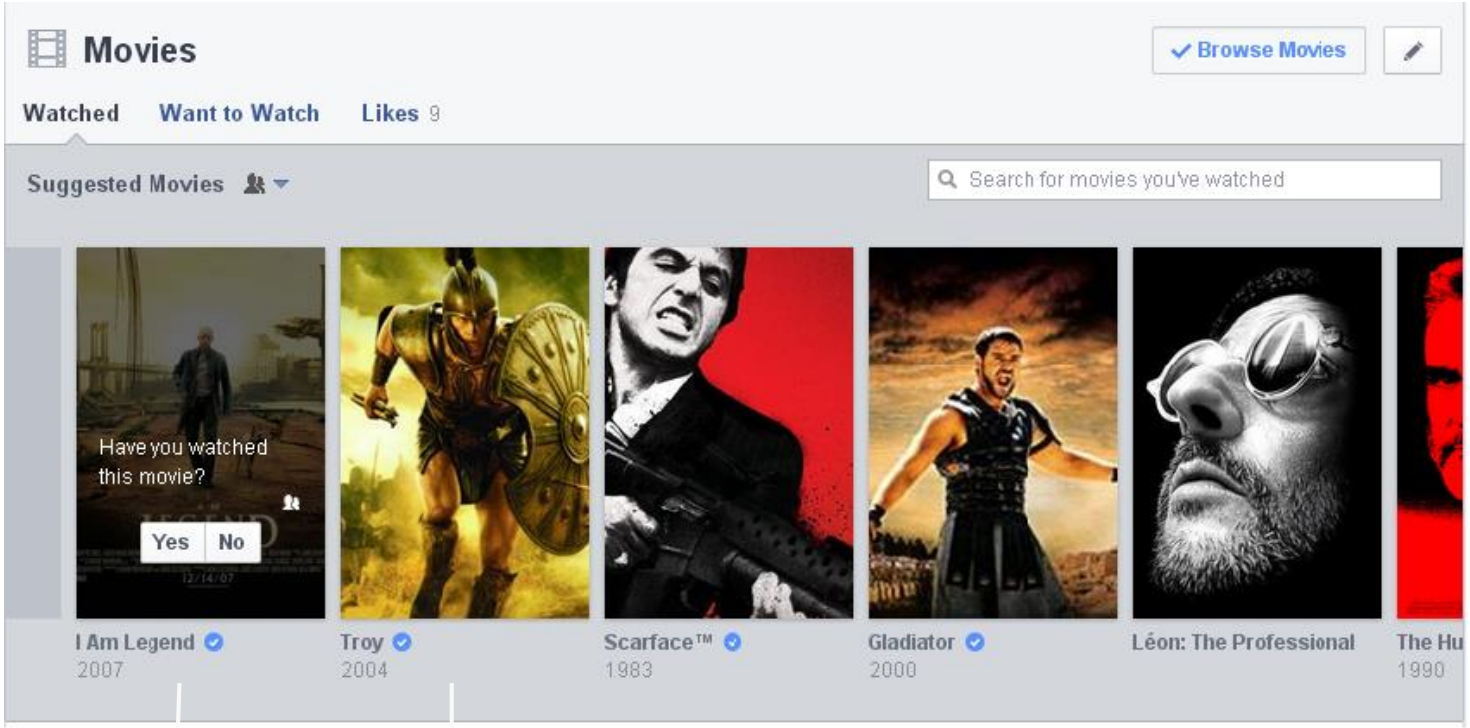
V. C. Ostuni et al., Sound and Music Recommendation with Knowledge Graphs. ACM Transactions on Intelligent Systems and Technology (TIST) - 2016 - http://sisinflab.poliba.it/publications/2016/OODSD16/

Direct item Linking



dbr:I_Am_Legend_(film)

Direct item Linking

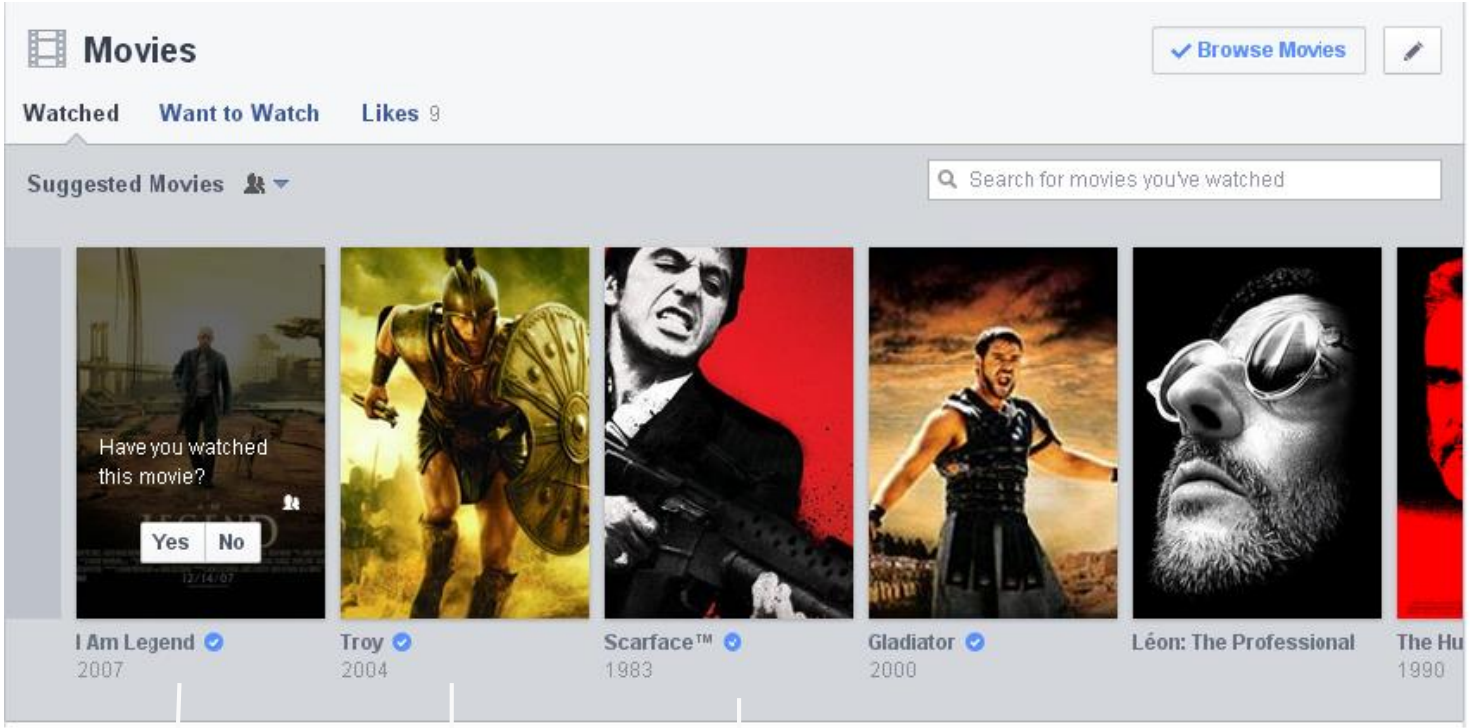


dbr:I_Am_Legend_(film)

dbr:Troy_(film)

~~dbr:Troy~~

Direct item Linking



dbr:I_Am_Legend_(film)

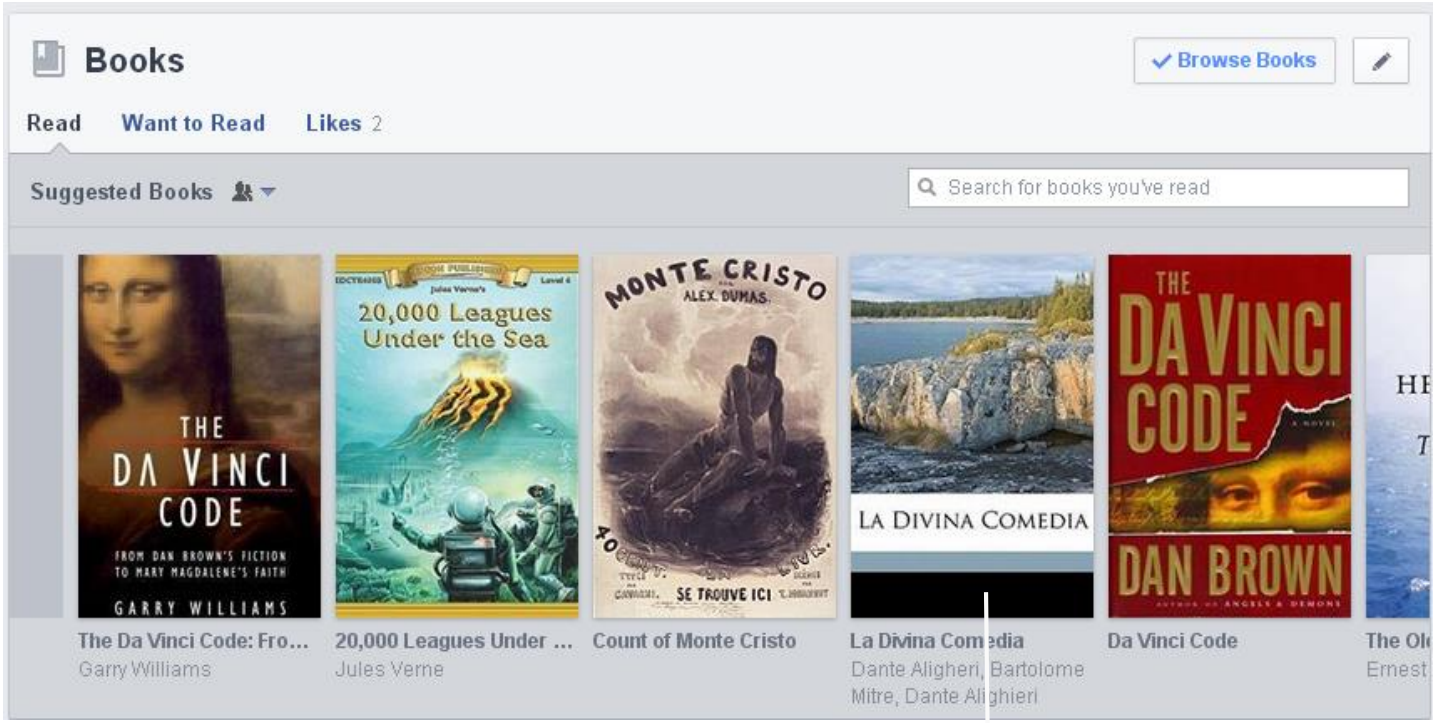
dbr:Scarface_(1983_film)

~~dbr:Scarface:_The_World_Is_Yours~~

dbr:Troy_(film)

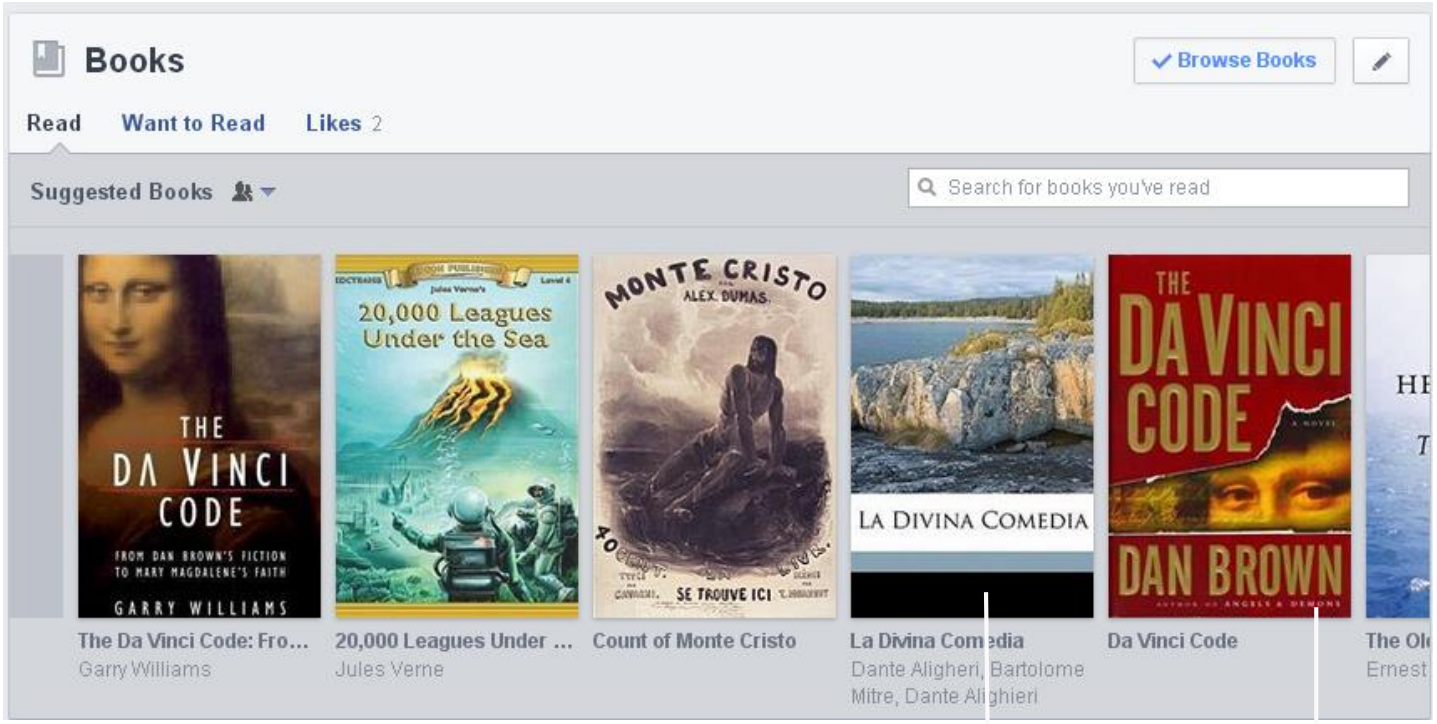
~~dbr:Troy~~

Direct item Linking



dbr:Divine_Comedy

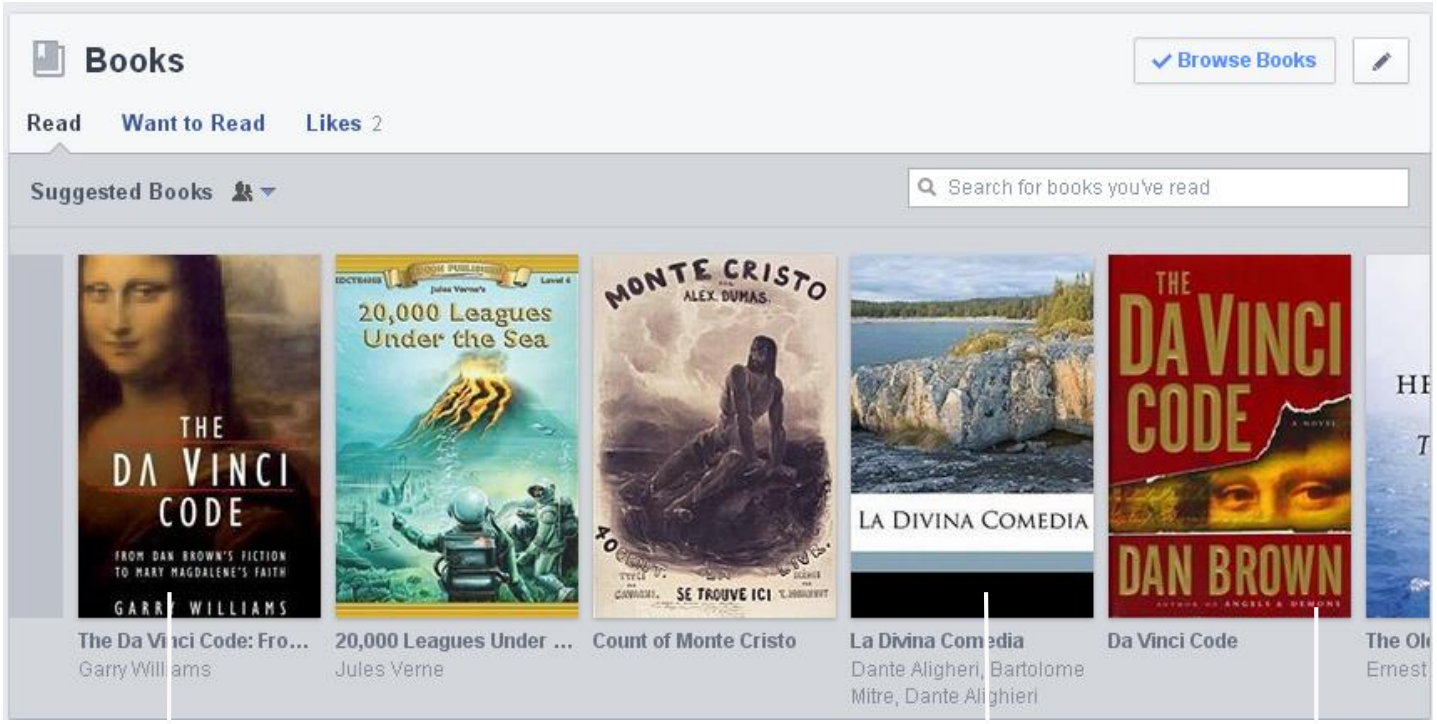
Direct item Linking



dbr:The_Da_Vinci_Code

dbr:Divine_Comedy

Direct item Linking



???

dbr:Divine_Comedy

dbr:The_Da_Vinci_Code

Direct Item Linking

How? SPARQL Query

```
SELECT DISTINCT ?uri, ?title WHERE {
    ?uri rdf:type dbpedia-owl:Film.
    ?uri rdfs:label ?title.
    FILTER langMatches(lang(?title), "EN") .
    FILTER regex(?title, "matrix", "i")
}
```

Other approaches

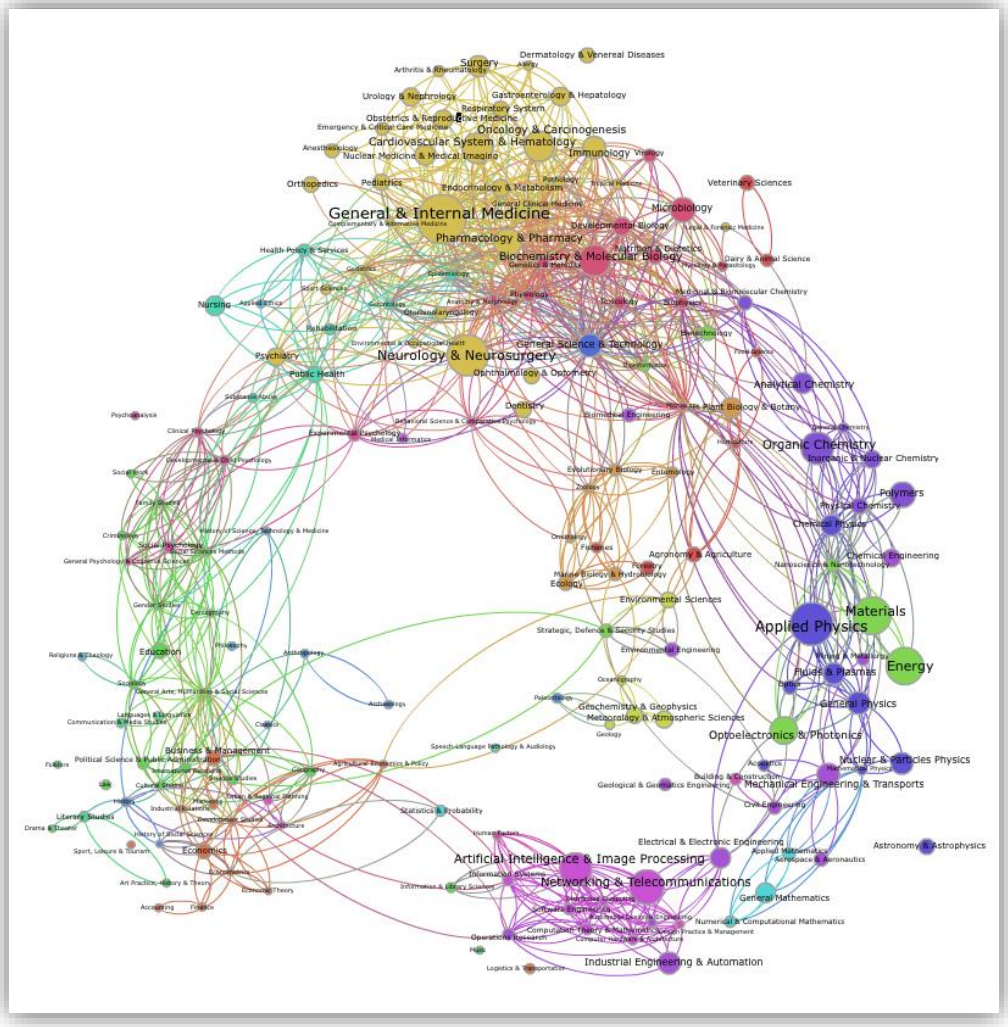
- ✓ **DBpedia Lookup**
  https://github.com/dbpedia/lookup

- ✓ **Silk Framework**
  http://silk-framework.com/

Ontologies

- used to describe **domain-specific knowledge**
- **hierarchies of concepts** with **attributes** and **relations**
- *"An **ontology** is a formal, explicit specification of a **shared** conceptualization"*

(e.g. the Science Ontology)



Guarino, Nicola. "Understanding, building and using ontologies." *International Journal of Human-Computer Studies* 46.2 (1997): 293-310.

Exogenous Semantics through Ontologies

why do we need an ontology?

to share common understanding of the structure of information

among people

among software agents

to enable reuse of domain knowledge

to avoid “re-inventing the wheel”

to introduce standards to allow **interoperability**

Exogenous Semantics through Ontologies

why do we need an ontology?

to share common understanding of the structure of information

among people

among software agents

to enable reuse of domain knowledge

to avoid “re-inventing the wheel”

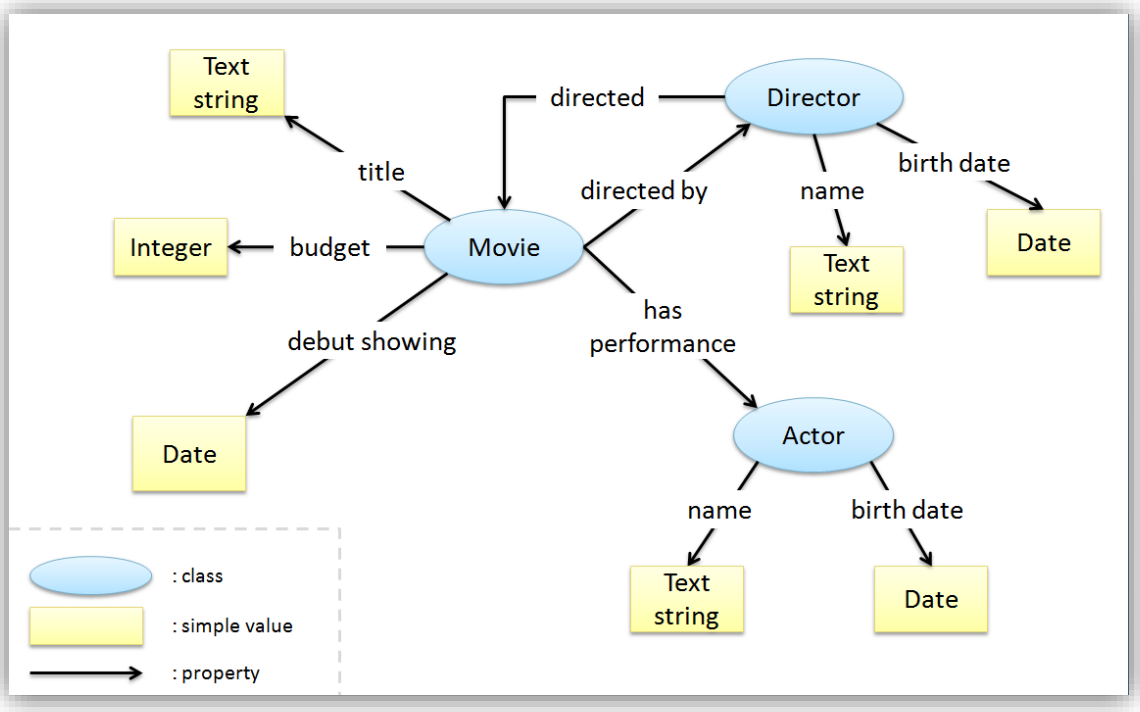
to introduce standards to allow **interoperability**

…let’s have an example!

Exogenous Semantics through Ontologies

A Movie Ontology

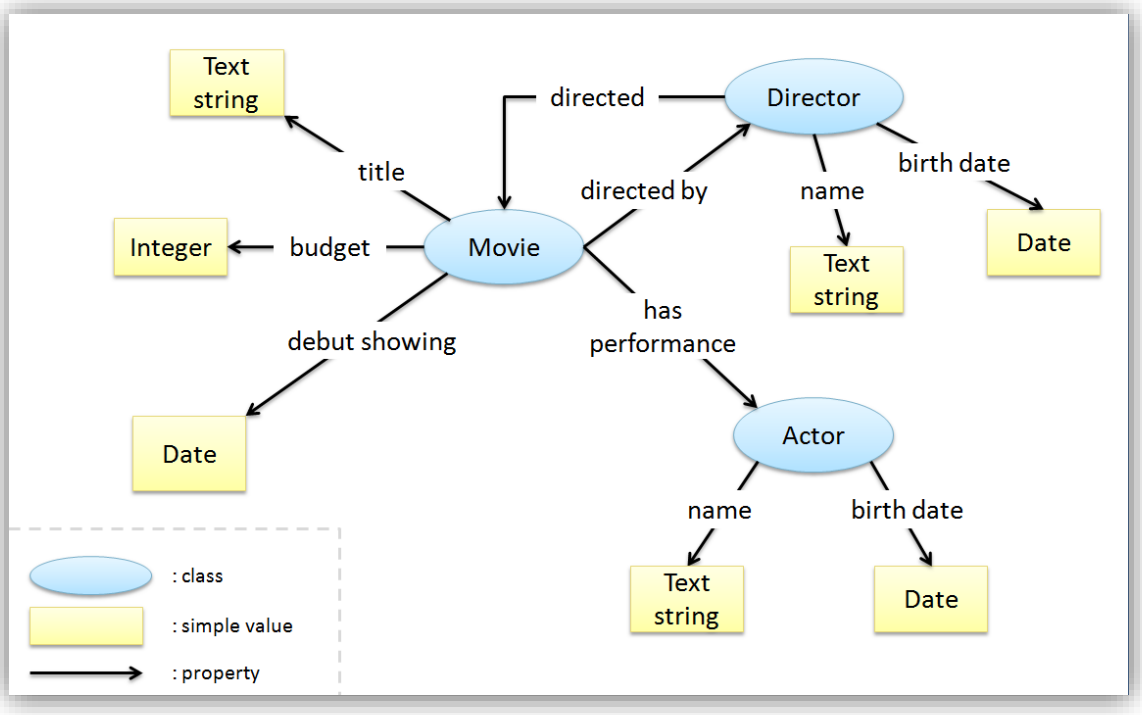
(a small portion, actually)



Exogenous Semantics through Ontologies

A Movie Ontology

(a small portion, actually)

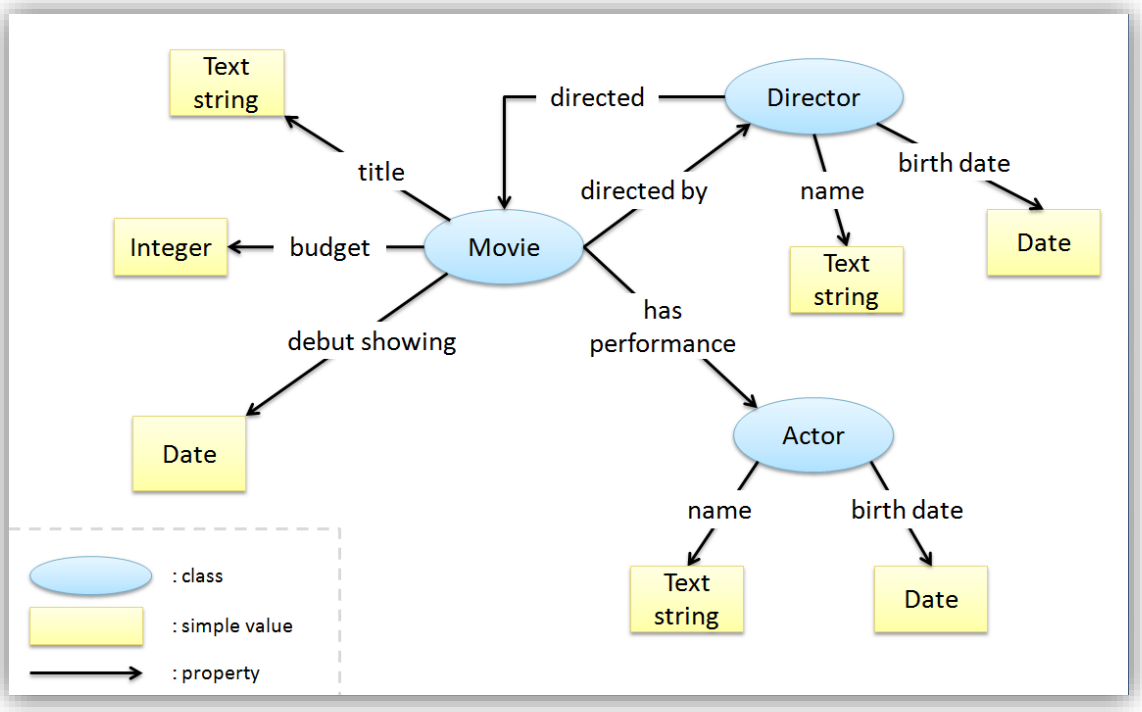


we formally encode **relevant aspects** and the **relationships among them**

Exogenous Semantics through Ontologies

A Movie Ontology

(a small portion, actually)



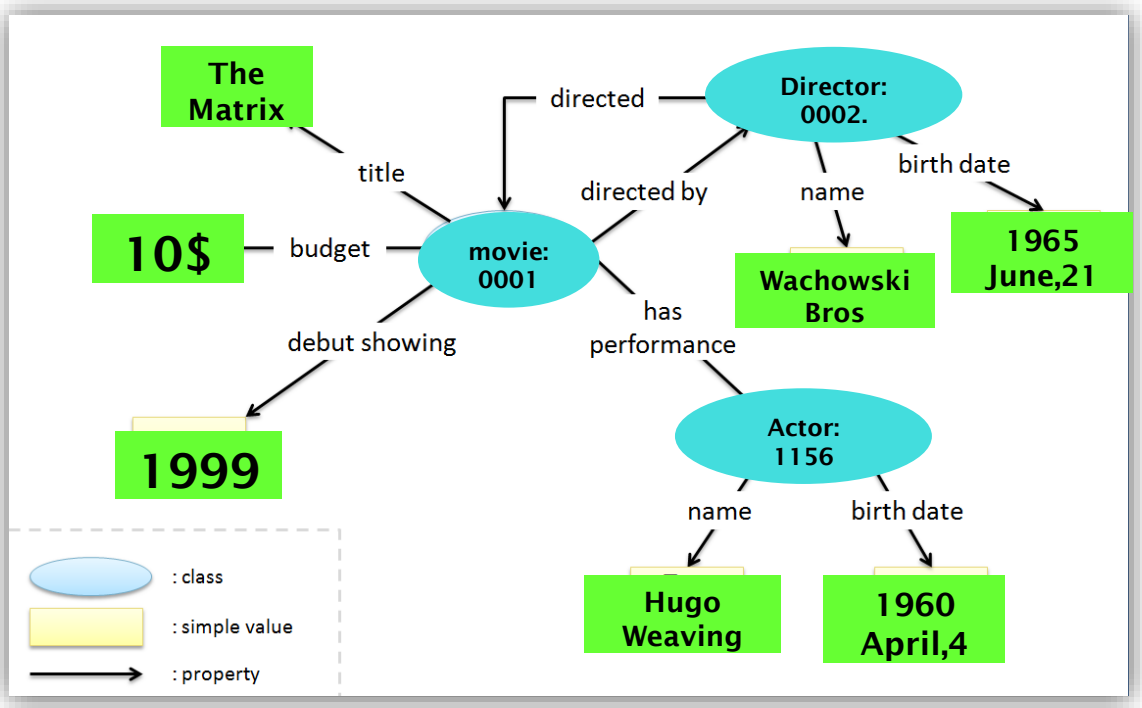
every item formally modeled **by following this structure**

and encoded through a Semantic Web language (e.g. OWL, RDF)

Exogenous Semantics through Ontologies

A Movie Ontology

(a small portion, actually)



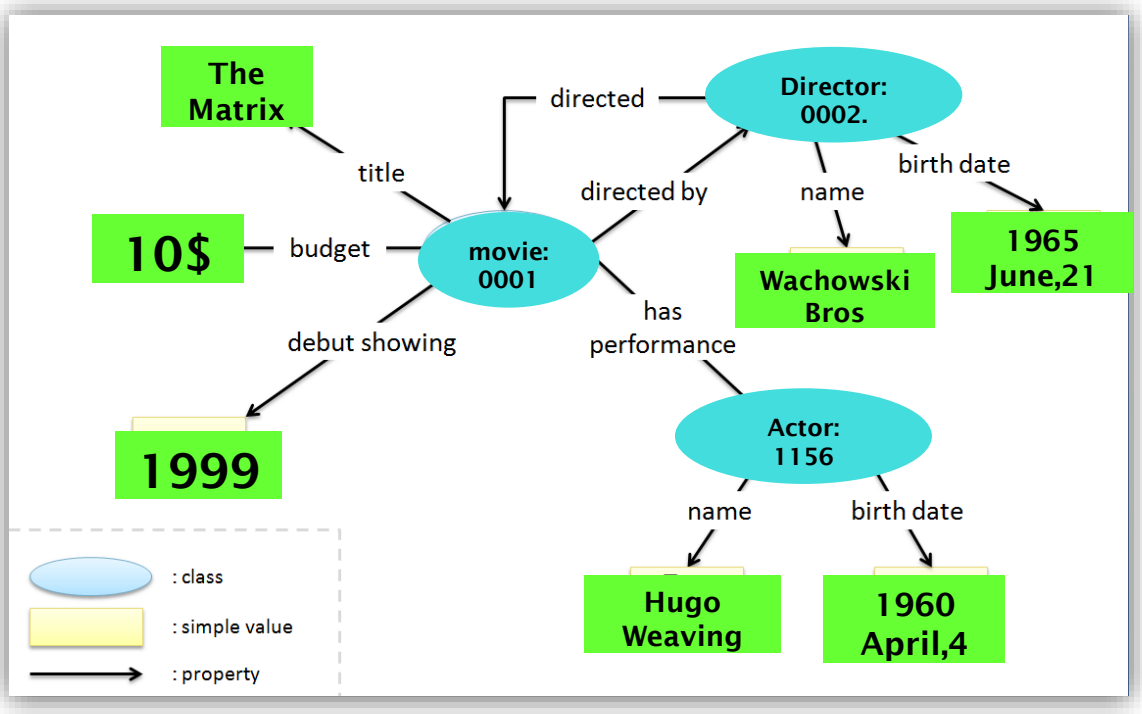
every item formally modeled **by following this structure**

and encoded through a Semantic Web language (e.g. OWL, RDF)

Exogenous Semantics through Ontologies

A Movie Ontology

(a small portion, actually)

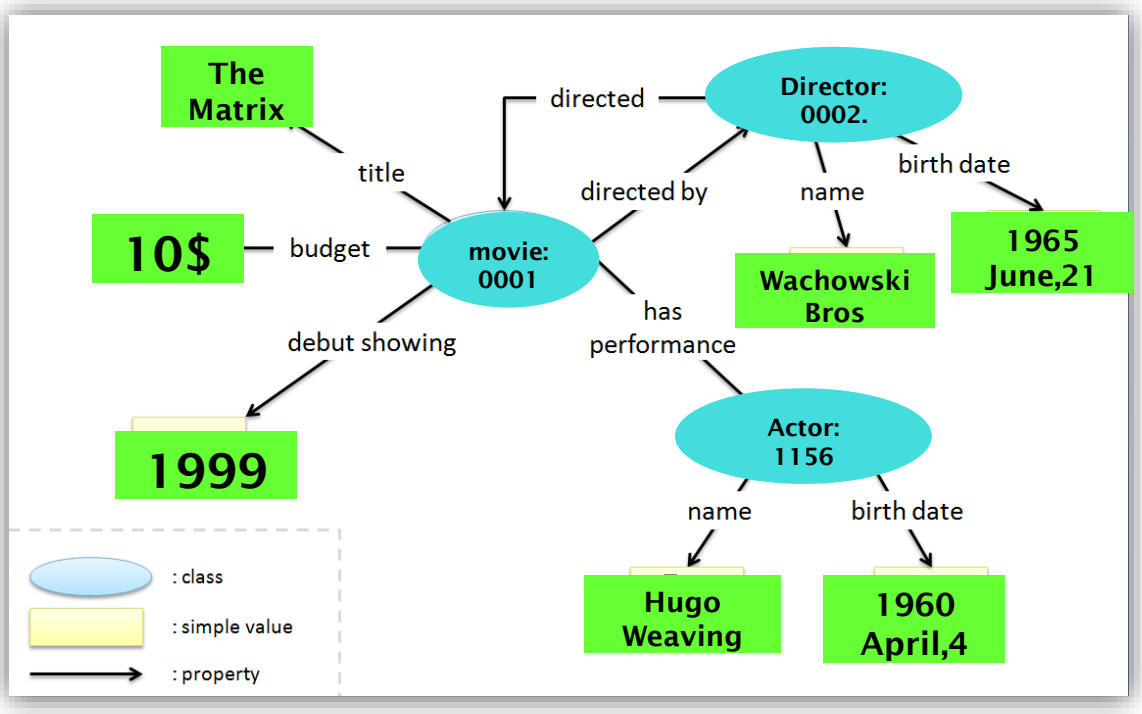


why is it useful?

Exogenous Semantics through Ontologies

A Movie Ontology

(a small portion, actually)



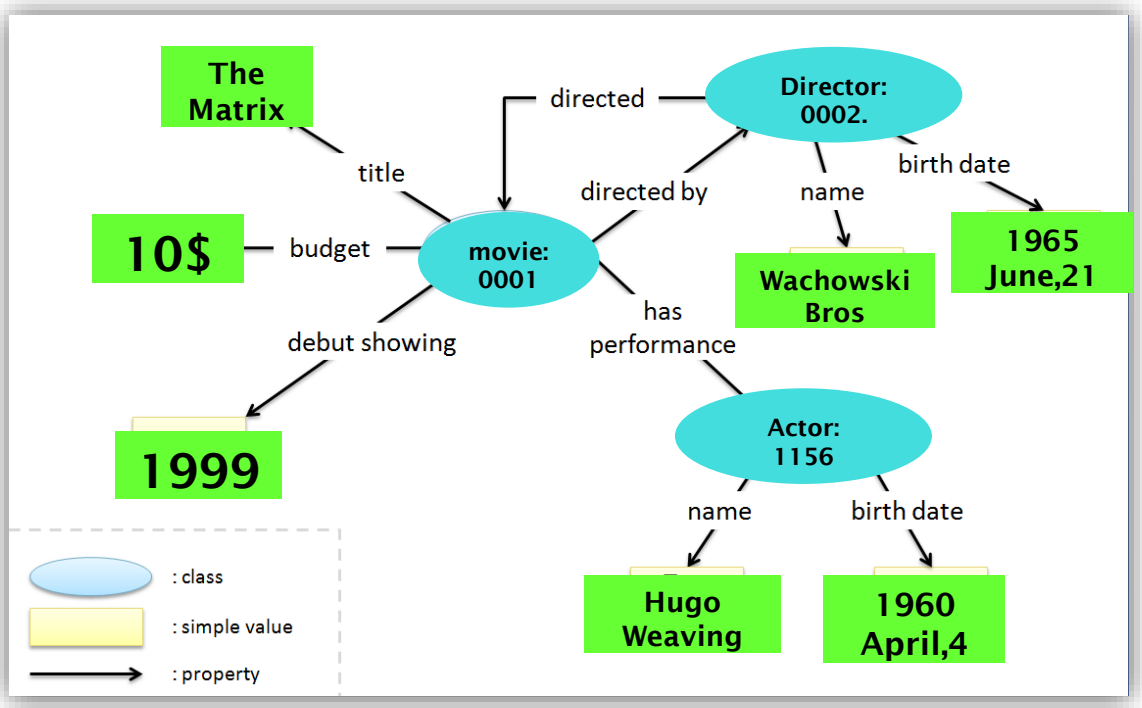
why is it useful?

each feature has a non-ambiguous «meaning»

Exogenous Semantics through Ontologies

A Movie Ontology

(a small portion, actually)



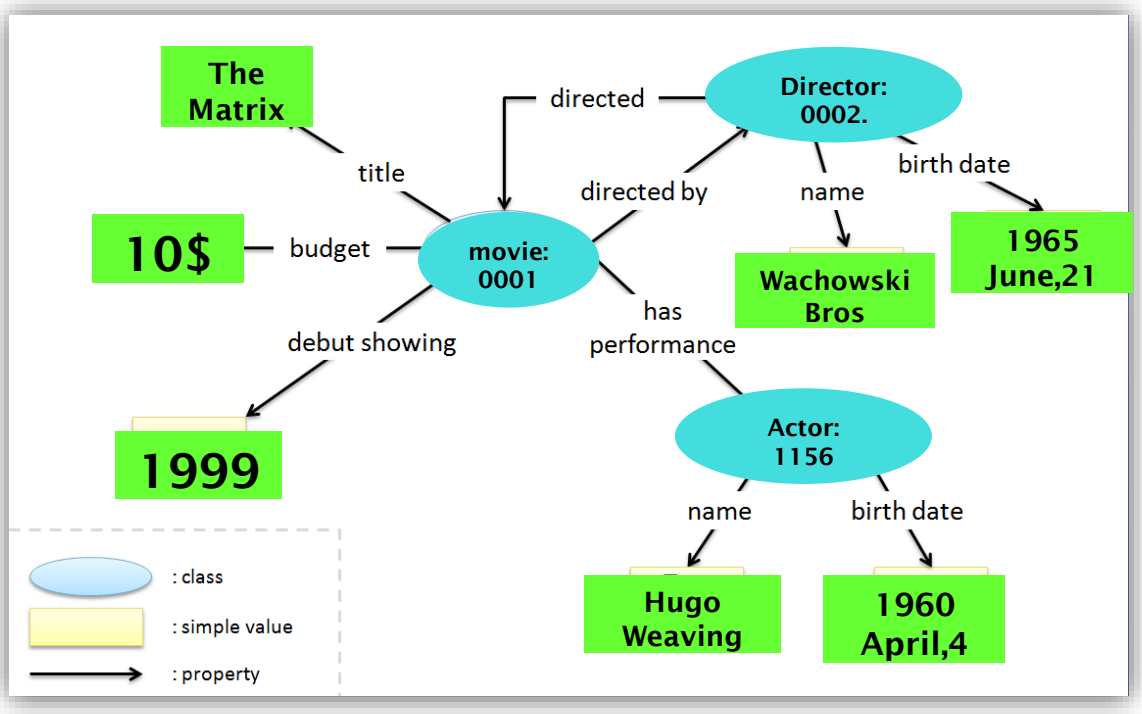
why is it useful?

we don't need to process unstructured content

Exogenous Semantics through Ontologies

A Movie Ontology

(a small portion, actually)



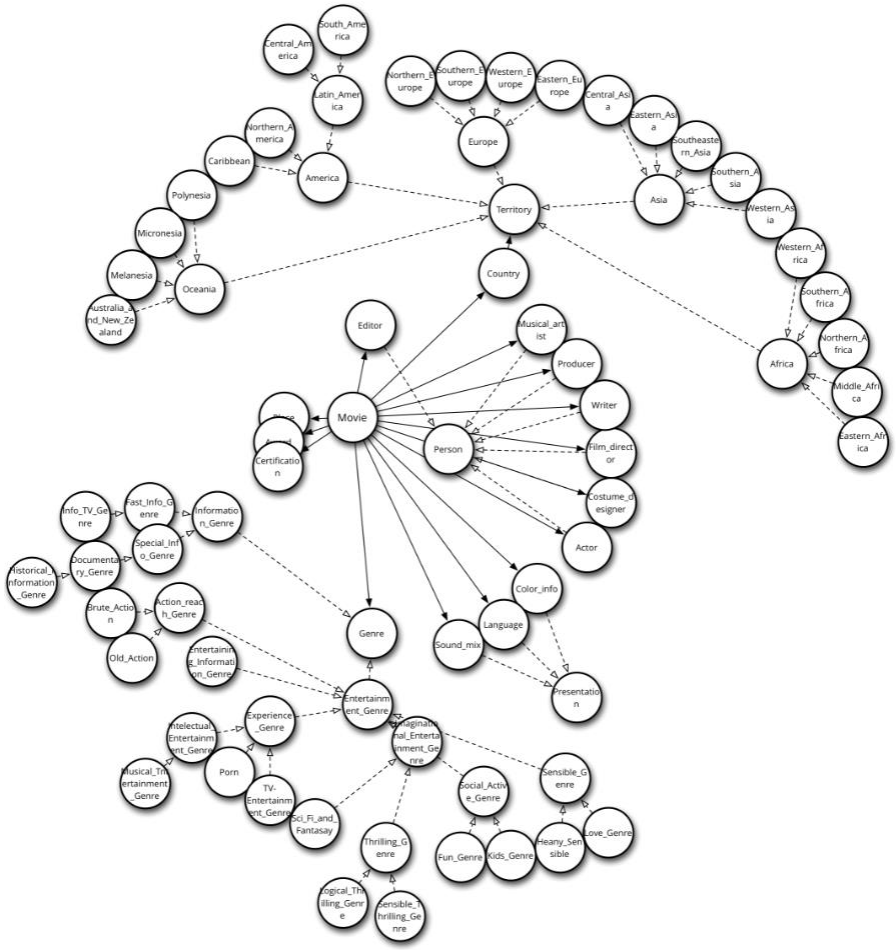
why is it useful?

we can perform some «reasoning» on user preferences. How?

Exogenous Semantics through Ontologies

The Movie Ontology

http://www.movieontology.org/



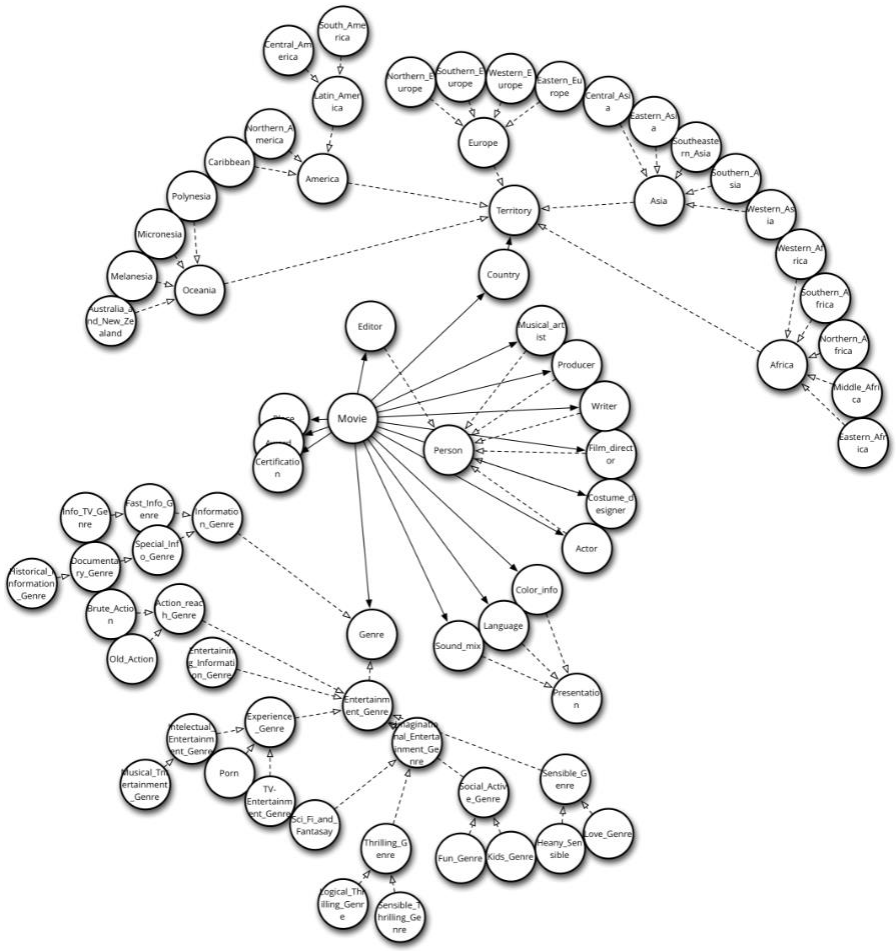
explicit Relationship between the classes are **encoded**.

e.g. Genre «Love» is a **subclass of** «Sensible Love» *that is a subclass of* «Entertainment Genre» which is a property of the **class Movie**.

Exogenous Semantics through Ontologies

The Movie Ontology

http://www.movieontology.org/



We can reason on the preferences and infer that a user interested in **The Matrix** (SciFi_and_Fantasy genre) is interested in **Imaginational_Entertainment** and potentially in **Logical_Thrilling**

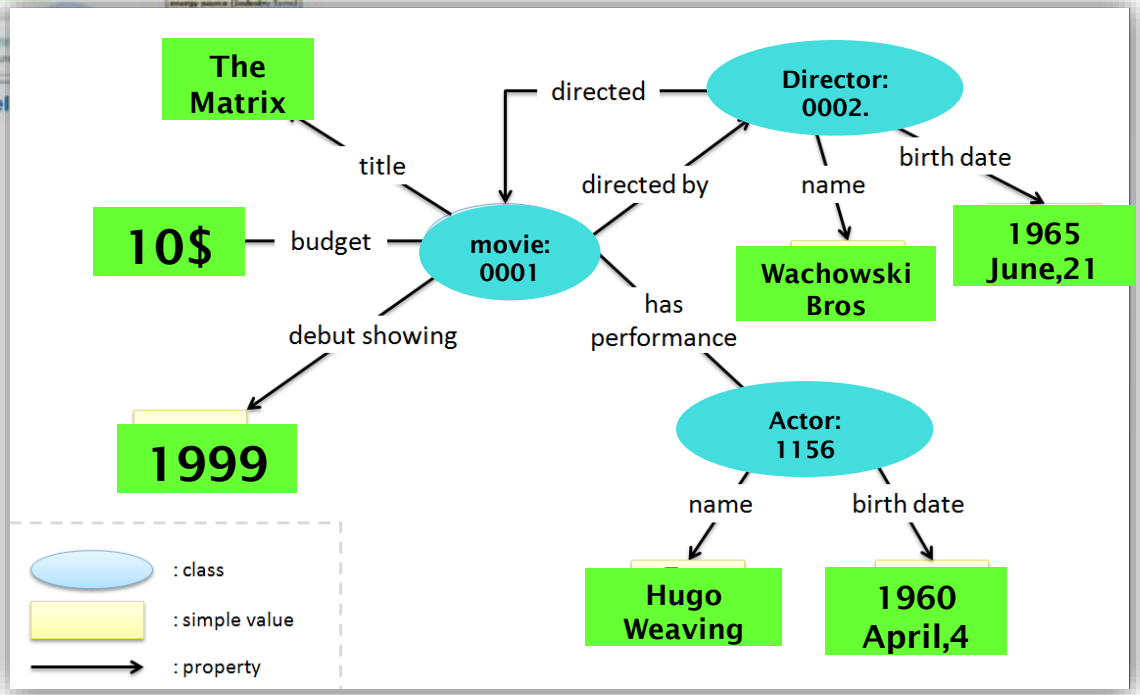
The Matrix representation

from a

flat representation

toward a

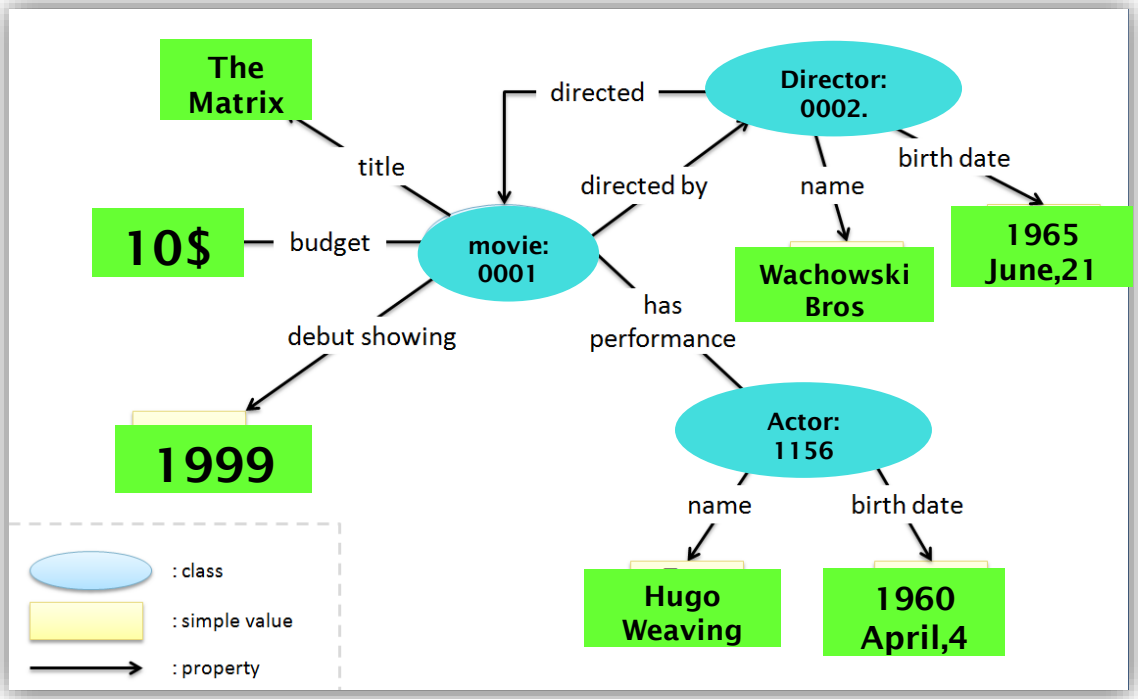
graph-based representation



The Matrix representation

from a
flat representation

toward a
graph-based representation



semantics explicitly encoded

explicit relations between concepts exist: reasoning to infer interesting information

ontologies typically domain-dependant

huge effort to build and mantain an ontology

very huge effort to populate an ontology!

Vision


WE NEED FORMAL SEMANTICS
WITHOUT THE EFFORT OF BUILDING AND POPULATING ONTOLOGIES
imgflip.com

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Ontologies

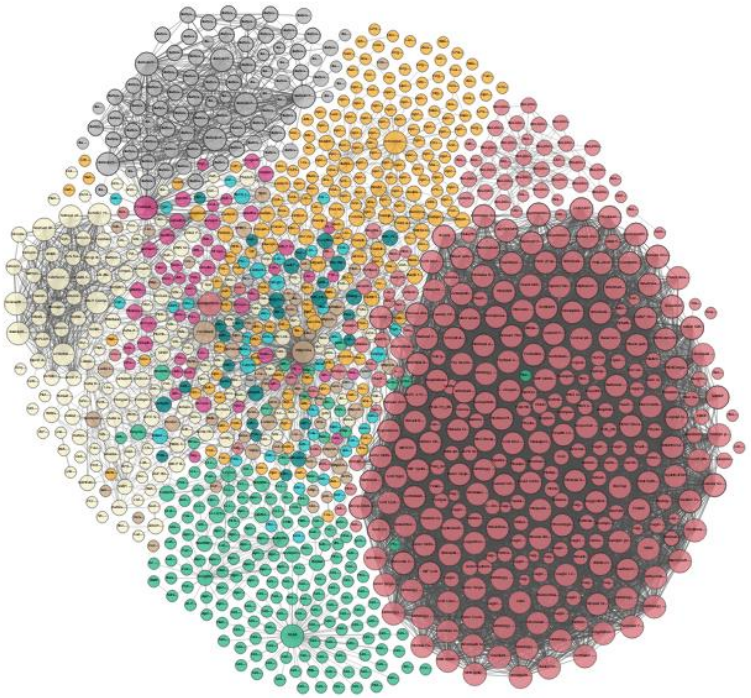
Linked Open Data

DBpedia

.....

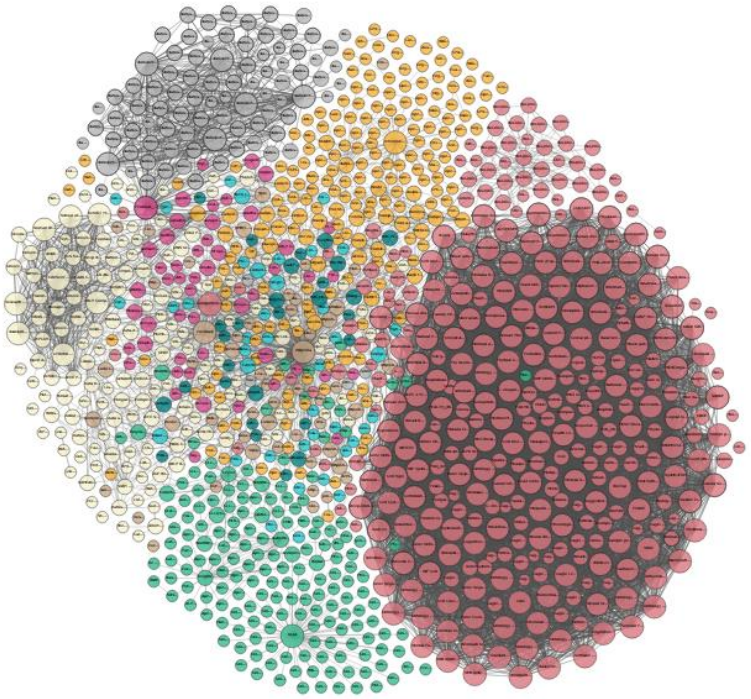
Linked Open Data

the **giant** **global graph**



Linked Open Data (cloud)

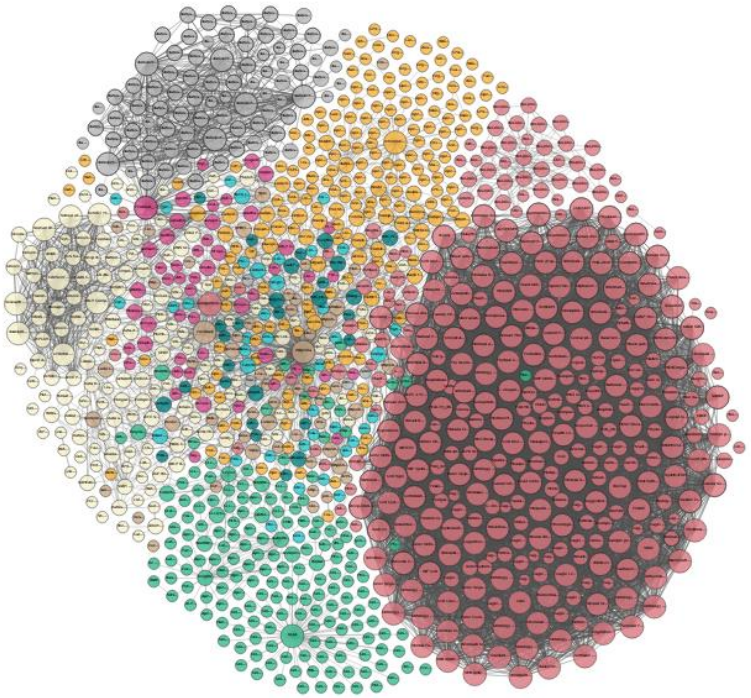
what is it?



(large) set of **interconnected** **semantic** datasets

Linked Open Data (cloud)

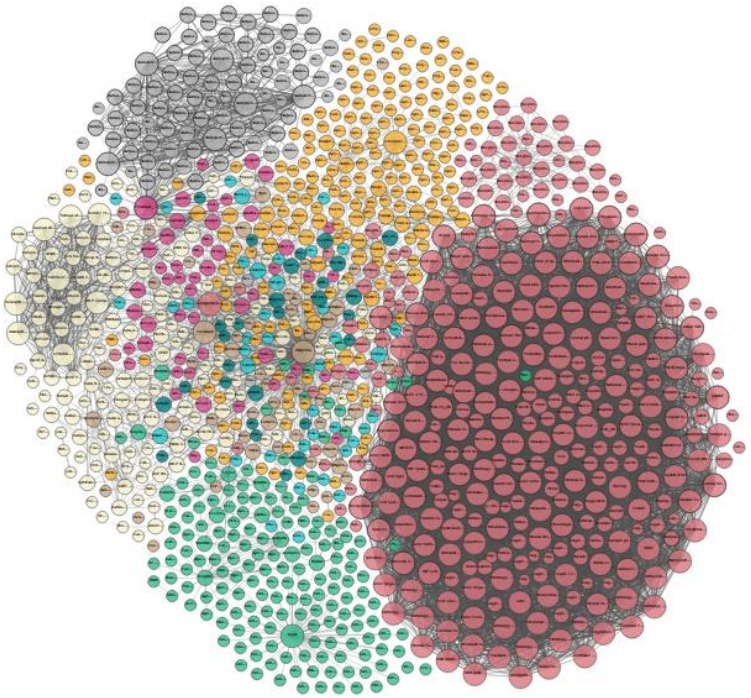
statistics



149 billions triples, **3,842** datasets (http://stats.lod2.eu)

Linked Open Data (cloud)

DBpedia



core of the **LOD cloud**

RDF mapping of **Wikipedia**

Linked Open Data

cornerstones

methodology to **publish**, **share** and **link**
structured data on the Web

use of **RDF** every resource/entity/relation identified by a (unique)
URI http://dbpedia.org/resource/Halifax

re-use of existing properties to express an
agreed semantics and **connect data sources**

Which vocabularies/ontologies?

Most popular on **http://prefix.cc** (June 6, 2016)

- ✓ **YAGO:** http://yago-knowledge.org/resource/
- ✓ **FOAF:** http://xmlns.com/foaf/0.1/
- ✓ **DBpedia Ontology:** http://dbpedia.org/ontology/
- ✓ **DBpedia Properties:** http://dbpedia.org/property/
- ✓ **Dublin Core:** http://dublincore.org/

Most popular on **http://lov.okfn.org** (June 6, 2016)

- ✓ **VANN:** http://purl.org/vocab/vann/
- ✓ **SKOS:** http://www.w3.org/2004/02/skos/core
- ✓ **FOAF**
- ✓ **DCTERMS**
- ✓ **DCE:** http://purl.org/dc/elements/1.1/

Linked Open Data

	Web	Linked Data
Analogy	File System	Database
Designed for	Men	Machines (Software Agents)
Main elements	Documents	Things
Links between	Documents	Things
Semantics	Implicit	Explicit

Courtesy of Prof. Enrico Motta, The Open University, Milton Keynes – Uk – Semantic Web: Technologies and Applications.

Linked Open Data

Direct Item Linking through SPARQL query

```
SELECT DISTINCT ?uri, ?title WHERE {
   ?uri rdf:type dbpedia-owl:Film.
   ?uri rdfs:label ?title.
   FILTER langMatches(lang(?title), "EN") .
   FILTER regex(?title, "matrix", "i")
}
```

Other approaches: Dbpedia Lookup

https://github.com/dbpedia/lookup

Linked Open Data (cloud)

The Matrix

From Wikipedia, the free encyclopedia

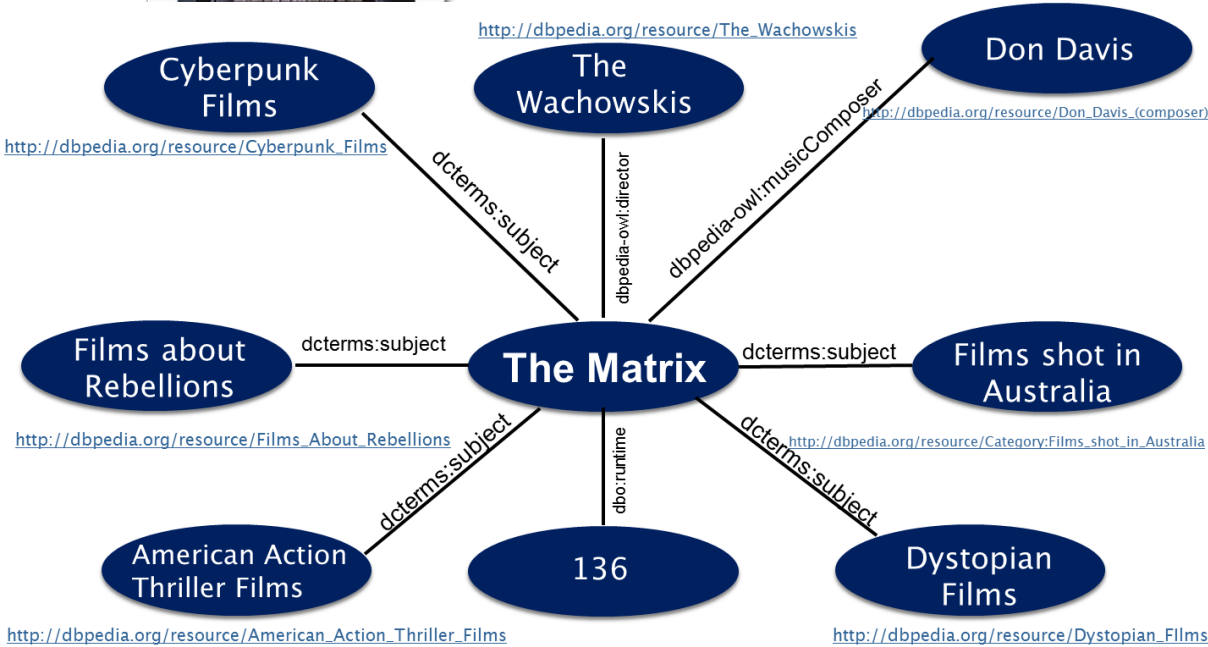
This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

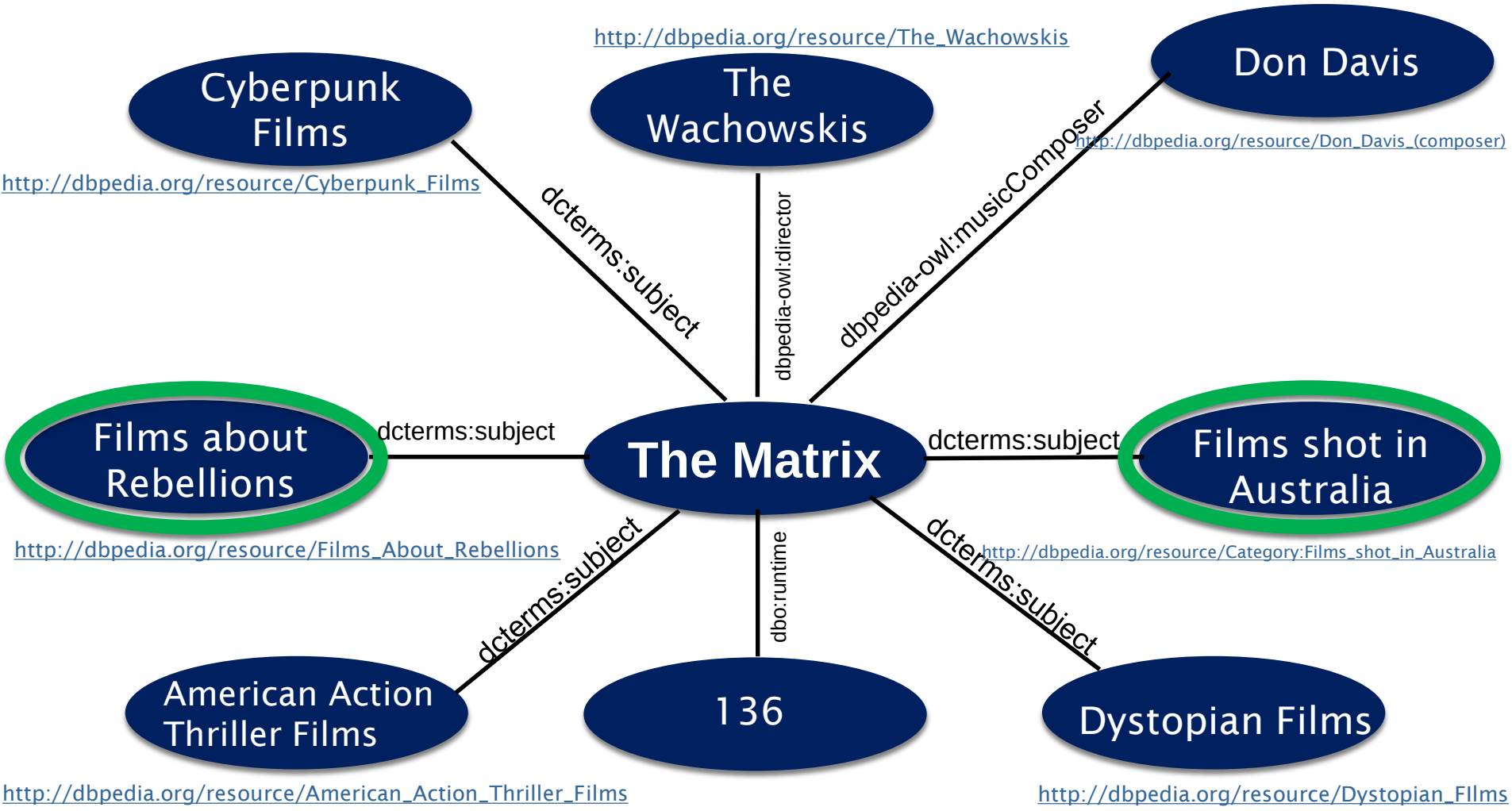
The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million wo

The Matrix



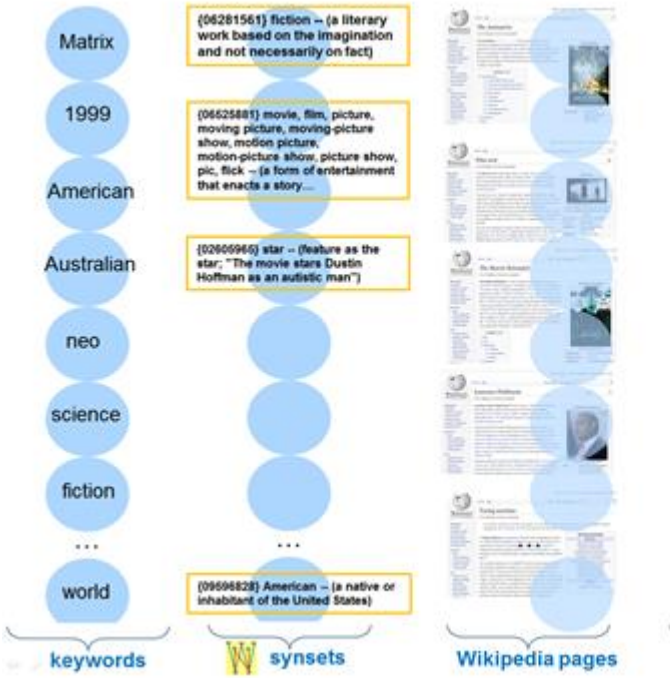
Linked Open Data (cloud)

representation



interesting non-trivial features come into play

The Matrix representation

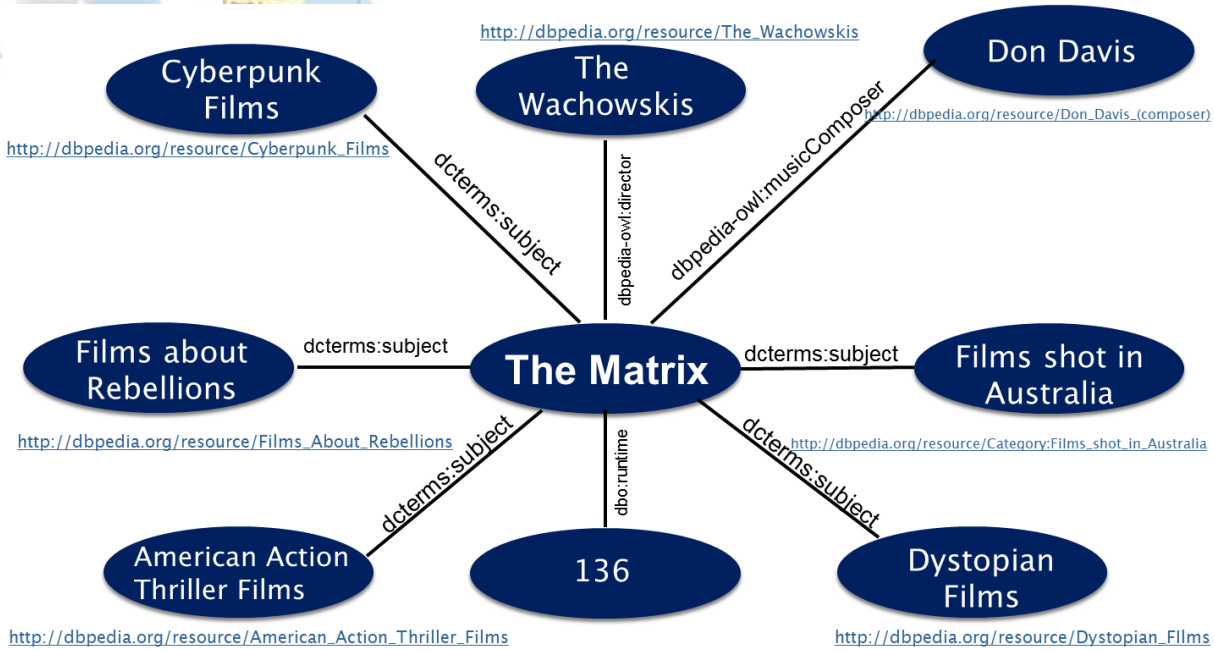


from a

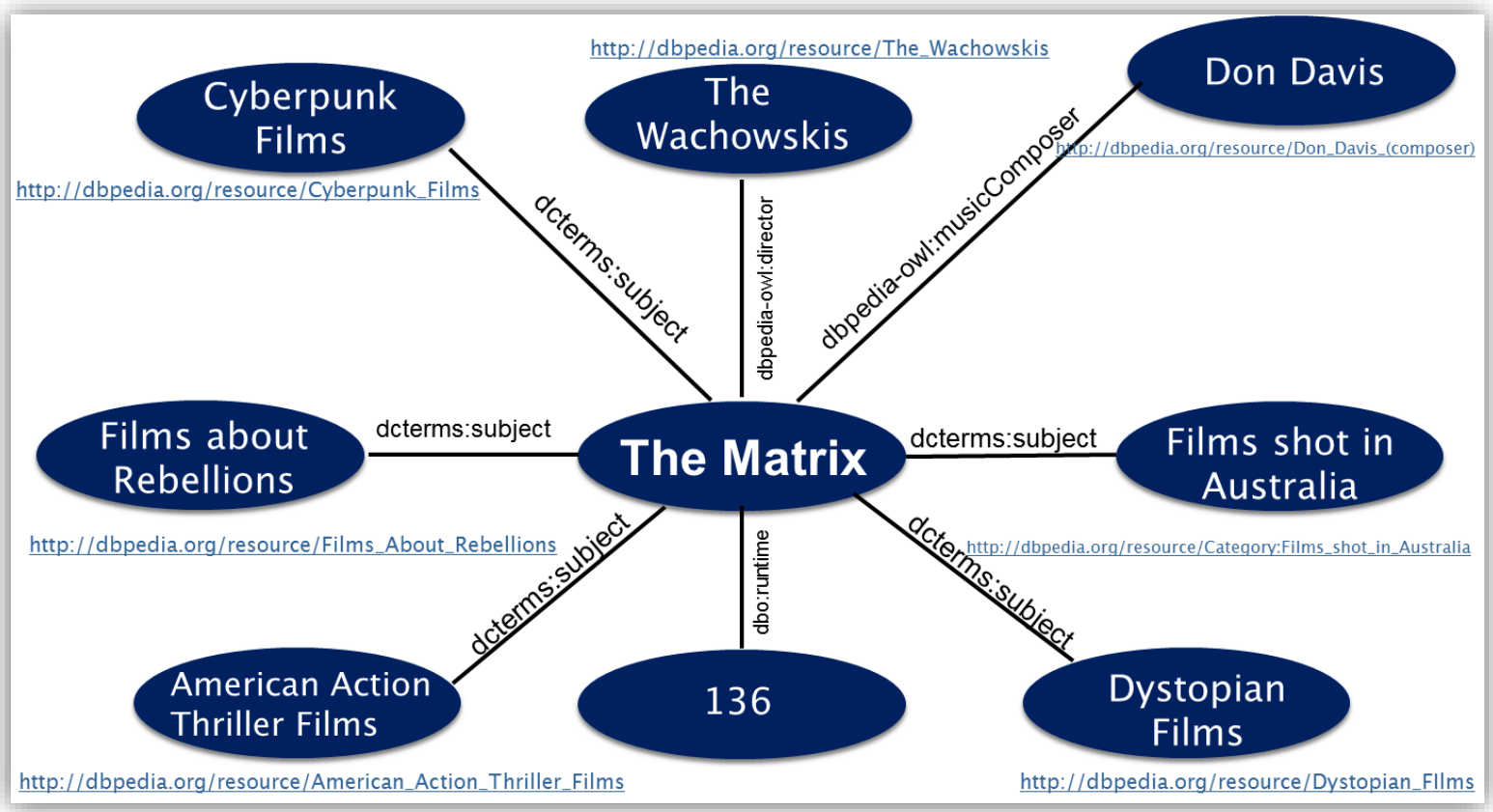
flat representation

toward a

(richer) **graph-based representation**



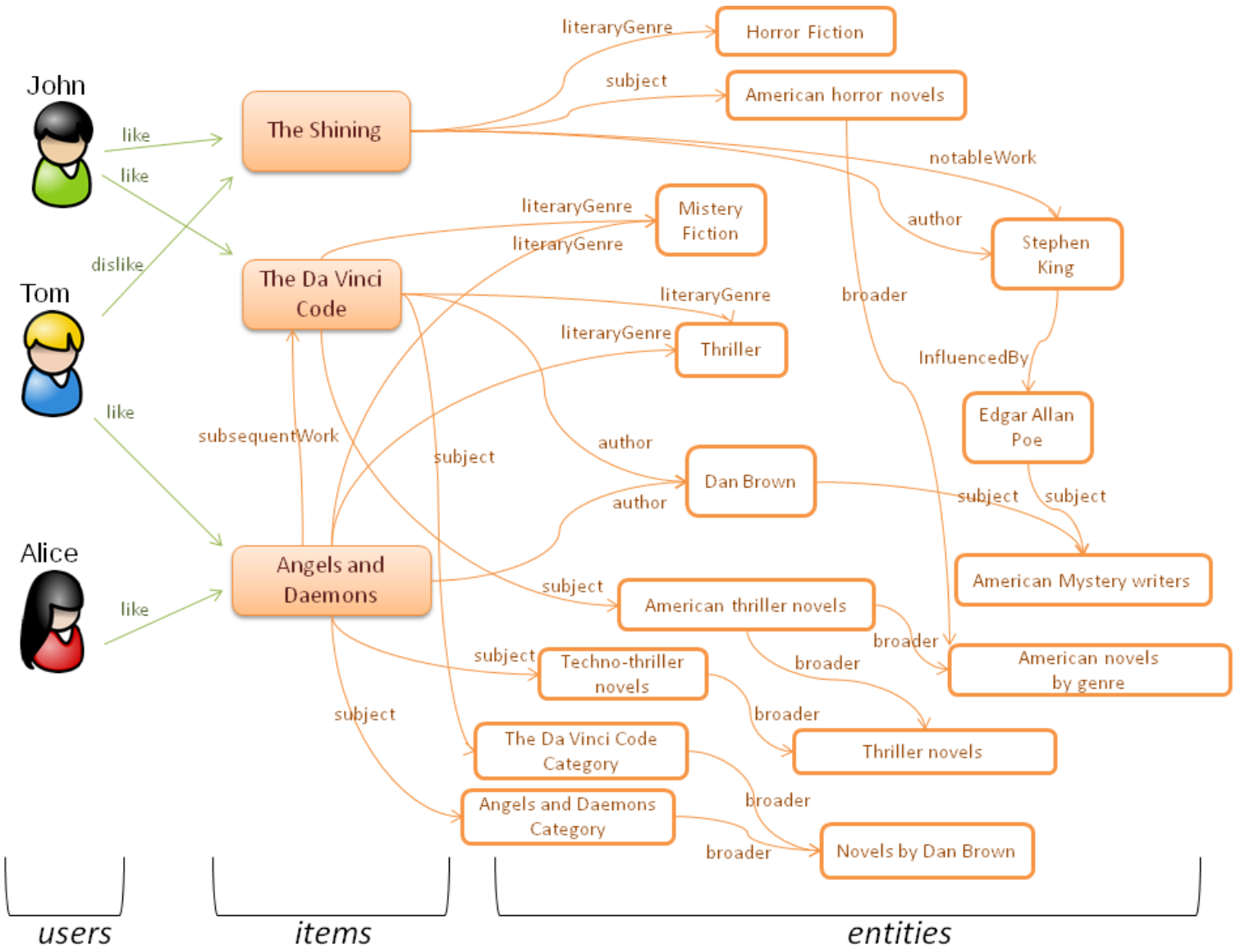
The Matrix representation



we have the advantage of **formal semantics defined in RDF**, with **interesting features coming from Wikipedia**

without the need of building and **manually populating** an ontology

Direct Item Linking



Ontologies vs Linked Open Data

Ontologies

domain-specific

less ambiguous

Linked Open Data

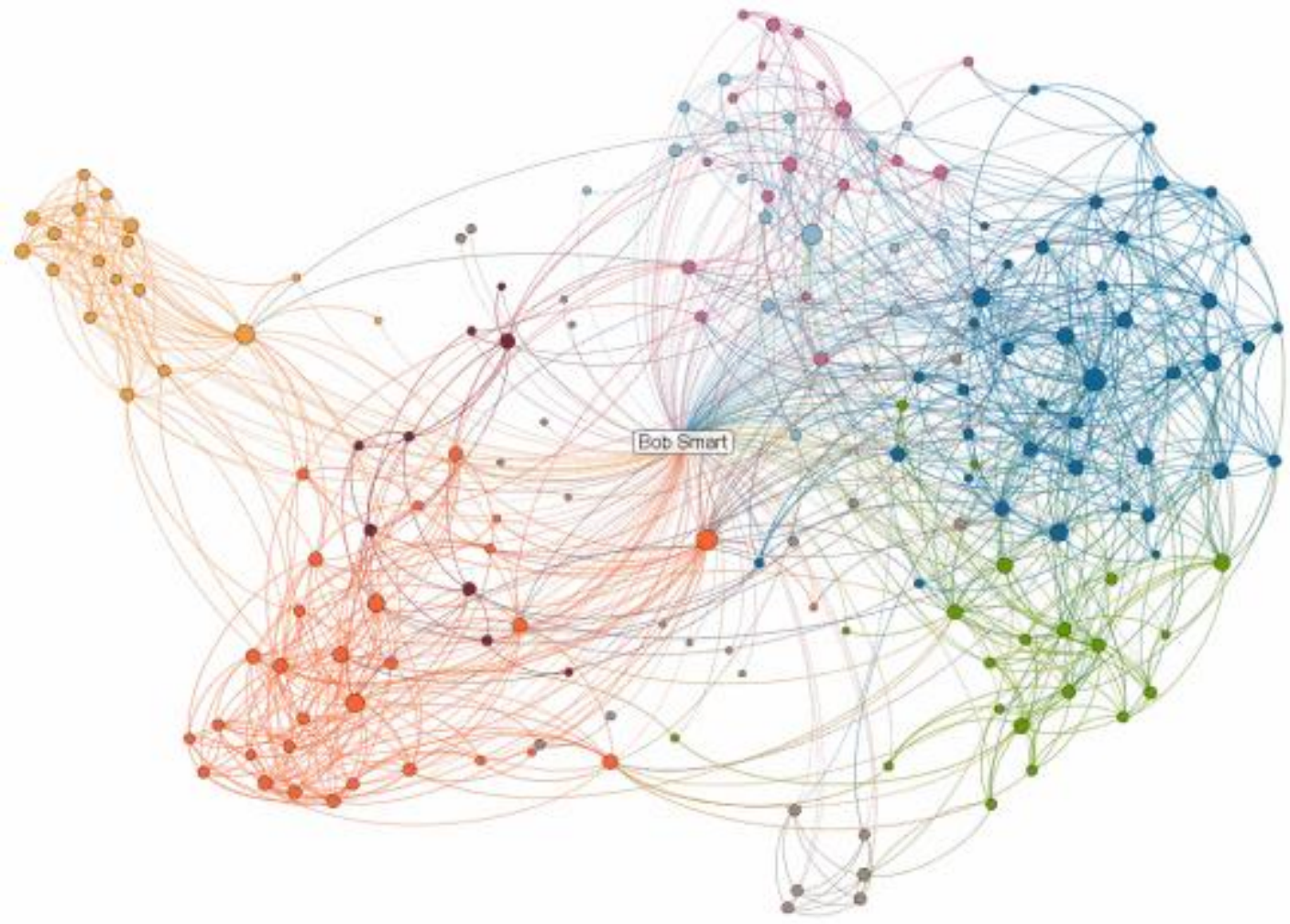
free and covers **many domains**

under **constant development**

based on Wikipedia: highly **accurate**

[Giles05]

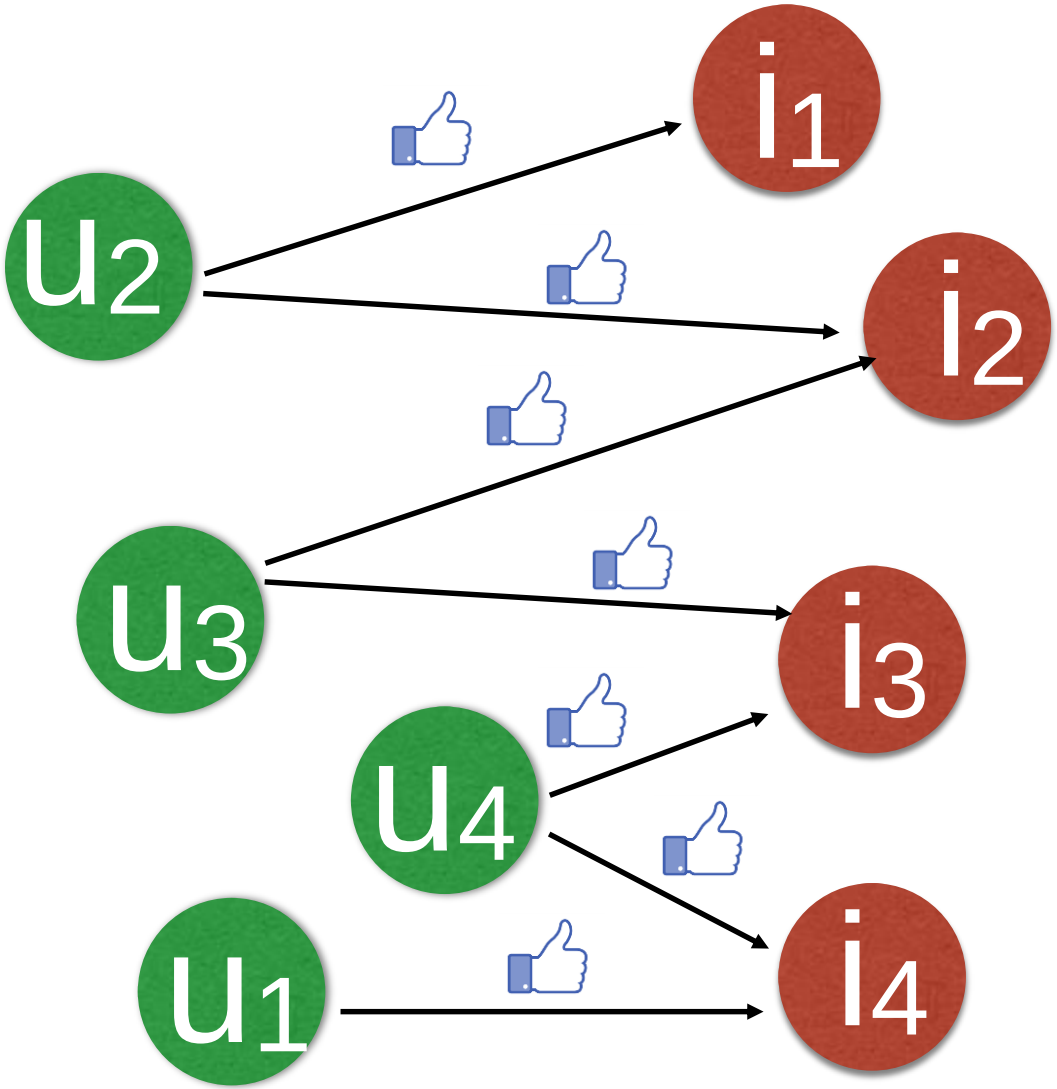
[Giles05] J. Giles. Internet Encyclopaedias Go Head to Head. *Nature*, 438:900–901, 2005.



another advantage

graph-based data models can be exploited to define more **semantic features** based on **graph topology**

Graph-based Data Model

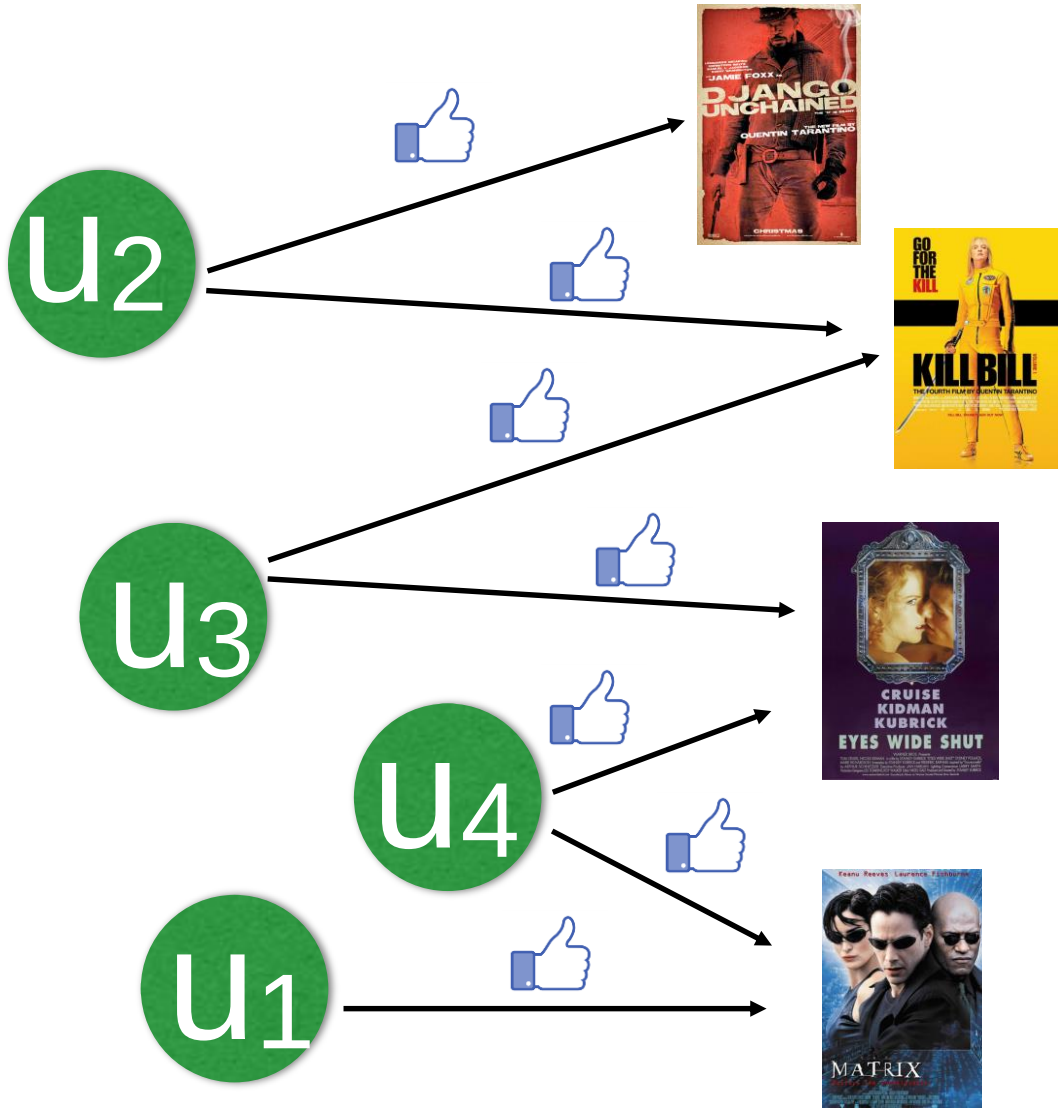


users = **nodes**
items = **nodes**
preferences = **edges**

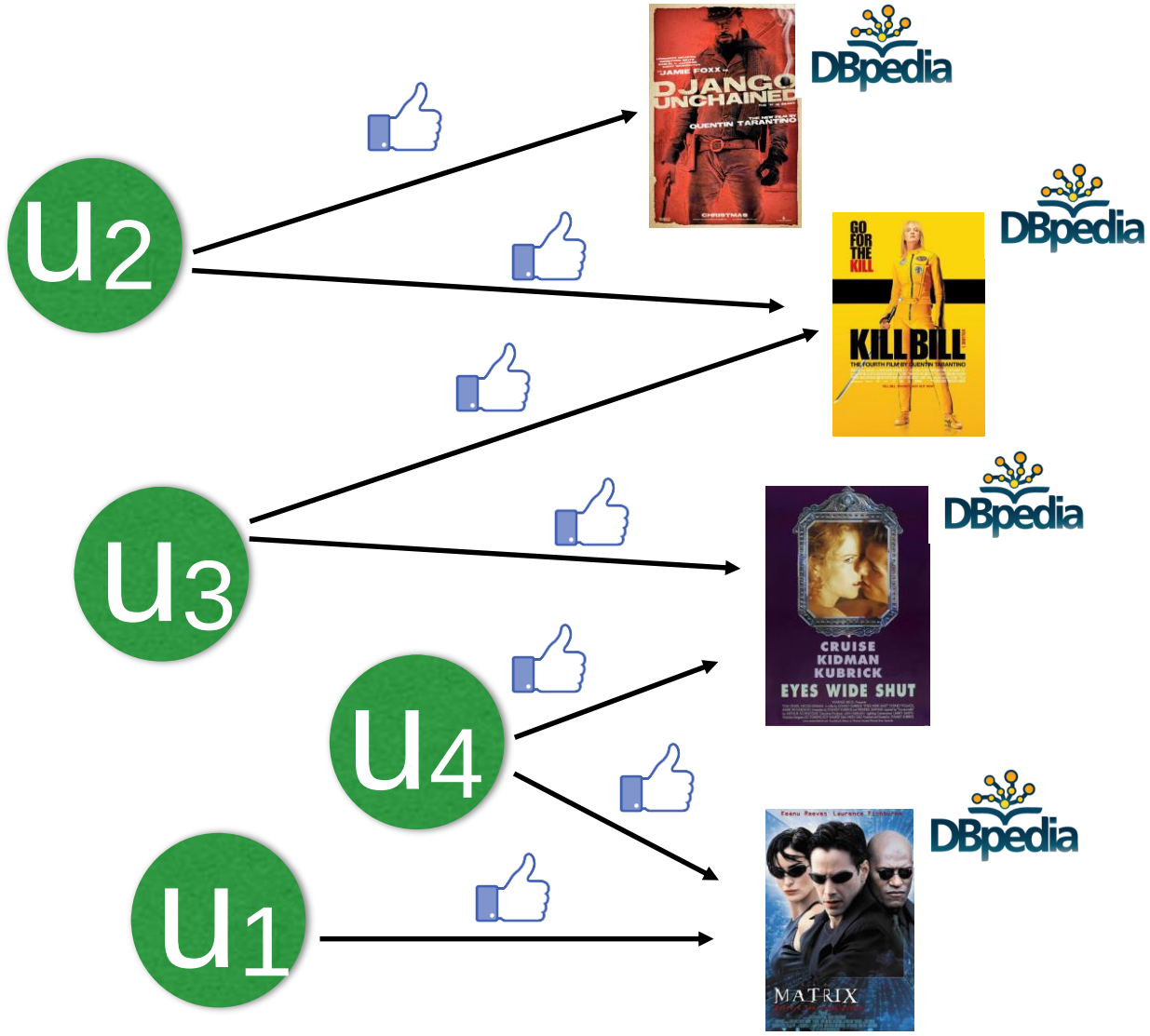
(bipartite graph)

Very intuitive representation!

Semantic Graph-based Data Model



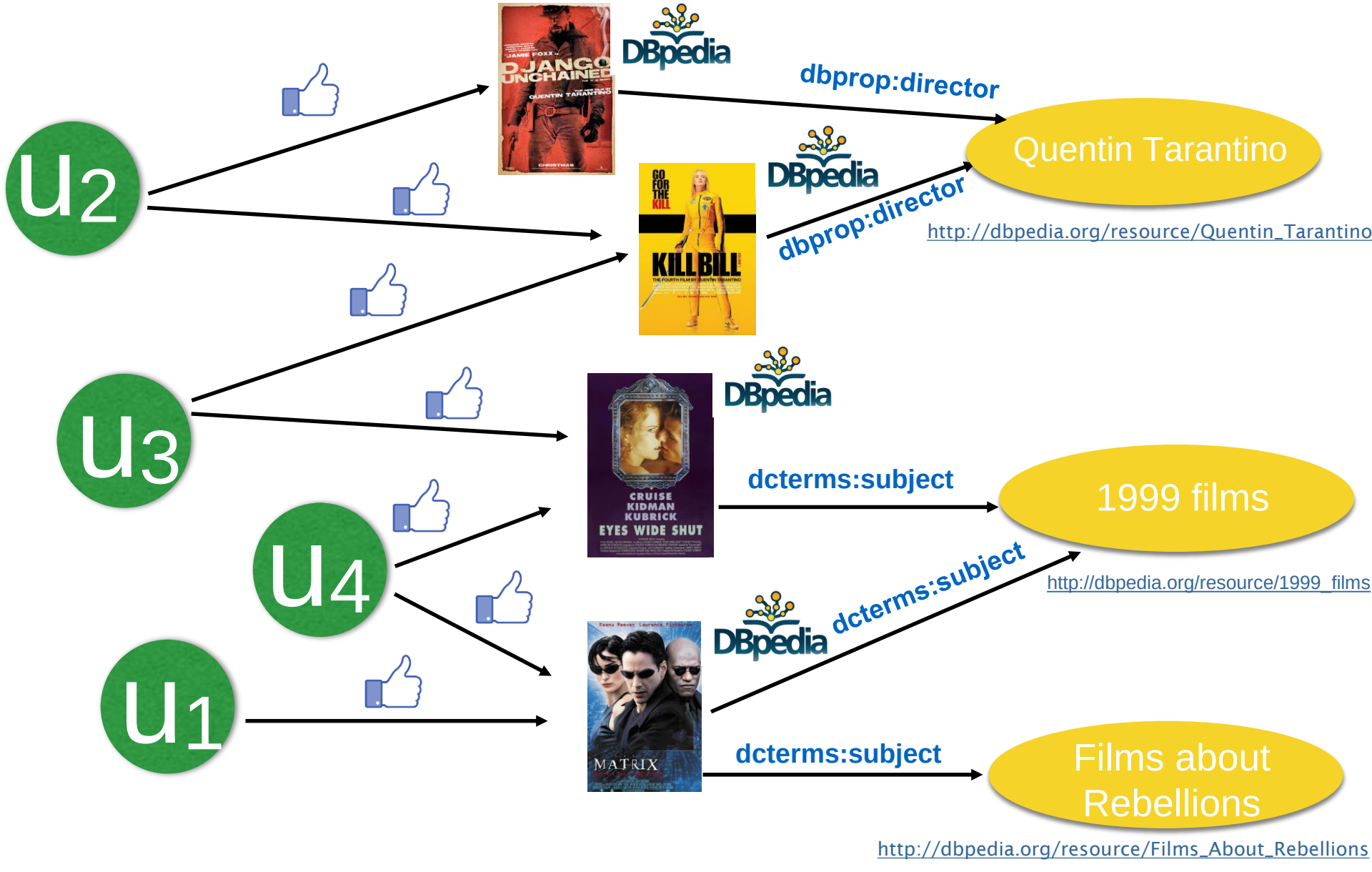
Semantic Graph-based Data Model



DBpedia mapping

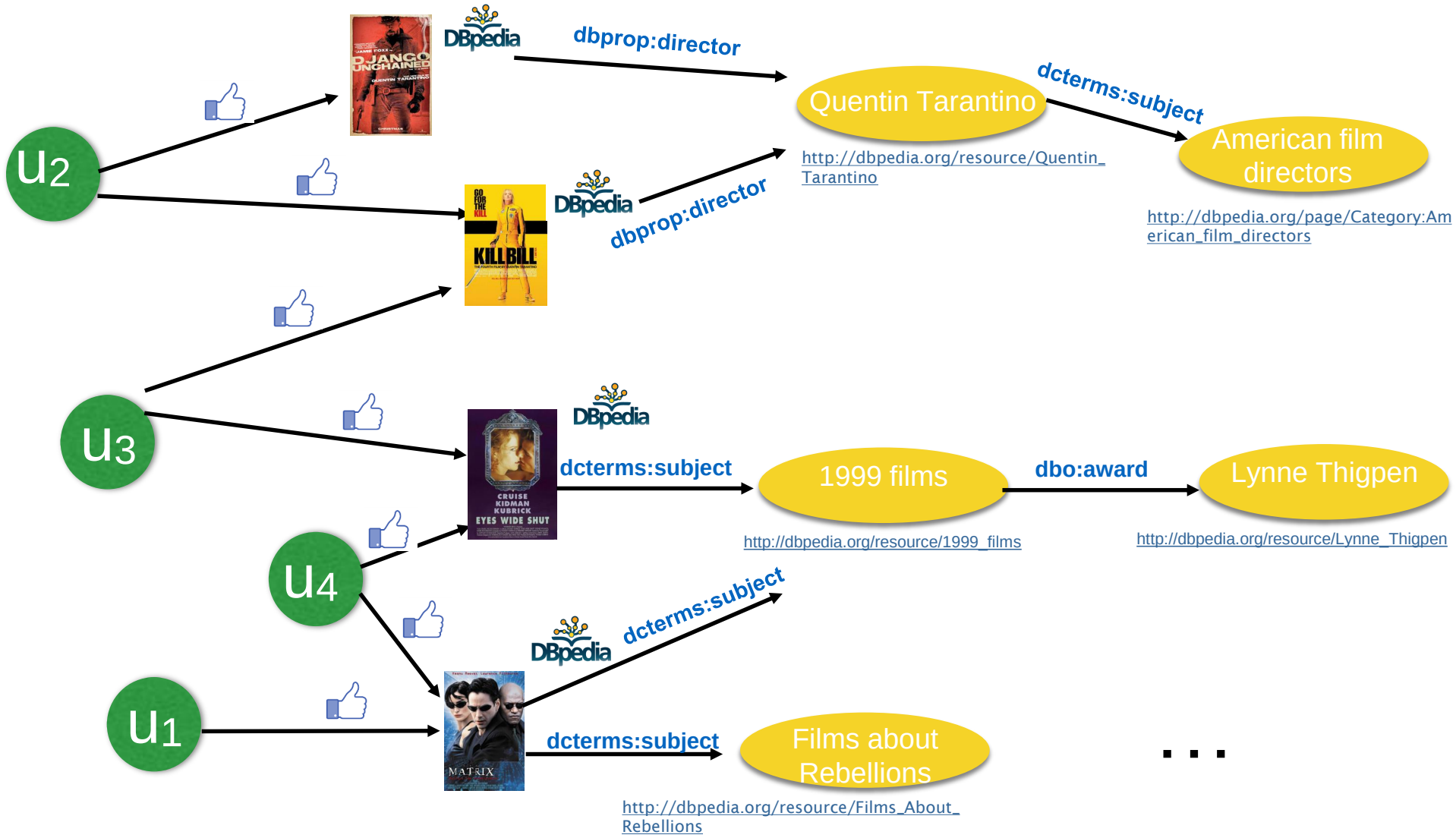
Semantic Graph-based Data Model

(1-hop)



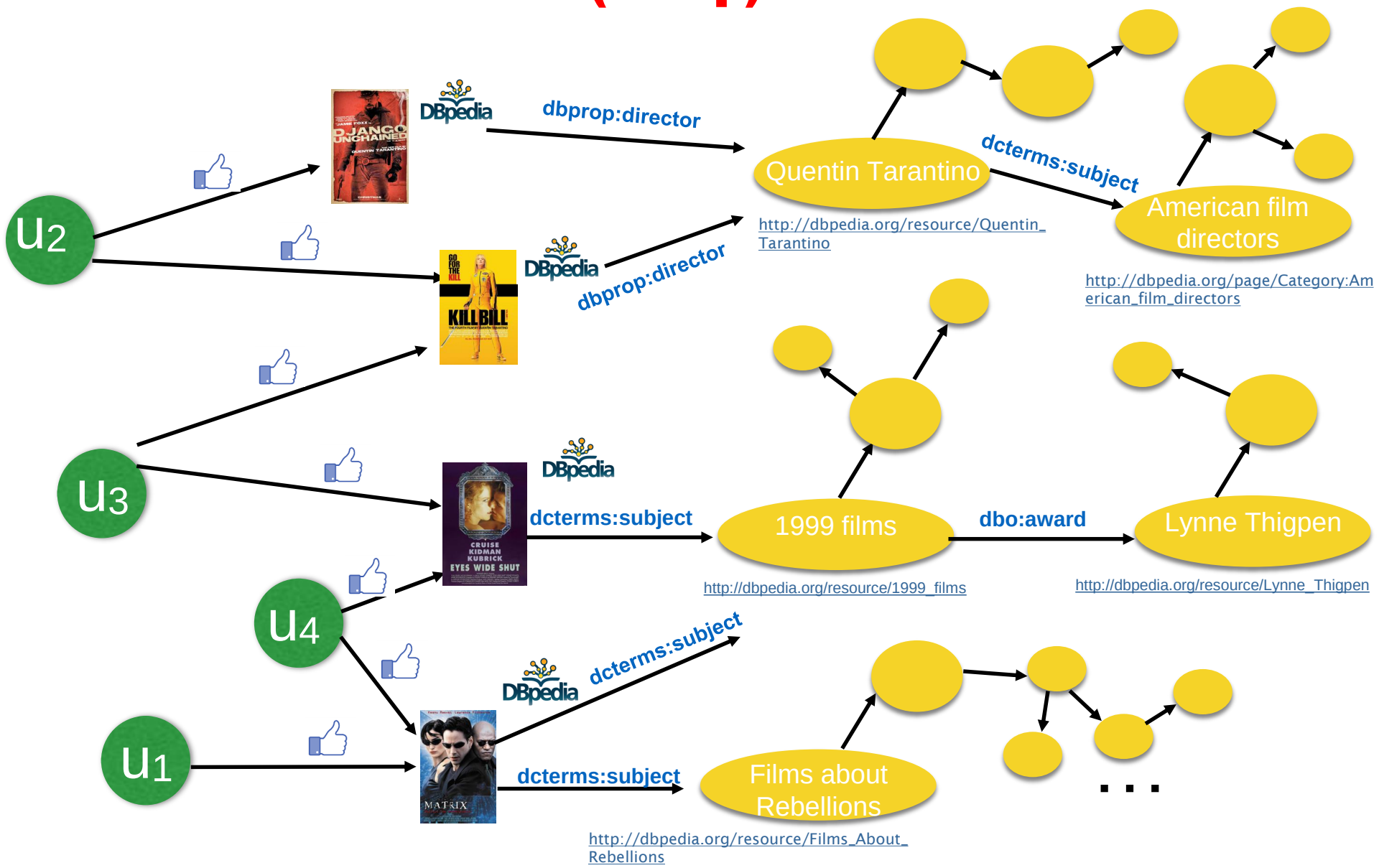
Semantic Graph-based Data Model

(2-hop)

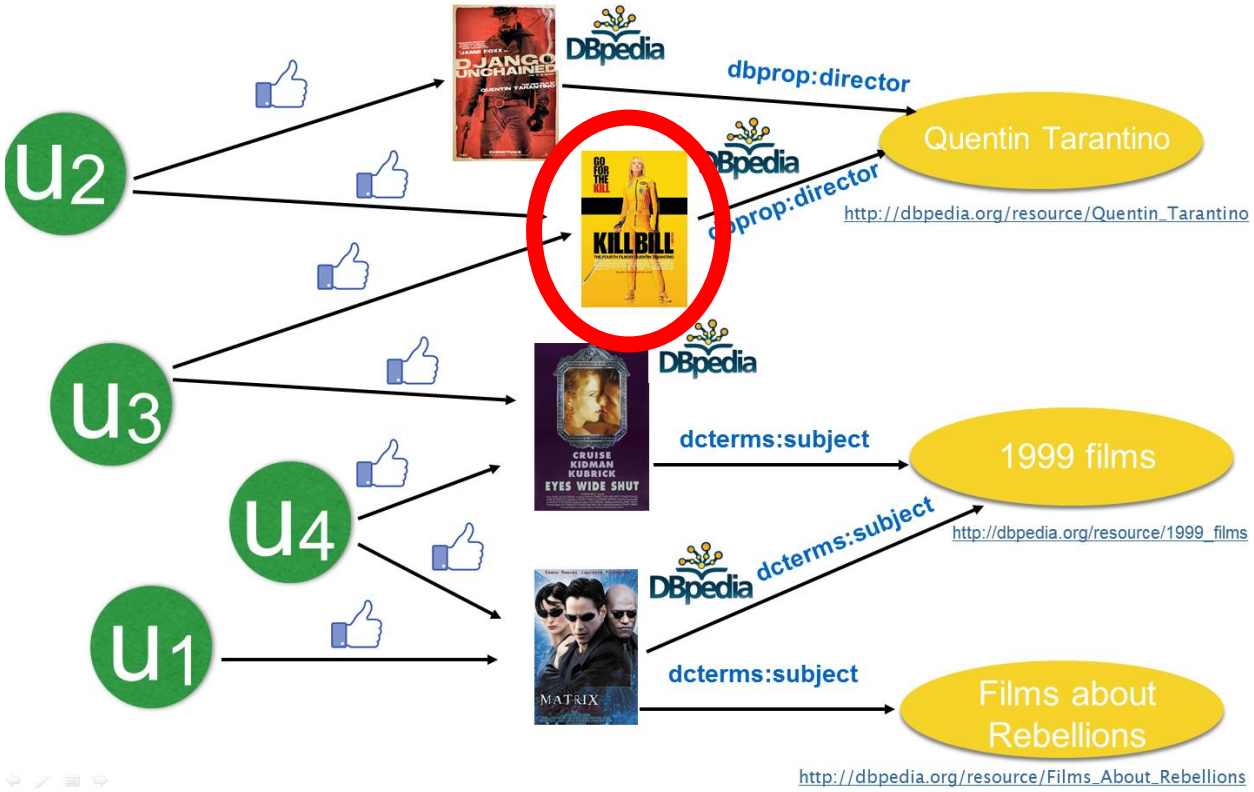


Semantic Graph-based Data Model

(n-hop)



Semantic Graph-based Data Model
(Feature Generation)



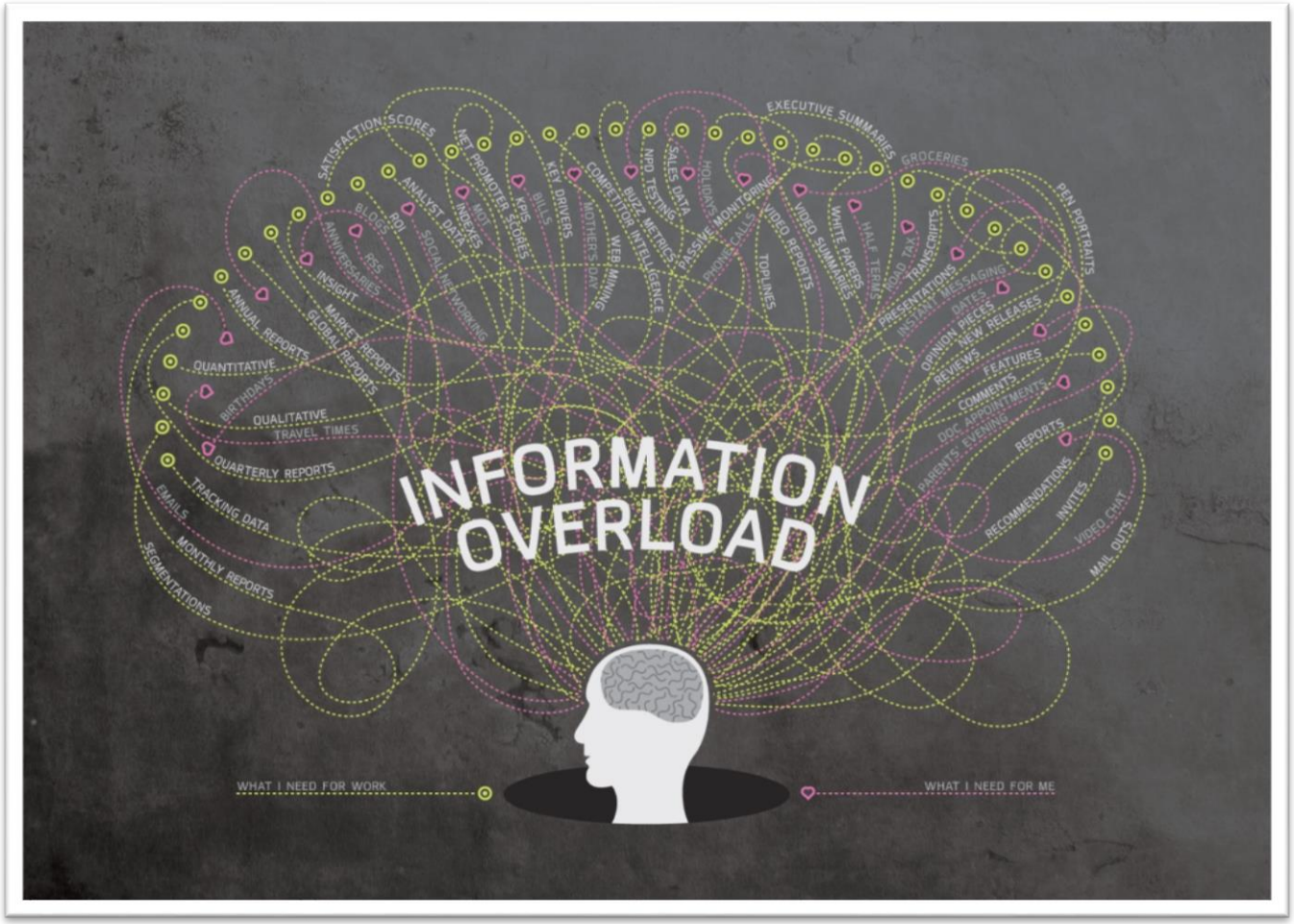
new semantic features describing the item can be **inferred** by **mining** the **structure** of the **tripartite graph**

PageRank
Spreading activation
Average Neighbors
Degree Centrality
…

How?

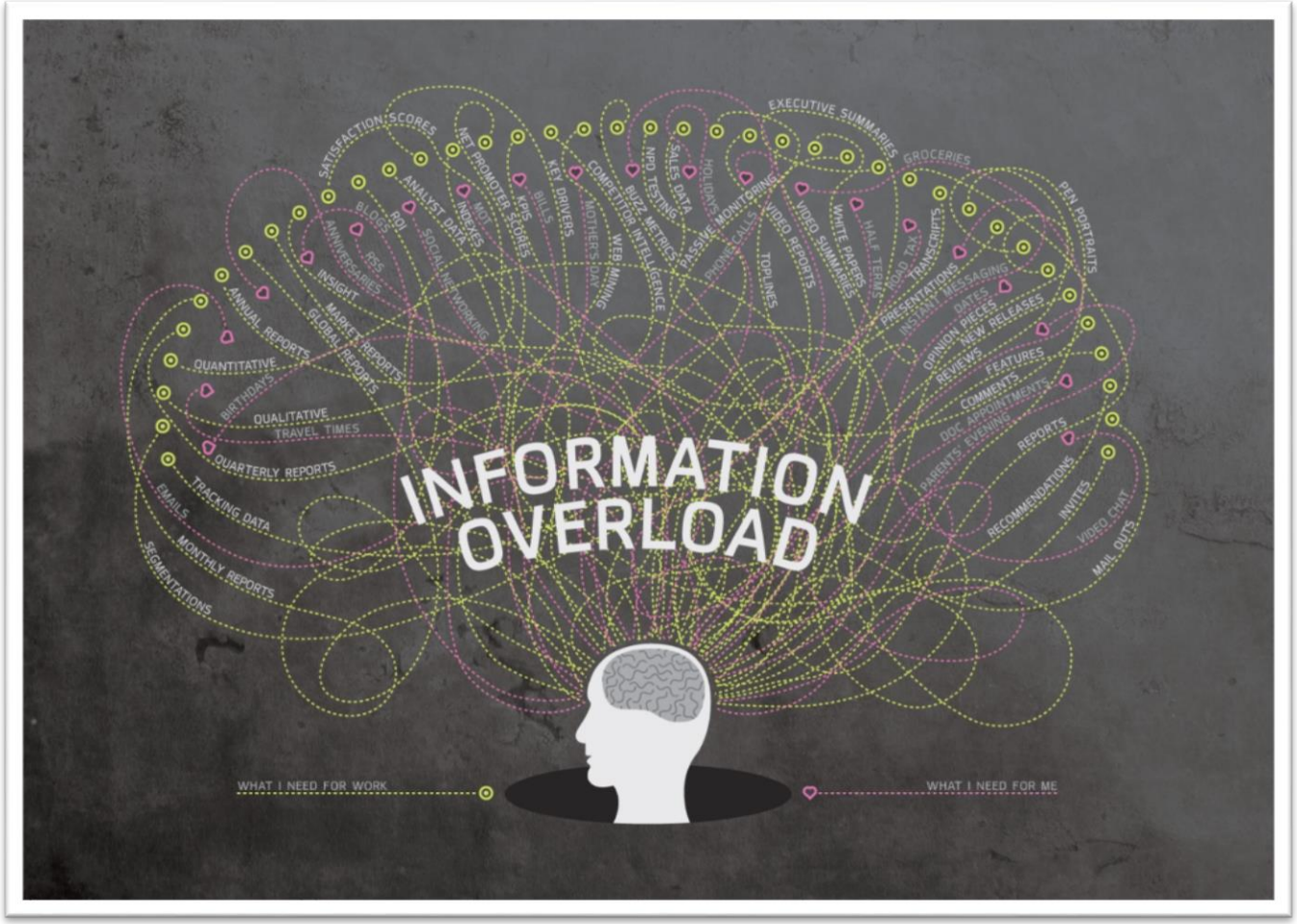
Encoding **endogenous semantics**
(bottom-up approaches)

Insight



Very huge availability of textual content

Insight



We can use this huge amount of content to directly learn a representation of words

Insight

Pass me a **Peroni!**
I like **Peroni**
Football and **Peroni**, what a perfect Saturday!

What is «Peroni» ?

Insight

Pass me a **Budweiser!**
I like **Budweiser**
Football and **Budweiser**, what a perfect Saturday!

What is «Budweiser» ?

Insight

Pass me a **Budweiser!**
I like **Budweiser**
Football and **Budweiser**, what a perfect Saturday!

What is «Budweiser» ?

Insight

Pass me a **Peroni!**
I like **Peroni**
Football and **Peroni**, what a perfect Saturday!

What is «Peroni» ?

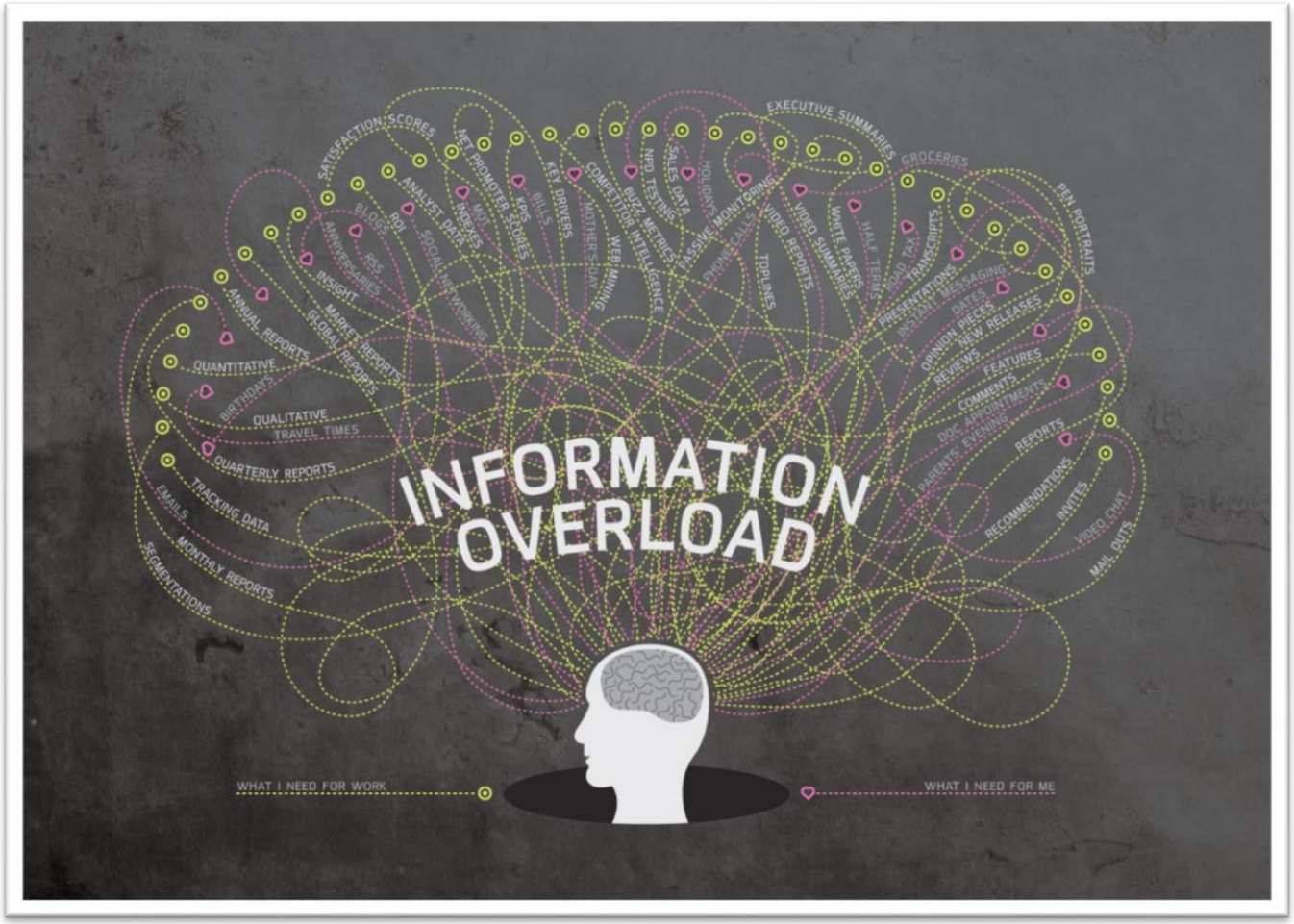
Insight

Pass me a **Peroni!**
I like **Peroni**
Football and **Peroni**, what a perfect Saturday!

What is «Peroni» ?

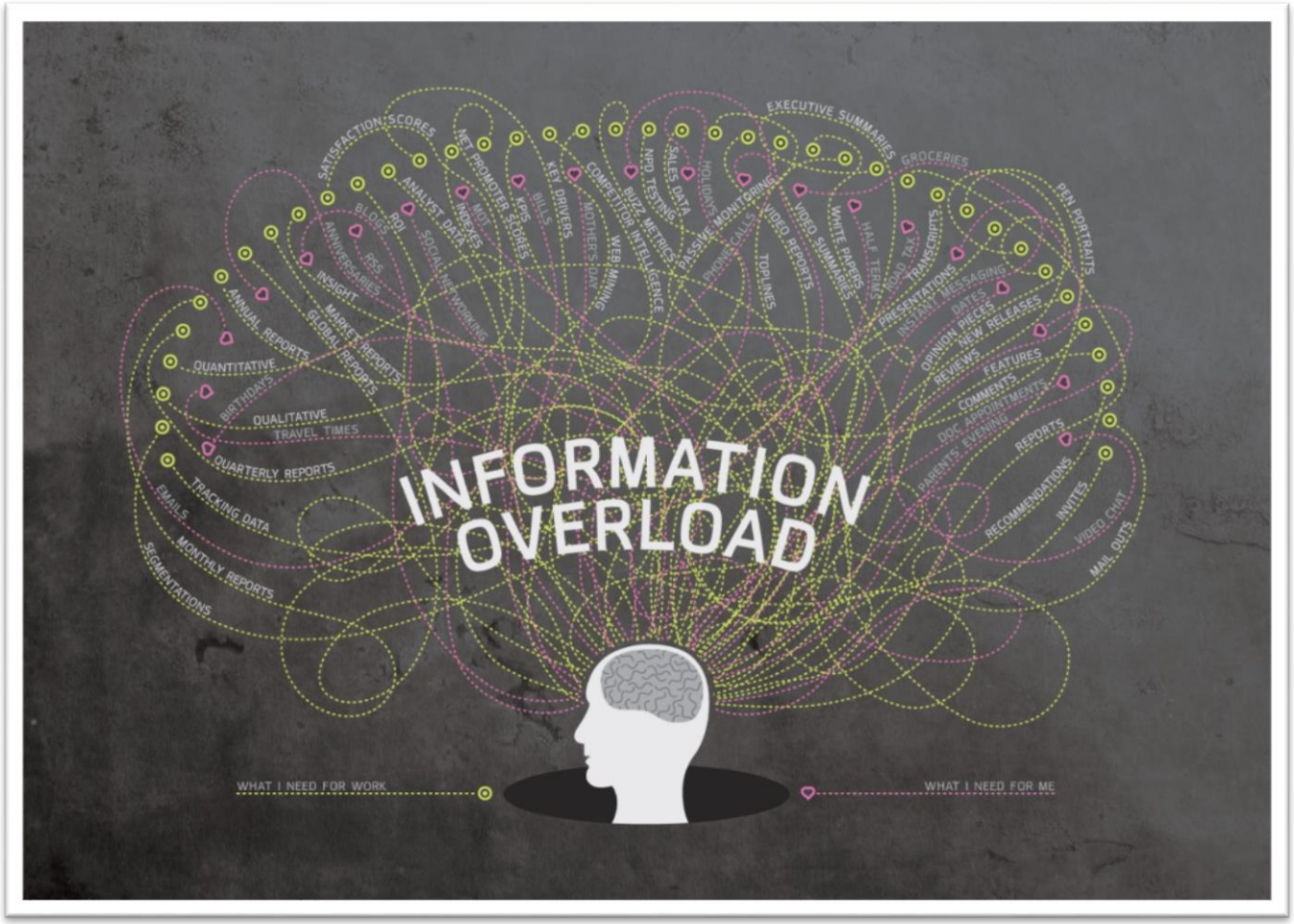
The most famous beer in Bari !

Insight



The semantics learnt according to terms usage is called «distributional»

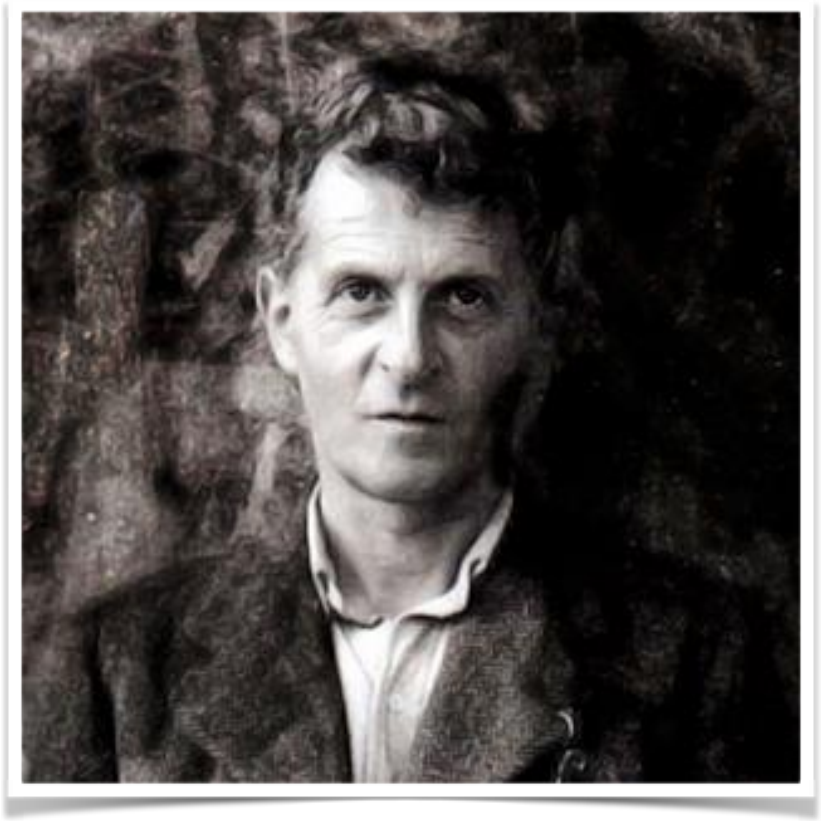
Insight



Distributional Hypothesis

«Terms used in similar contexts share a similar meaning»

Distributional Semantics



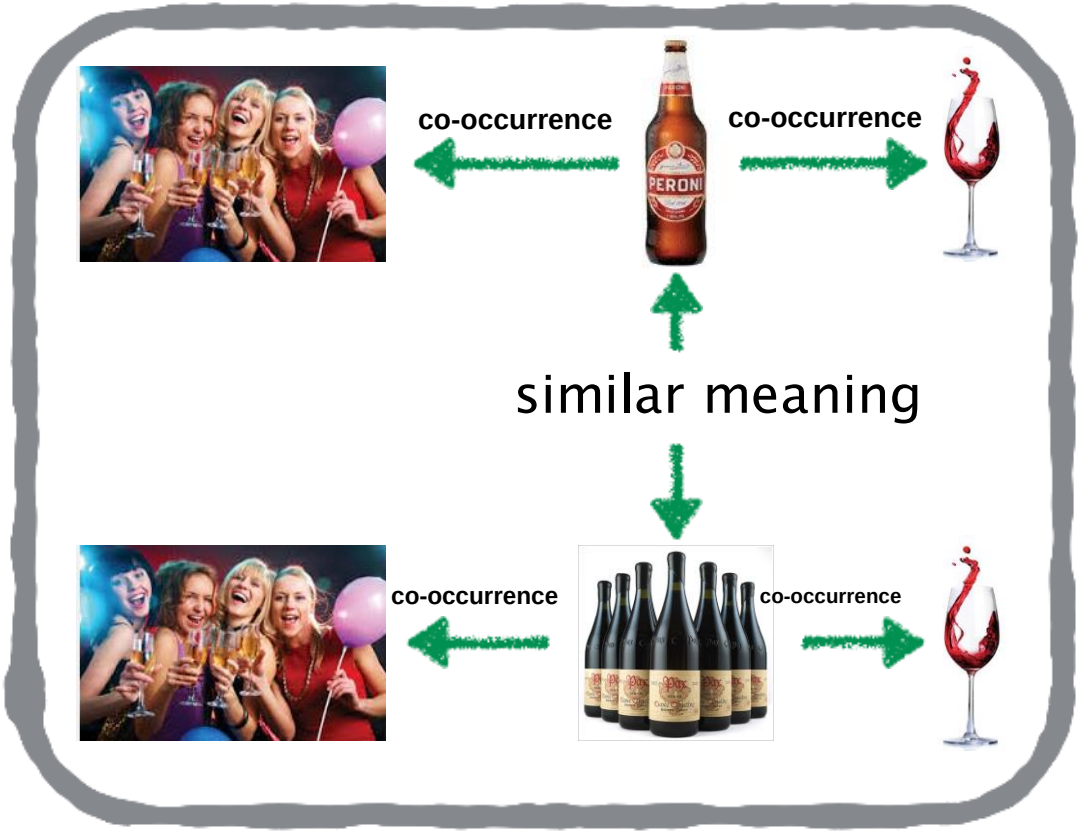
L.Wittgenstein
(Austrian philosopher)

“meaning is its use”

Distributional Semantics

Definition

by analyzing large corpora of textual data it is possible to infer information about the usage (about the meaning) of the terms



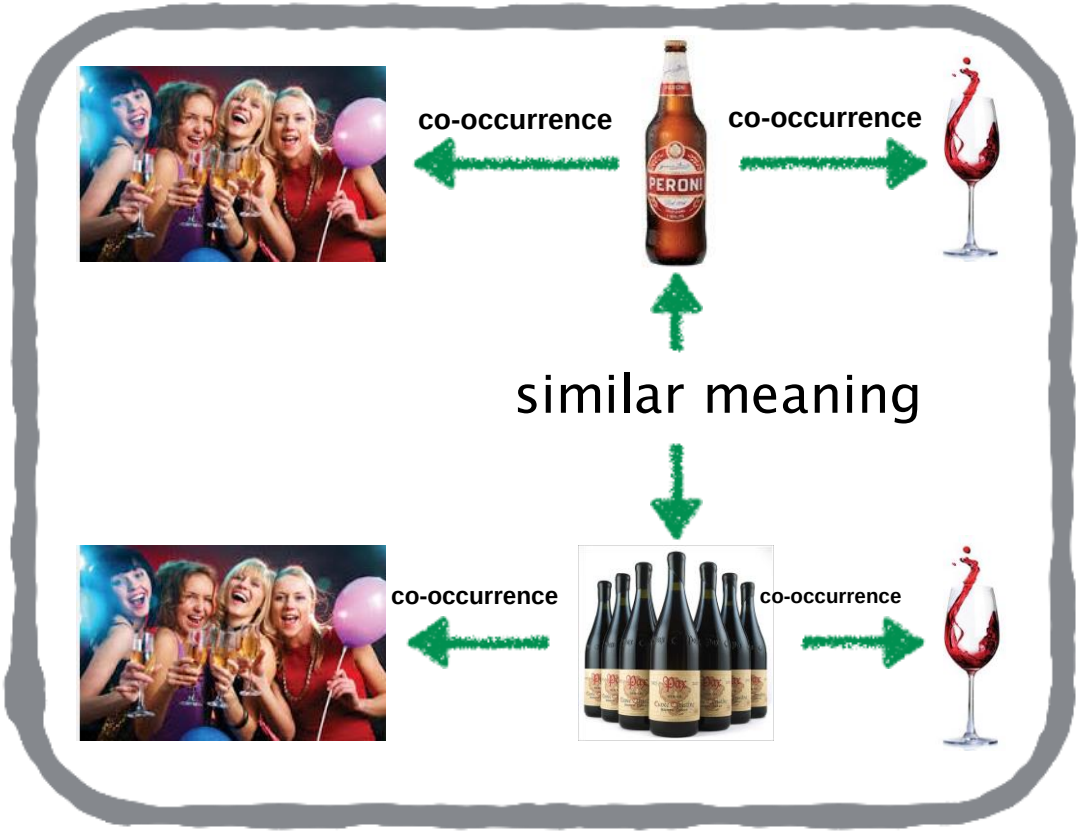
(*) Firth, J.R. A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32, 1957.

Distributional Semantics

Definition

by analyzing large corpora of textual data it is possible to infer information about the usage (about the meaning) of the terms

(*) Firth, J.R. A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp. 1-32, 1957.



Beer and wine share a similar meaning since they are often used in similar contexts

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

A **vector-space representation** is learnt
by encoding in which context each term is used

Each row of the matrix is a vector!

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

beer vs wine: good overlap

Similar!

Distributional Semantics

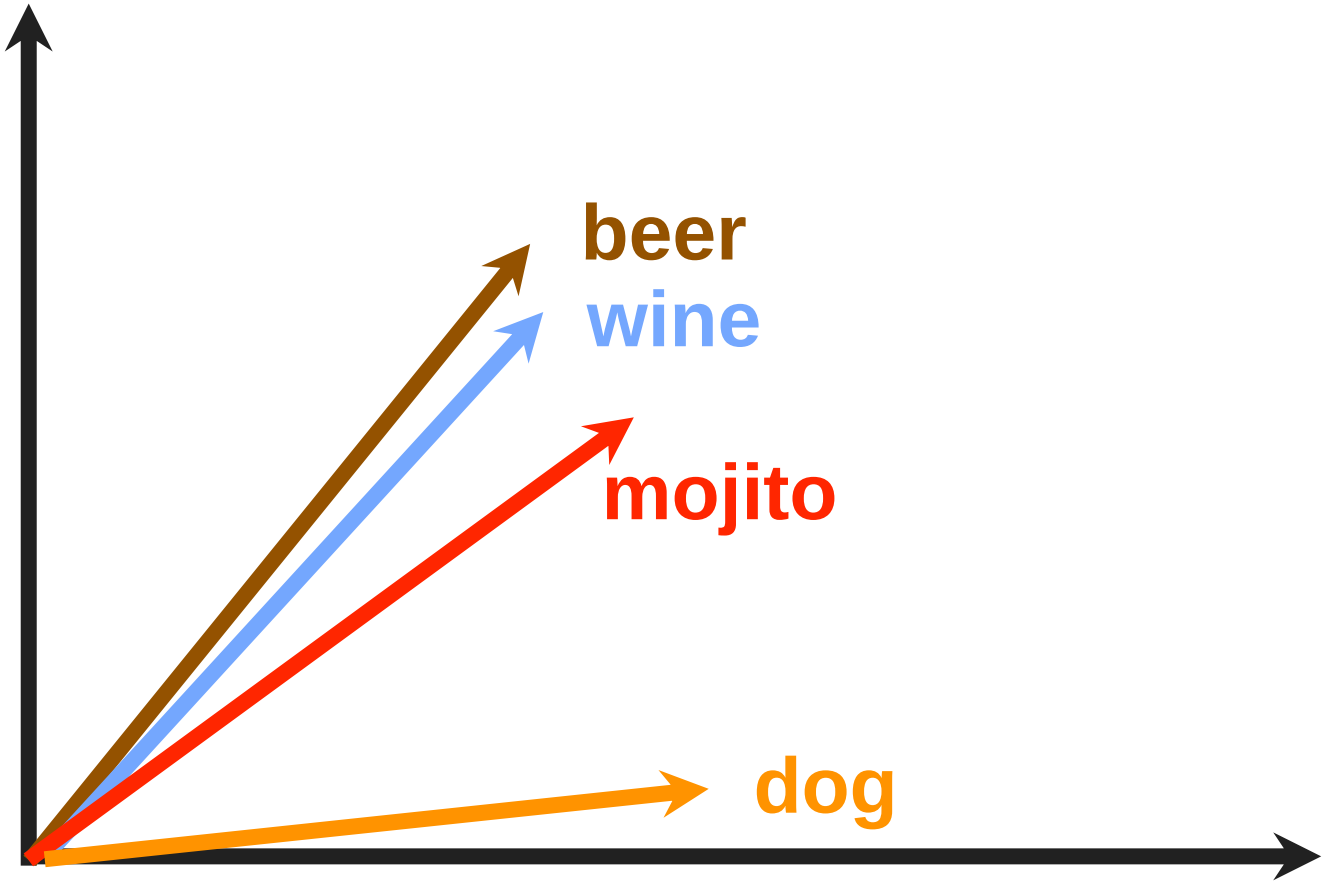
Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

beer vs spoon: no overlap

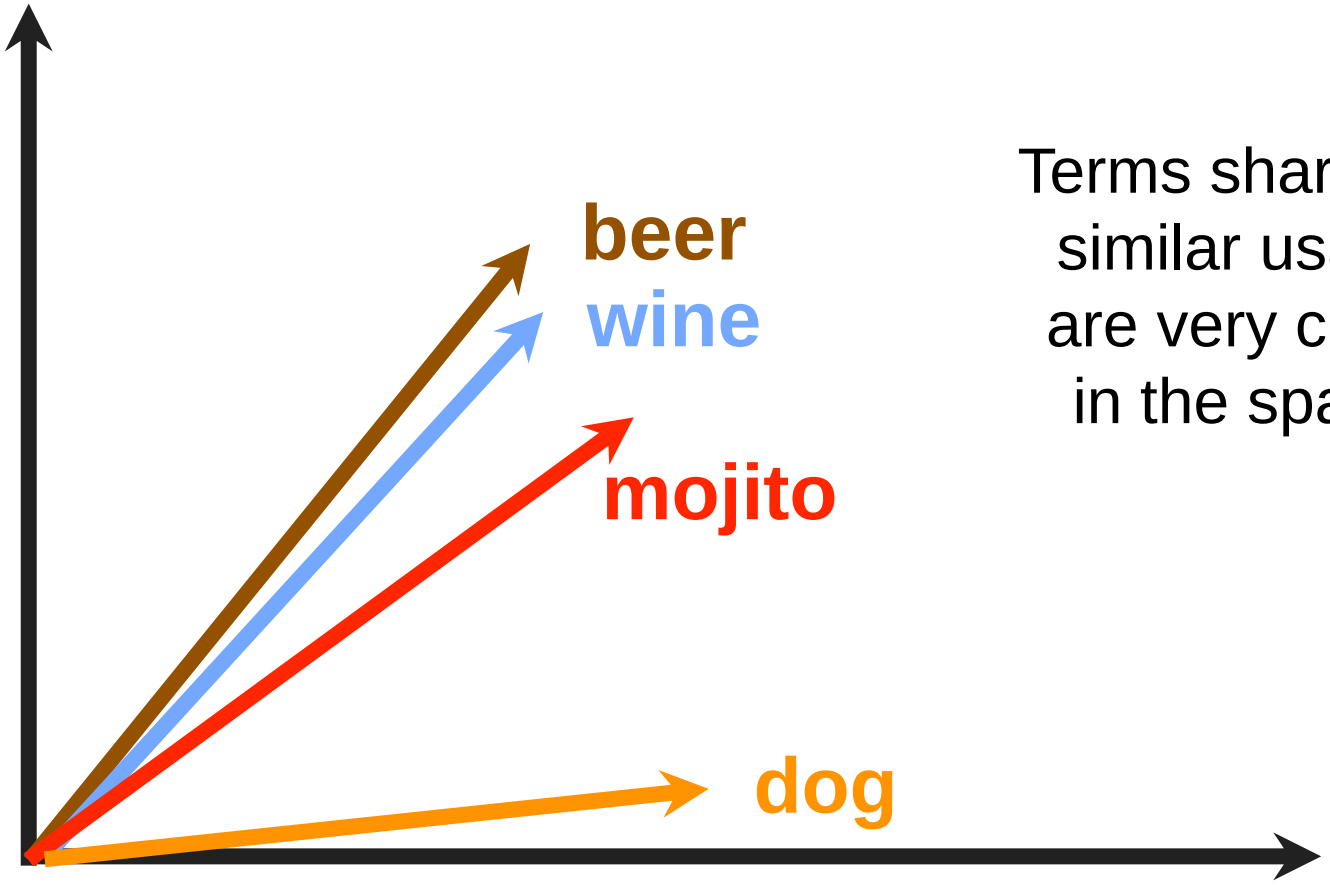
Not Similar!

WordSpace



A vector space representation (called WordSpace) is learnt according to terms usage in contexts

WordSpace



Terms sharing a similar usage are very close in the space

A vector space representation (called WordSpace) is learnt according to terms usage in contexts

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

Key question: **what is the context?**

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

Key question: **what is the context?**

These approaches are very flexible since the «context» can be set **according to the granularity required by the representation**

Distributional Semantics

Term-Contexts Matrix

	d1	d2	d3	d4	d5	d6	d7	d8	d9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

Key question: **what is the context?**

Coarse-grained granularity:

context=whole document

Distributional Semantics

Term-Contexts Matrix = Term-Document Matrix

	d1	d2	d3	d4	d5	d6	d7	d8	d9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

Key question: **what is the context?**

(This is Vector Space Model!)

Vector Space Model is a Distributional Model

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

Key question: **what is the context?**

Fine-grained granularities:

context=paragraph, sentence, window of words

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

Fine-grained granularities:

PROs: the more fine-grained the representation, **more precise the vectors**
CONs: the more fine-grained the representation, **the bigger the matrix**

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
beer		✓	✓			✓	✓		
wine		✓	✓			✓	✓	✓	
spoon	✓			✓				✓	✓
glass	✓	✓	✓		✓				✓

The flexibility of distributional semantics models
also regards the rows of the matrix

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✓	✓			✓	✓		
concept2		✓	✓			✓	✓	✓	
concept3	✓			✓				✓	✓
concept4	✓	✓	✓		✓				✓

The flexibility of distributional semantics models
also regards the rows of the matrix

Keywords can be replaced with concepts
(as synsets or entities!)

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✓		✓		✓	✓		✓
Al Pacino			✓			✓			
American Writers	✓			✓				✓	✓
Laurence Fishburne	✓		✓		✓				✓

The flexibility of distributional semantics models
also regards the rows of the matrix

Keywords can be replaced with concepts
(as synsets or entities!)

Distributional Semantics

Term-Contexts Matrix

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✓		✓		✓	✓		✓
Al Pacino			✓			✓			
American Writers	✓			✓					
Laurence Fishburne	✓		✓		✓				

Drama film

From Wikipedia, the free encyclopedia

This article **possibly contains original research**. Please improve it citations. Statements consisting only of original research should be r

A **drama film** is a film genre that depends mostly on in-depth development of realistic characters alcoholism, drug addiction, infidelity, moral dilemmas, racist prejudice, religious intolerance, sexua corruption put the characters in conflict with themselves, others, society and even natural phenom subgenres such as romantic drama, sport films, period drama, courtroom drama and crime.[1]

At the center of a drama is usually a character or characters who are in conflict at a crucial mome *Ordinary People* dig under the skin of everyday life to ask big questions and touch on the deepest tragic or at least painful resolutions and concern the survival of some tragic crisis, like the death o *Kramer*). Some of the greatest screen performances come from dramas, as there is ample opport afford.[2]

Drama films have been nominated frequently for the Academy Award (particularly Best Picture) -

Contents [hide]

1 Sub-genres
2 Early film-1950s
3 1960s-1970s
4 1980s-1990s

Keanu Reeves and Al Pacino are «connected» because they both acted in **Drama Films**

Distributional Semantics

Representing Documents

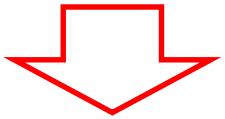
	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✓		✓		✓	✓		✓
Al Pacino			✓			✓			
American Writers	✓			✓				✓	✓
Laurence Fishburne	✓		✓		✓				✓

Given a WordSpace, a vector space representation of documents (called DocSpace) is typically built as the centroid vector of word representations

Distributional Semantics

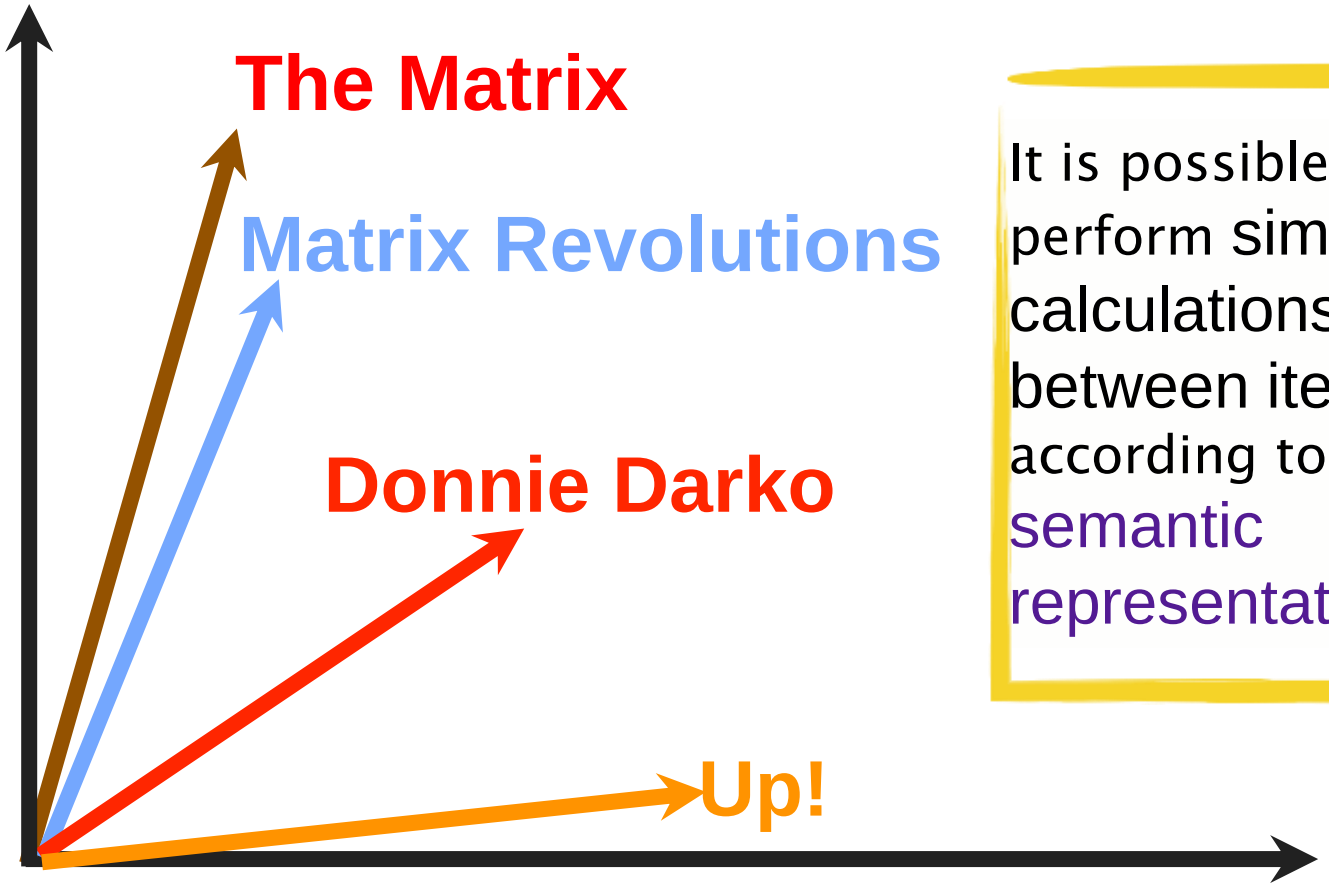
Representing Documents

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Keanu Reeves		✓		✓		✓	✓		✓
Al Pacino			✓			✓			
American Writers	✓			✓				✓	✓
Laurence Fishburne	✓		✓		✓				✓



The Matrix	✓	✓	✓	✓	✓	✓		✓	✓

DocSpace



It is possible to perform similarity calculations between items according to their semantic representation

Given a WordSpace, a vector space representation of documents (called DocSpace) is typically built as the centroid vector of word representations

Distributional Semantics

- We can exploit the (big) corpora of data to **directly learn a semantic vector-space representation of the terms** of a language
- **Lightweight semantics,** not formally defined
- **High flexibility:** everything is a vector: term/term similarity, doc/term, term/doc, etc..
- **Context can have different granularities**
- **Huge amount of content** is needed
- **Matrices are particularly huge** and difficult to build
  - **Too many features: need for dimensionality reduction**

Semantic representations

- Explicit (Exogenous) Semantics
  - Introduce semantics **by mapping the features describing** the item with semantic **concepts**
  - Introduce semantics **by linking the Item to a knowledge graph**
- Implicit (Endogenous) Semantics
  - **Distributional semantic** models
    - Explicit Semantic Analysis
    - Random Indexing
    - Word2Vec
    - ……

Distributional Semantics models share the same insight **but have important distinguishing aspects**

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Distributional Semantics models share the same insight **but have important distinguishing aspects**

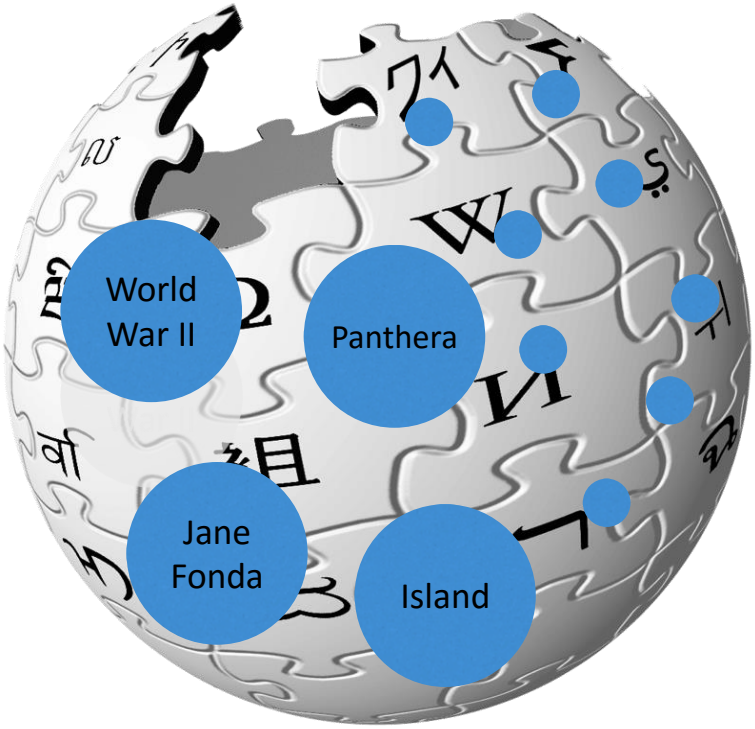
Explicit Semantic Analysis

Random Indexing

Word2Vec

……

Explicit Semantic Analysis (ESA)



ESA builds a vector-space **semantic representation** of natural language texts in a high-dimensional space of **comprehensible concepts** derived from Wikipedia [Gabri06]

[Gabri06] E. Gabrilovich and S. Markovitch. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Proceedings of the 21th National Conf. on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, pages 1301–1306. AAAI Press, 2006.

Explicit Semantic Analysis (ESA)

ESA matrix

ESA is a Distributional Semantic model which uses **Wikipedia articles as context**

Wikipedia articles

ESA	Concept 1	...	Concept n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

Terms

Explicit Semantic Analysis (ESA)

ESA matrix

semantic relatedness
between a word and a concept
TF-IDF score

Wikipedia articles

ESA	Concept 1	...	Concept n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

Terms

Explicit Semantic Analysis (ESA)

ESA matrix

semantic relatedness

between a word and a concept

TF-IDF score

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

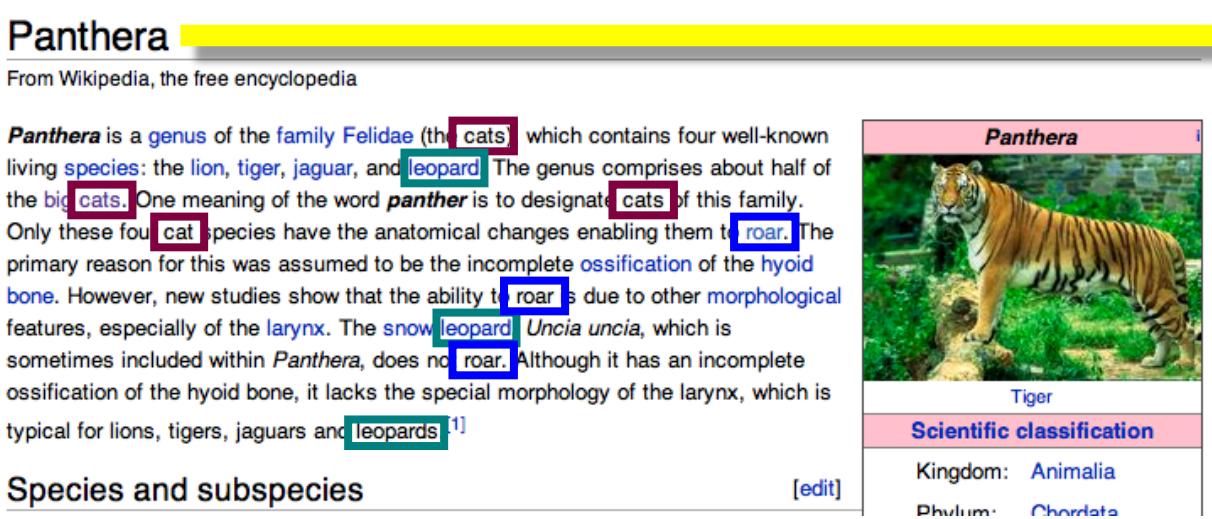
Wikipedia articles

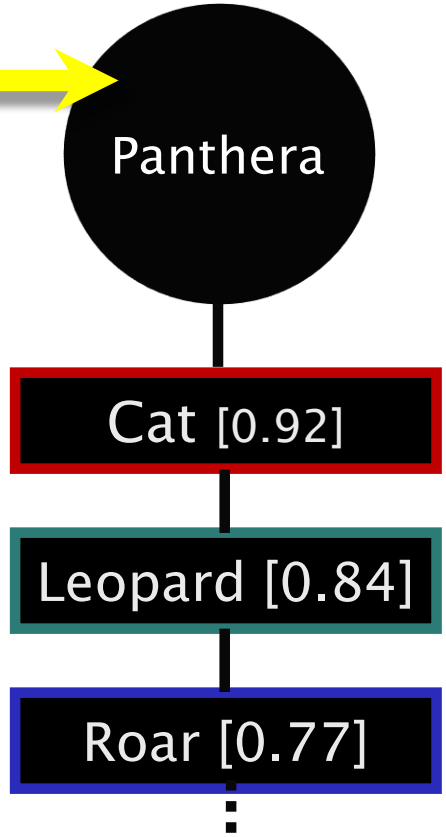
Terms

ESA	Concept 1	...	Concept n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

Explicit Semantic Analysis (ESA)

Every Wikipedia article represents a **context**





Article words are associated with the **concept** (TF-IDF)

Each Wikipedia page can be described in terms of the words with the highest TF/IDF score

(this is a column of ESA matrix)

Explicit Semantic Analysis (ESA)

ESA matrix

ESA	Concept 1	...	Concept n
term 1	TF-IDF	TF-IDF	TF-IDF
…	TF-IDF	TF-IDF	TF-IDF
term k	TF-IDF	TF-IDF	TF-IDF

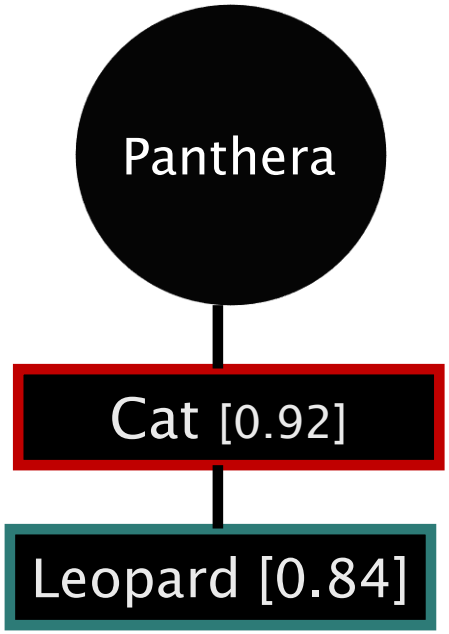
The vector-space representation of each term is called

semantic interpretation vector

Explicit Semantic Analysis (ESA)

Every Wikipedia article represents a **concept**
Article words are associated with the **concept** (TF-IDF)

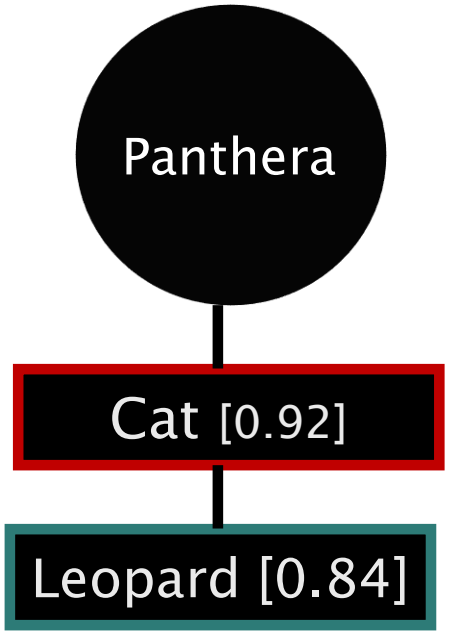
The **semantics** of a word is the **vector** of its **associations** with Wikipedia concepts

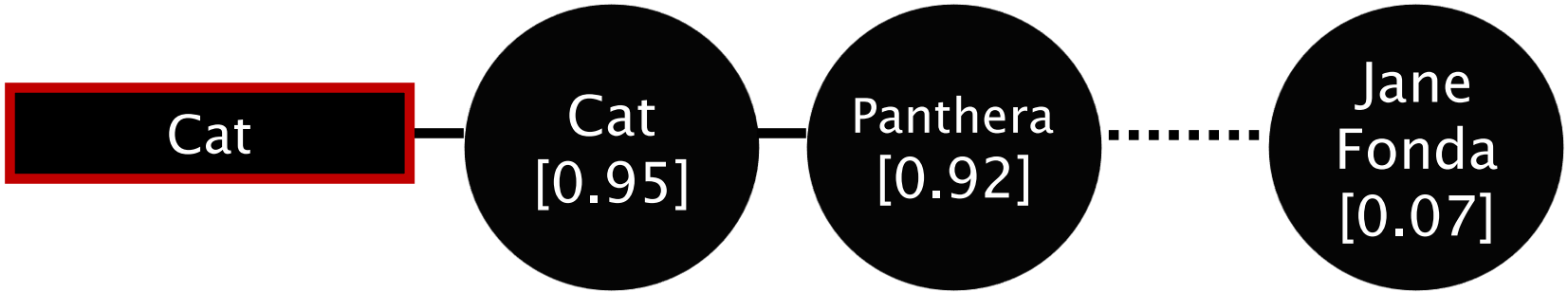


Explicit Semantic Analysis (ESA)

Every Wikipedia article represents a **concept**
Article words are associated with the **concept** (TF-IDF)

The **semantics** of a word is the **vector** of its **associations** with Wikipedia concepts





Explicit Semantic Analysis (ESA)

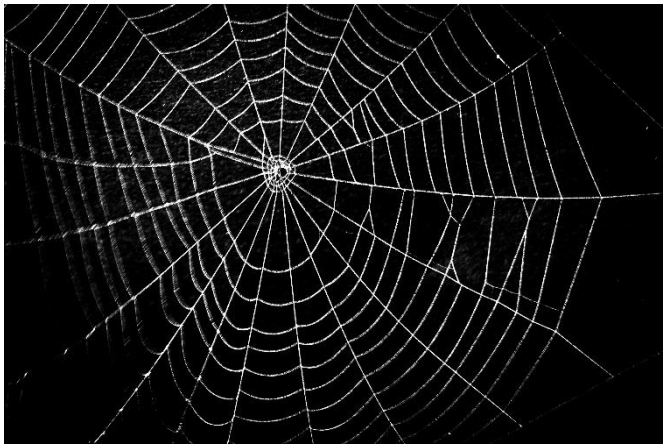
Important: the semantics of the words is not static.
It changes as Wikipedia articles are modified or
new articles are introduced.

ESA provides a representation which evolves over time!

Explicit Semantic Analysis (ESA)

Important: the semantics of the words is not static.
It changes as Wikipedia articles are modified or
new articles are introduced.

ESA provides a representation which evolves over time!



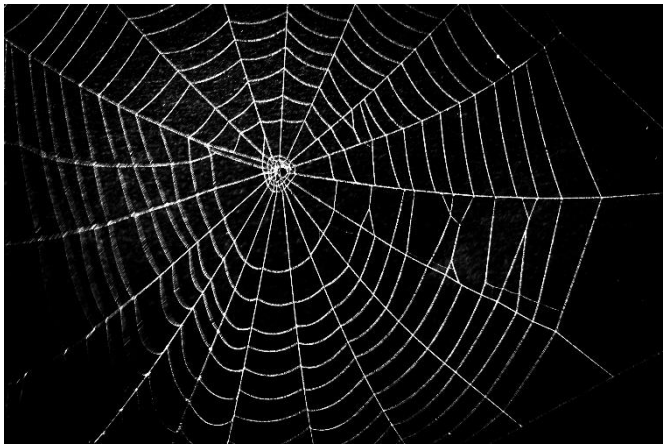
«web» in 1980

«web» in 2000

Explicit Semantic Analysis (ESA)

Important: the semantics of the words is not static.
It changes as Wikipedia articles are modified or
new articles are introduced.

ESA provides a representation which evolves over time!



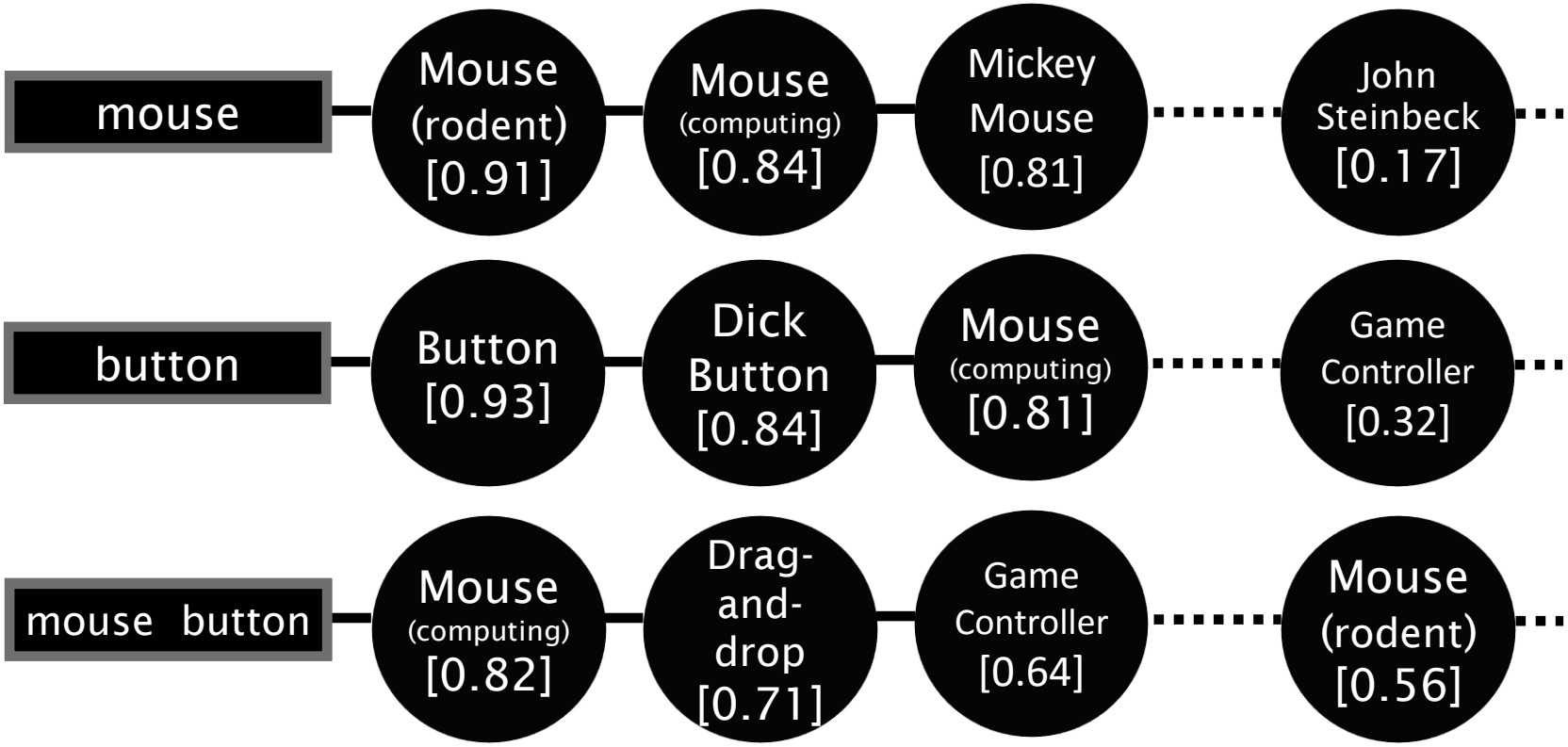
«web» in 1980

«web» in 2000

…what about text fragments?

Explicit Semantic Analysis (ESA)

The **semantics** of a **text fragment** is the **centroid** of the semantics of its **words**



Explicit Semantic Analysis (ESA)

The Matrix

From Wikipedia, the free encyclopedia

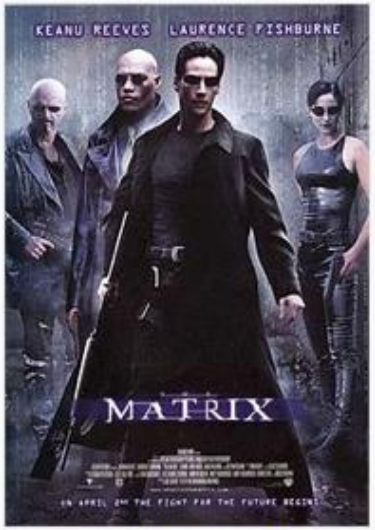
This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well received by critics [10][11] and won four Academy Awards as well as other accolades including BAFTA

The Matrix



Theatrical release poster

A semantic representation of an item can be built as the centroid vector of the **semantic interpretation vectors of the terms** in the plot.

Explicit Semantic Analysis (ESA)

The Matrix

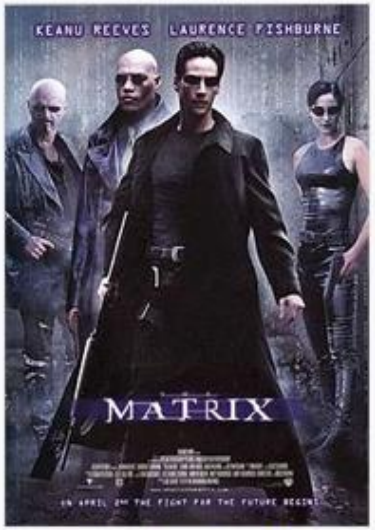
From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well received by critics [10][11] and won four Academy Awards as well as other accolades including BAFTA



Theatrical release poster

A semantic representation of an item can be built as the centroid vector of the **semantic interpretation vectors of the terms** in the plot.

Explicit Semantic Analysis (ESA)

The Matrix

From Wikipedia, the free encyclopedia

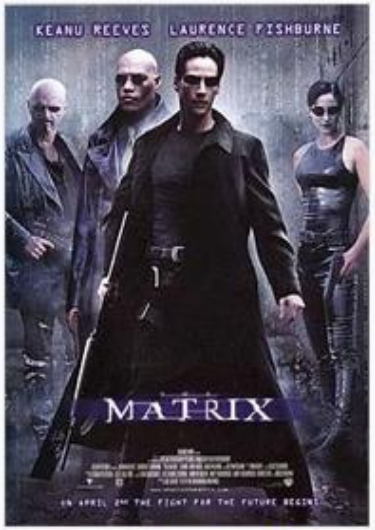
This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was generally well received by critics [10][11] and won four Academy Awards as well as other accolades including BAFTA

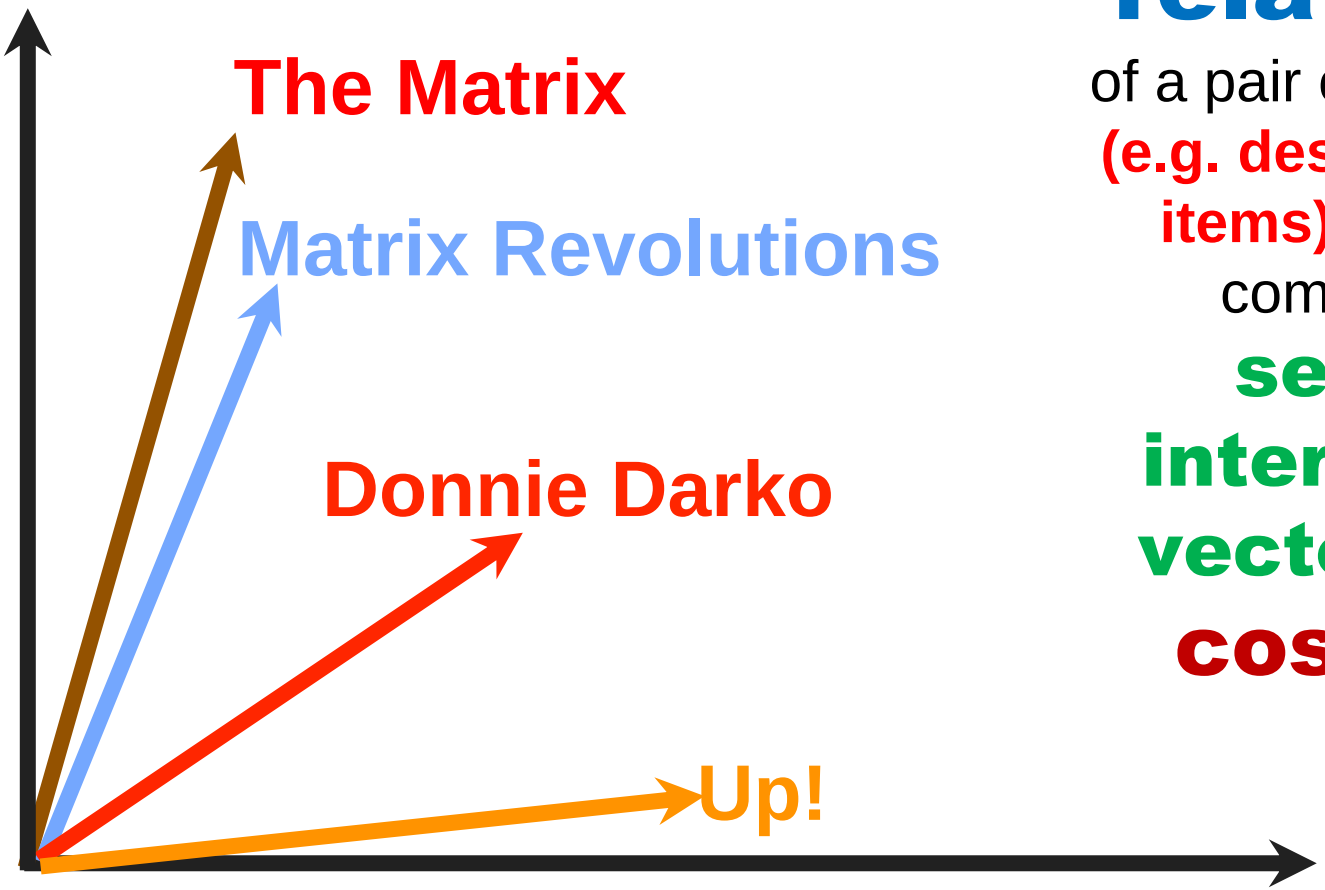
The Matrix



Theatrical release poster

Representation can be further improved and enriched by processing content through exogenous techniques (e.g. entity linking) in order to catch complex concepts

Explicit Semantic Analysis (ESA)



semantic relatedness of a pair of text fragments **(e.g. description of two items)** computed by comparing their **semantic interpretation vectors** using the **cosine** metric

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used to generate a set of relevant extra concepts describing the items.

How?

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used to generate a set of relevant extra concepts describing the items.

How?

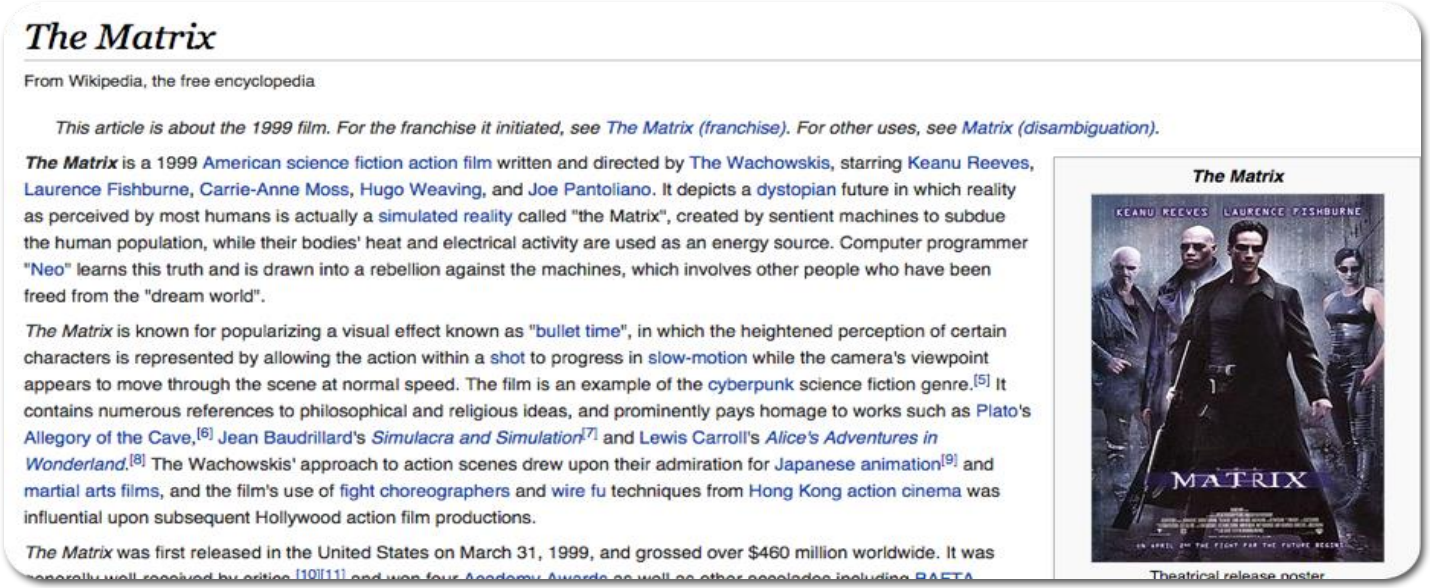
The Wikipedia pages with the highest TF/IDF score in the semantic interpretation vector of the item!

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used to generate a set of relevant extra concepts describing the items.

How?

The Wikipedia pages with the highest TF/IDF score in the semantic interpretation vector of the item!



The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's Allegory of the Cave,[6] Jean Baudrillard's *Simulacra and Simulation*[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over $460 million worldwide. It was

The Matrix

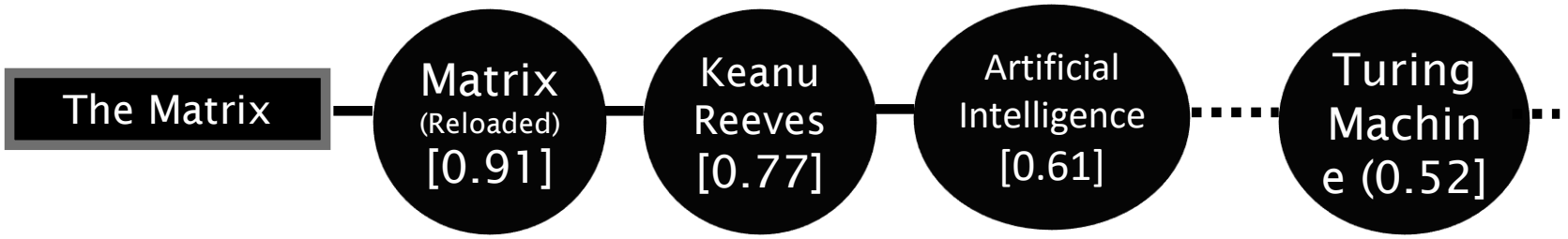
Theatrical release poster

Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used to generate a set of relevant extra concepts describing the items.

How?

The Wikipedia pages with the highest TF/IDF score in the semantic interpretation vector of the item!

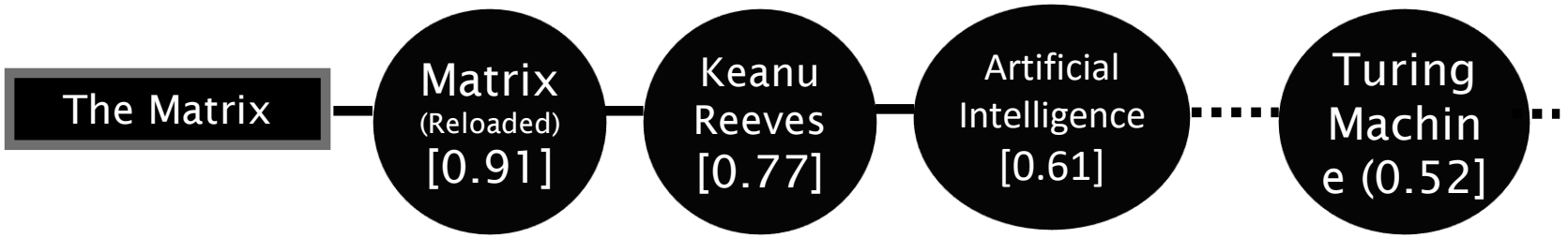


Explicit Semantic Analysis (ESA)

Another advantage: ESA can be also used to generate a set of relevant extra concepts describing the items.

How?

The Wikipedia pages with the highest TF/IDF score in the semantic interpretation vector of the item!



Extra features related to the item

Explicit Semantic Analysis (ESA)

ESA can be also exploited to build **a Multi-lingual item representation,** thanks to cross-lingual links in Wikipedia

Explicit Semantic Analysis (ESA)

ESA can be also exploited to build **a Multi-lingual item representation,** thanks to cross-lingual links in Wikipedia

For each text fragment, we can build a ESA representation in a language L1. Thanks to cross-lingual links we cap «map» the representation in different languages

Explicit Semantic Analysis (ESA)

Distributional Model which uses
Wikipedia Article as context

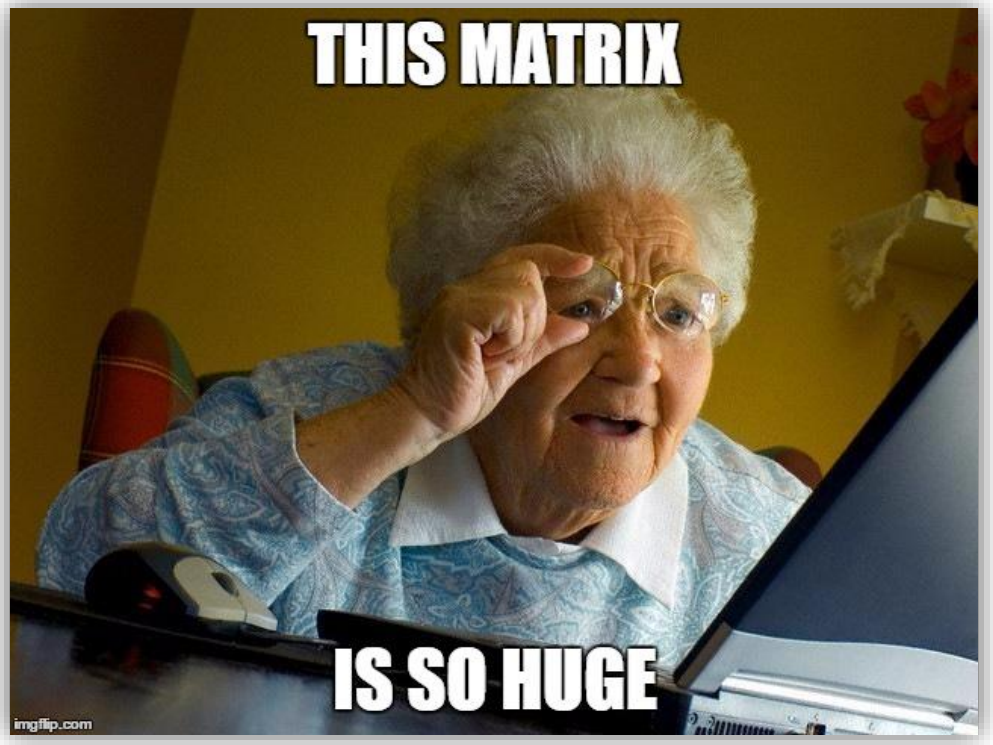
Very Transparent representation
(columns have an explicit meaning)

Representation can evolve over time!

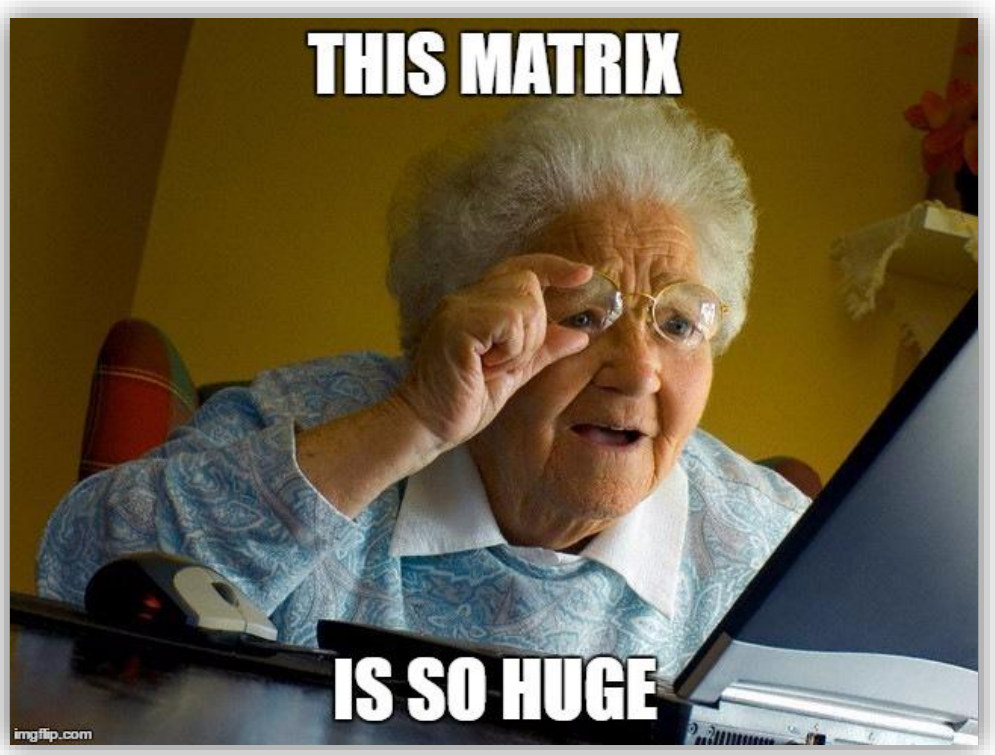
Also language-independent, thanks to cross-language Wikipedia links

The whole matrix is very huge

«Empirical» tuning of the parameters: how many articles? How many terms? What is the thresholding?

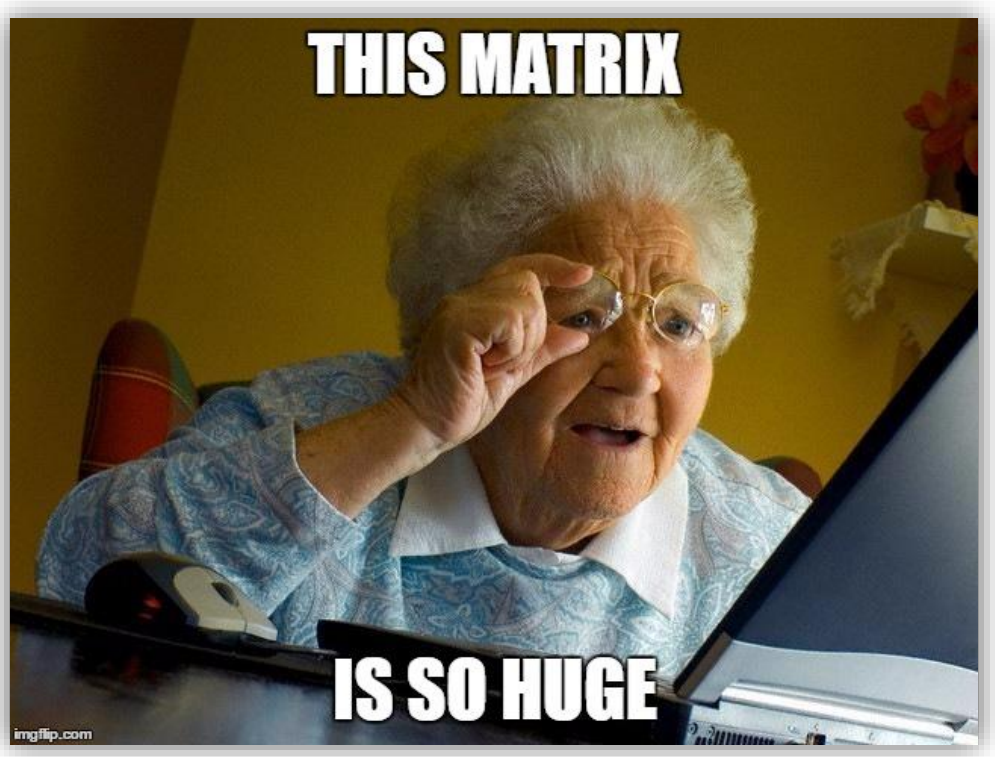


When transparency is not so important,
it is possible to learn a more compact vector-space representation of terms and items



When transparency is not so important,
it is possible to learn a more compact vector-space representation of terms and items

Dimensionality Reduction techniques

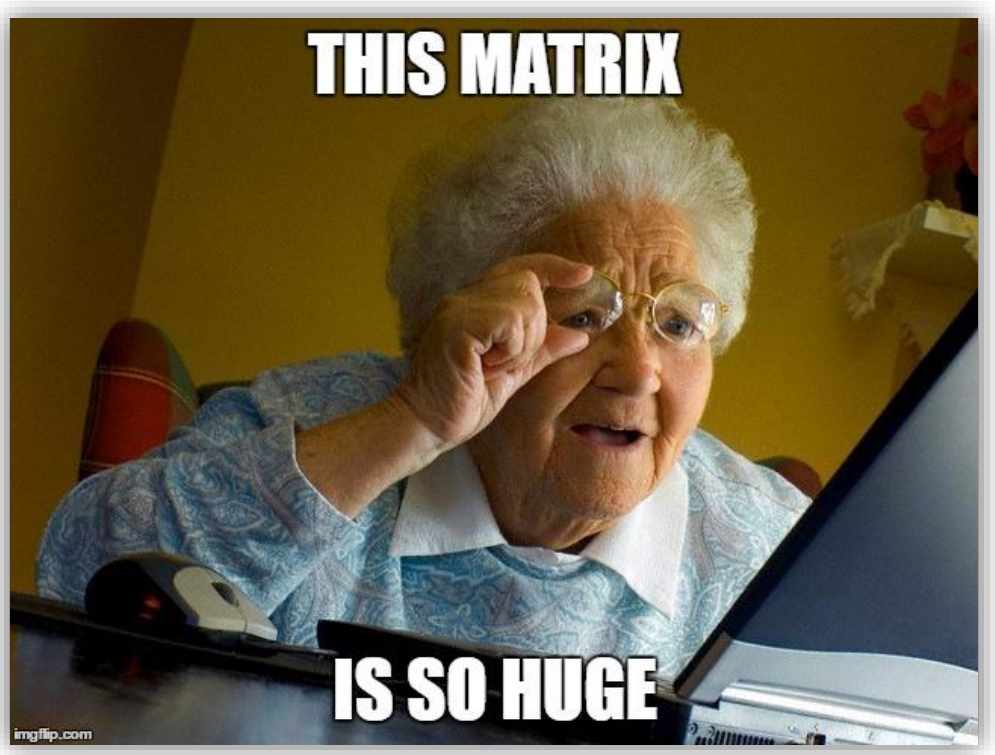


When transparency is not so important,
it is possible to learn a more compact vector-space representation of terms and items

a.k.a. Word embedding techniques

Embedding = a smaller representation of words

(more recent – equivalent - buzzword ☺)



When transparency is not so important,
it is possible to learn a more compact vector-space representation of terms and items

a.k.a. Word embedding techniques

Embedding = a smaller representation of words

Is this new?

Dimensionality reduction techniques

Latent Semantic Analysis (LSA) is a widespread distributional semantics model **which builds a term/context matrix** and calculates SVD over that matrix.

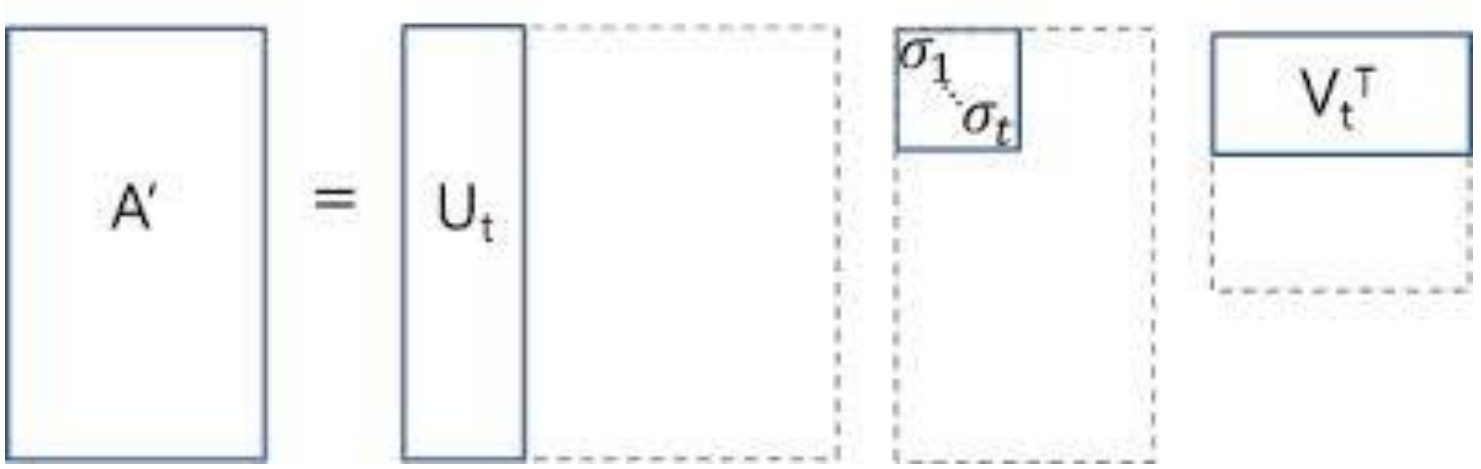
Dumais, Susan T. "Latent semantic analysis." *Annual review of information science and technology* 38.1 (2004): 188-230.

Dimensionality reduction techniques

Latent Semantic Analysis (LSA) is a widespread distributional semantics model **which builds a term/context matrix** and calculates SVD over that matrix.

Dumais, Susan T. "Latent semantic analysis." *Annual review of information science and technology* 38.1 (2004): 188-230.

Truncated Singular Value Decomposition



induces **higher-order (paradigmatic)** relations through the truncated SVD

Dimensionality reduction techniques

Singular Value Decomposition

PROBLEM

the **huge** co-occurrence matrix

SOLUTION

Use incremental and scalable techniques

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

Random Indexing

Word2Vec

……

Dimensionality reduction

Random Indexing

It is an incremental and scalable technique for dimensionality reduction.

M. Sahlgren. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD thesis, Stockholm University, 2006.

Dimensionality reduction

Random Indexing

It is an incremental and scalable technique for dimensionality reduction.

Insight

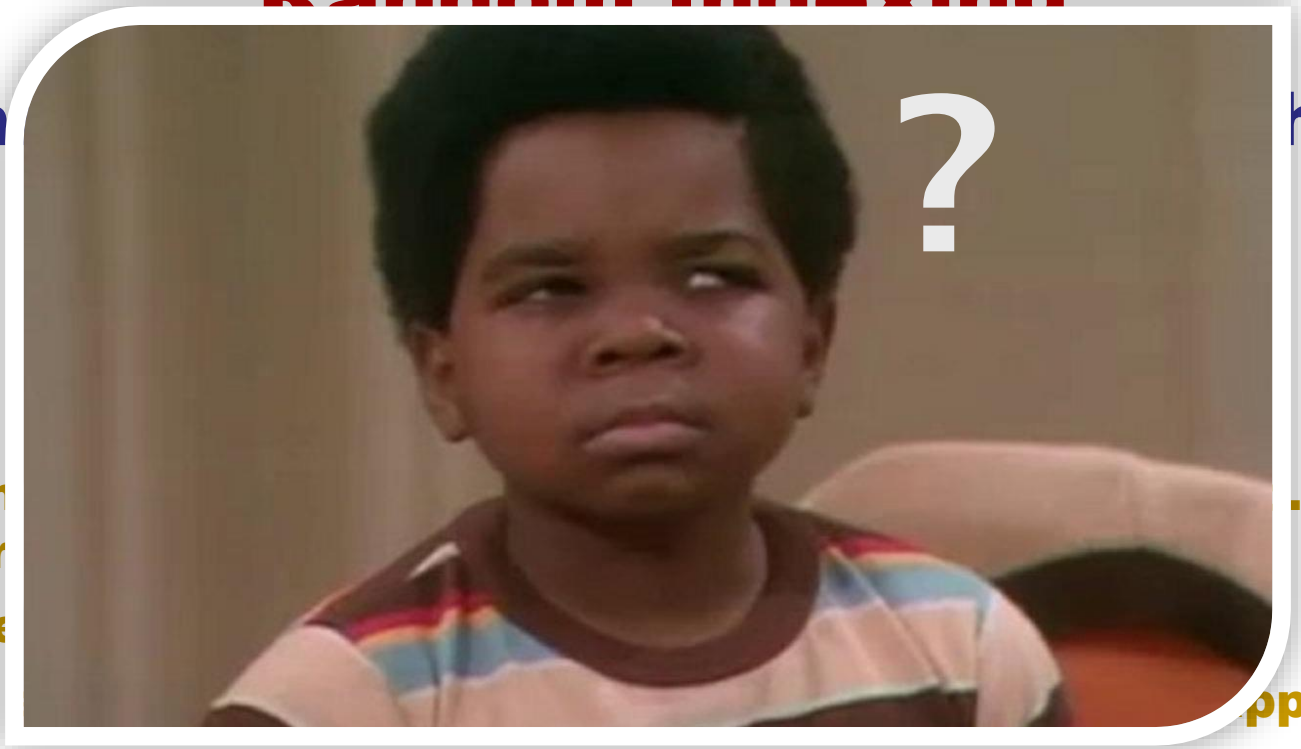
1. **Assign a vector to each context (word, documents, etc.).** The vector can be as big as you want.
2. Fill the vector **with (almost) randomly assigned values.**
3. Given a word, **collect the contexts where that word appears.**
4. **Update the representation according to term co-occurrences. The final representation is the «sum» of the contexts.**
5. **Obtain a (smaller but equivalent) vector space representation of the terms**

M. Sahlgren. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD thesis, Stockholm University, 2006.

Dimensionality reduction

Random Indexing

It is a … hnique



1. Assign … .). The vector …
2. Fill the …
3. Given … ppears.
4. Update the representation according to term co-occurrences. The final representation is the «sum» of the contexts.
5. Obtain a (smaller but equivalent) vector space representation of the terms

M. Sahlgren. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD thesis, Stockholm University, 2006.

Random Indexing

Algorithm

Step 1 - definition of the context granularity:

Document? Paragraph? Sentence? Word?

Step 2 – building the random matrix R

each **'context'** (e.g. sentence) is assigned a **context vector**

- ✓ dimension = **k**
- ✓ allowed values = **{-1, 0, +1}**
- ✓ small # of non-zero elements, i.e. **sparse vectors**
- ✓ values distributed in a **random way**

Random Indexing

Context vectors of dimension k = 8

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
r_n	…							

Each row is a «context»

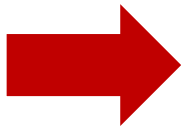
Random Indexing

Algorithm

Step 3 – building the reduced space

The **vector space representation** of a **term t** obtained by **combining** the **random vectors** of the **context in which it occurs in**

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
…								
r_n	…							



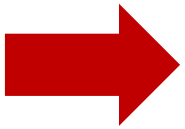
t1 ∈ {c1, c2, c5}

Random Indexing

Algorithm

Step 3 – building the reduced space B

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
...								
r_n	...							



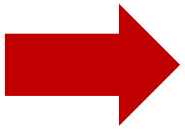
$t1 \in \{c1, c2, c5\}$

Random Indexing

Algorithm

Step 3 – building the reduced space B

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
...								
r_n	...							



t1 ∈ {c1, c2, c5}

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_5	1,	0,	0,	-1,	1,	0,	0,	0
t1	2,	0,	-1,	0,	1,	0,	0,	-1

Output: **WordSpace**

Random Indexing

Algorithm

Step 4 – building the document space

the **vector space representation** of a **document d** obtained by **combining** the **vector space representation** of the **terms that occur in the document**

Output: **DocSpace**

WordSpace and DocSpace

WordSpace

	c_1	c_2	c_3	c_4	...	c_k
t_1						
t_2						
t_3						
t_4						
...						
t_m						

DocSpace

	c_1	c_2	c_3	c_4	...	c_k
d_1						
d_2						
d_3						
d_4						
...						
d_n						

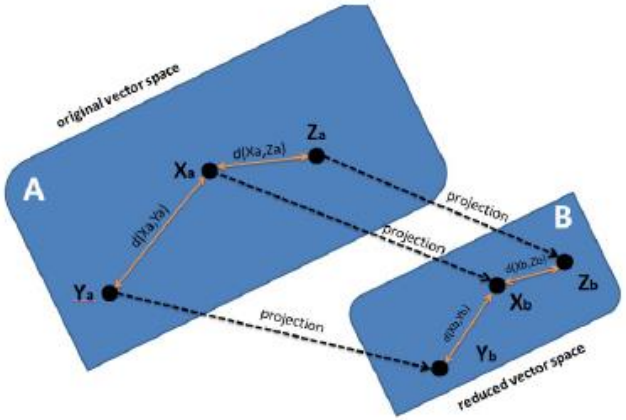
k is a simple parameter of the model

Uniform representation

Dimensionality reduction

..even if it sounds weird

theory: Johnson-Lindenstrauss' lemma [*]



$$B^{m,k} \approx A^{m,n}\, R^{n,k} \quad k << n$$

distances between the points in the reduced space
approximately preserved if

context vectors are nearly orthogonal

(and they are)

[*] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics, 26(189-206), 1.

Random Indexing

.... is also multilingual!

Why?

Random Indexing

.... is also multilingual!

Why?

Language-based geometric spaces are built on the ground of **the co-occurrences between terms**

Random Indexing

…. is also multilingual!

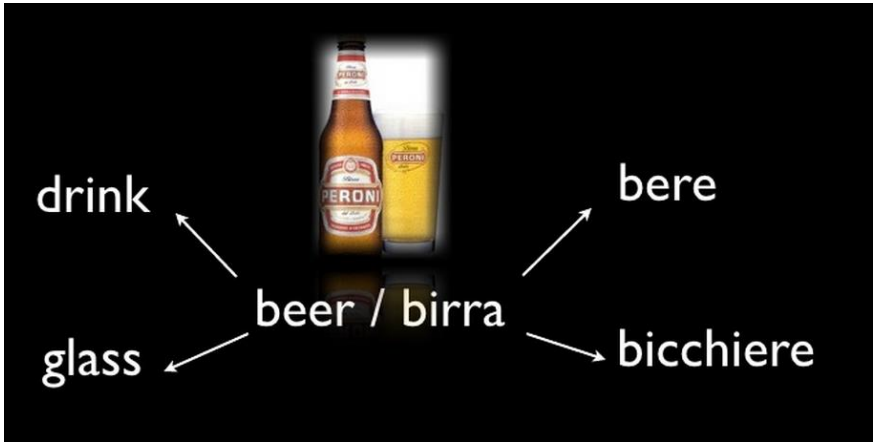
Why?

Language-based geometric spaces are built on the ground of **the co-occurrences between terms**

As the language changes, it only changes **the lexicalization of the co-occurrences**, their nature doesn't.

Random Indexing

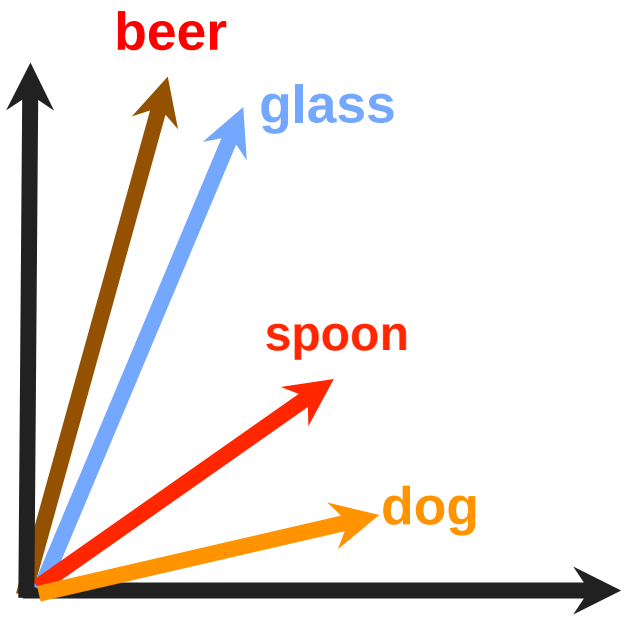
.... is also multilingual!



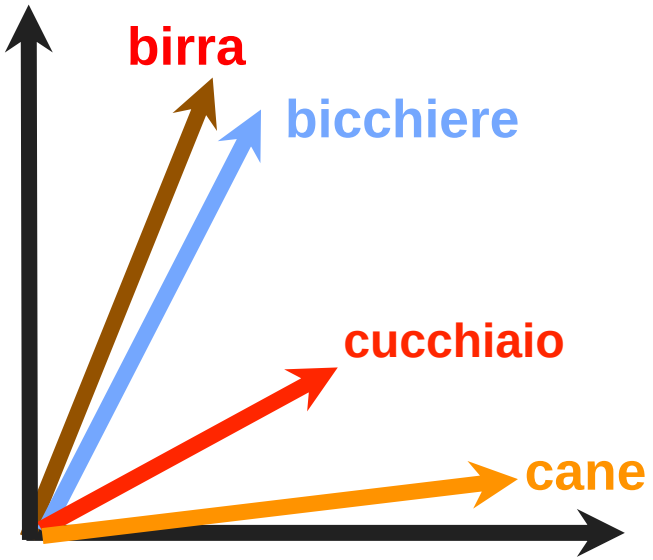
- ✓ the same concept, expressed in **different languages**, assumes the **same position** in language-based geometric spaces
- ✓ the position of **beer** in a geometric space based on **English** and the position of **birra** in a geometric space based on **Italian** are (almost) the same

Random Indexing

.... is also multilingual!



Italian WordSpace



English WordSpace

The position in the space can be slightly different, but the relations similarity between terms **still hold**

Random Indexing

Incremental and Scalable technique for learning word embeddings

Smaller vector space representation

Dimension of the space can be arbitrarly set

Incremental and Scalable

Multilingual!

Not transparent anymore

Proper tuning to find the optimal size of the embeddings

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

Random Indexing

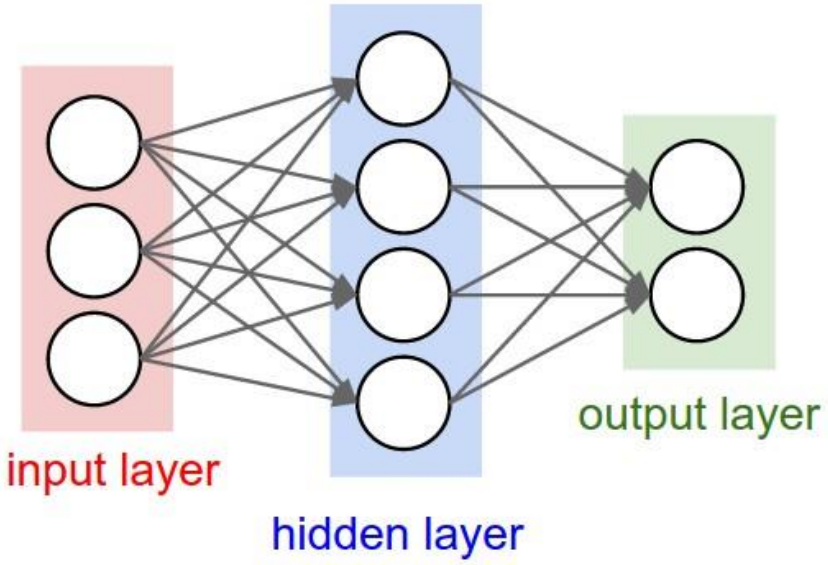
Word2Vec

……

Word2Vec

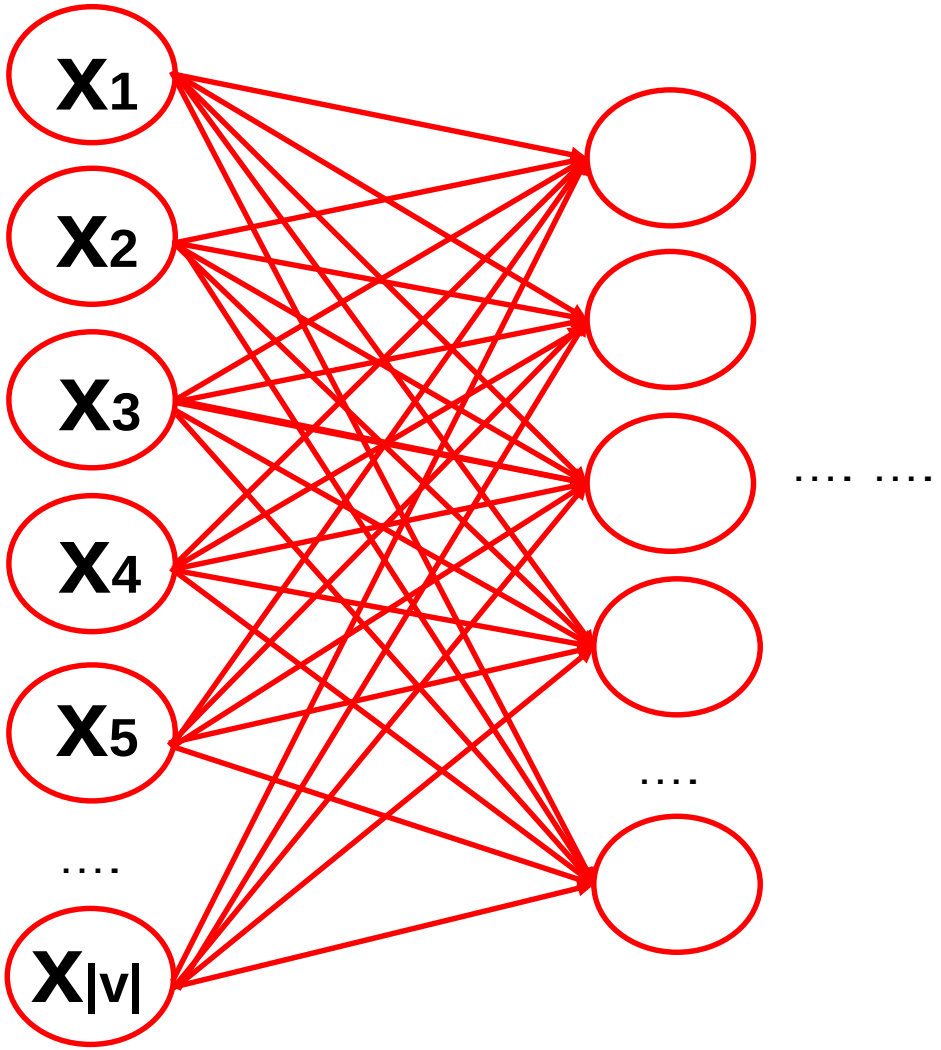
In a nutshell

- **Distributional Model** to learn Word Embeddings.
- Uses a **two-layers neural networks**
- Training based on the **Skip-Gram methodology**
- **Update of the network** through Mini-batch or Stochastic Gradient Descent



Word2Vec

Structure of the network

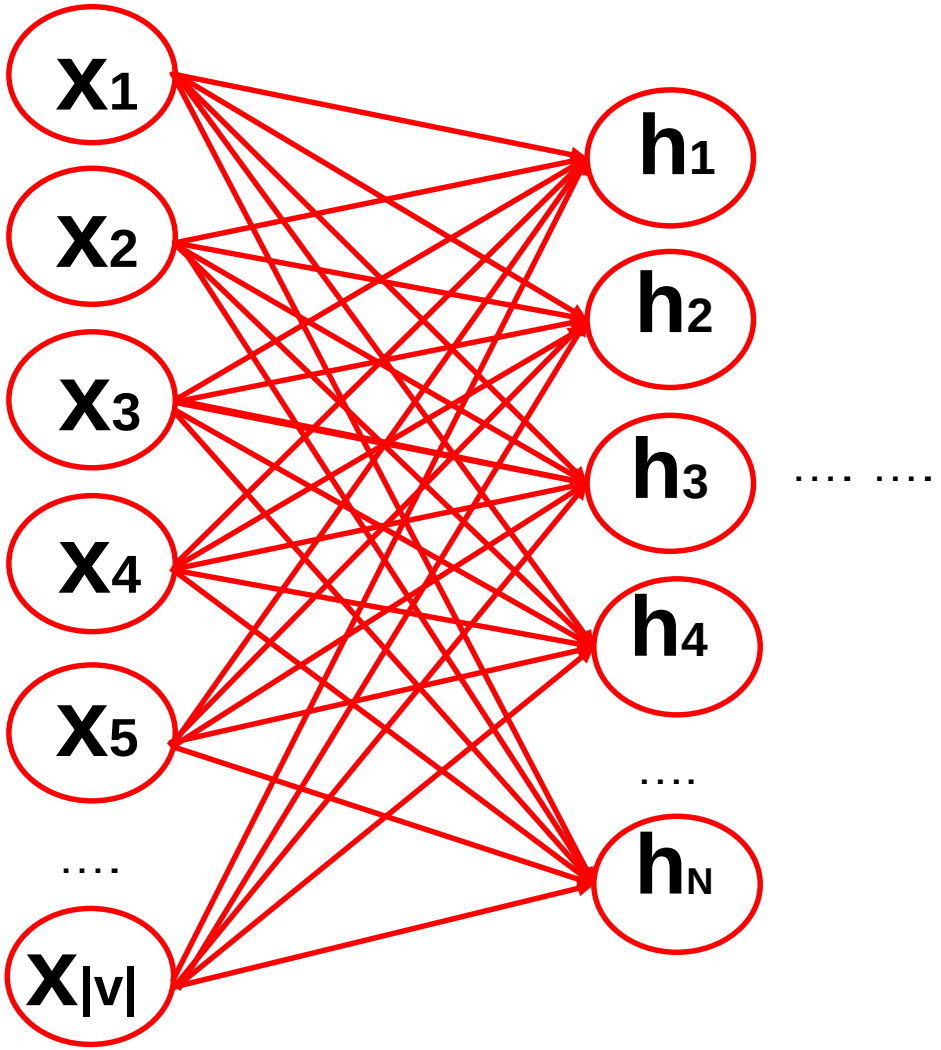


Input Layer:

- **Vocabulary V**
  - |V| number of terms
  - |V| nodes
  - **Each term is represented through a «one hot representation»**

Word2Vec

Structure of the network



Input Layer:

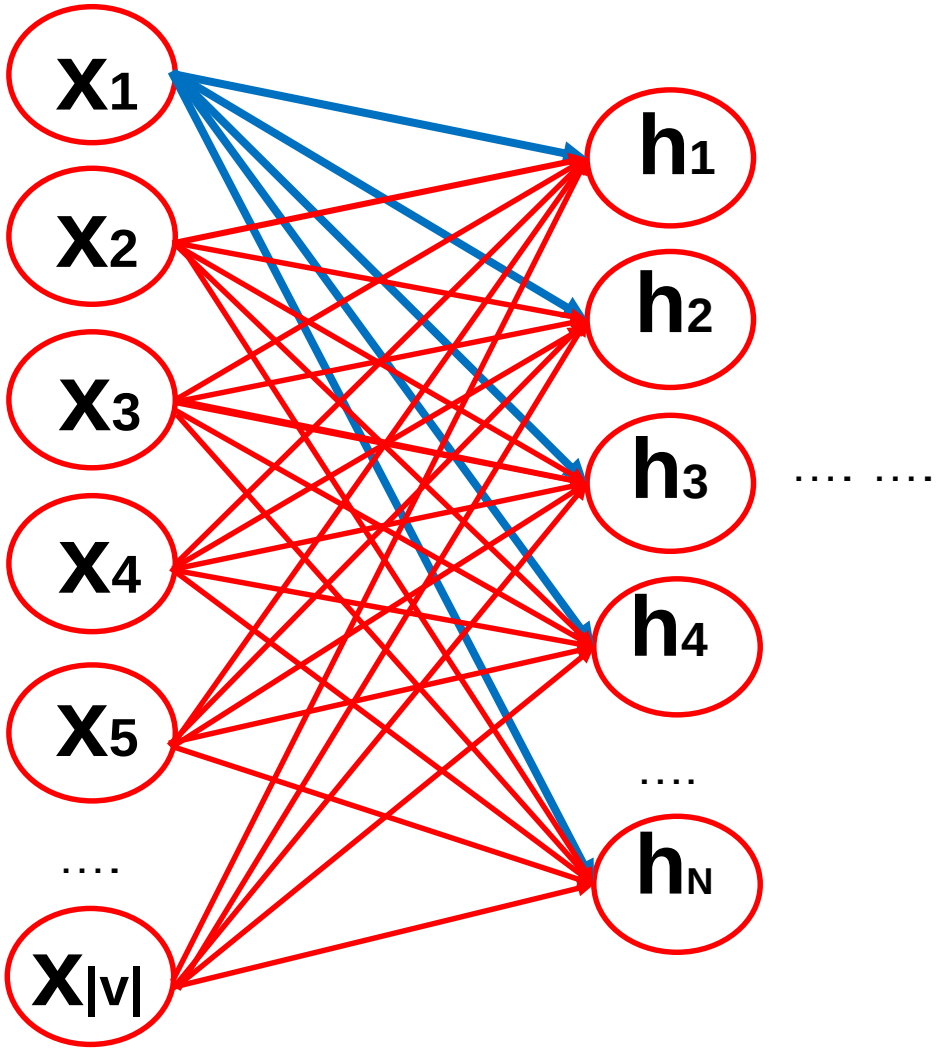
- **Vocabulary V**
  - |V| number of terms
  - |V| nodes
  - **One-hot representation**

Hidden Layer:

- **N nodes**
  - N = size of the embeddings
  - Parameter of the model

Word2Vec

Structure of the network



Input Layer:

- **Vocabulary V**
  - |V| number of terms
  - |V| nodes
  - **One-hot representation**

Hidden Layer:

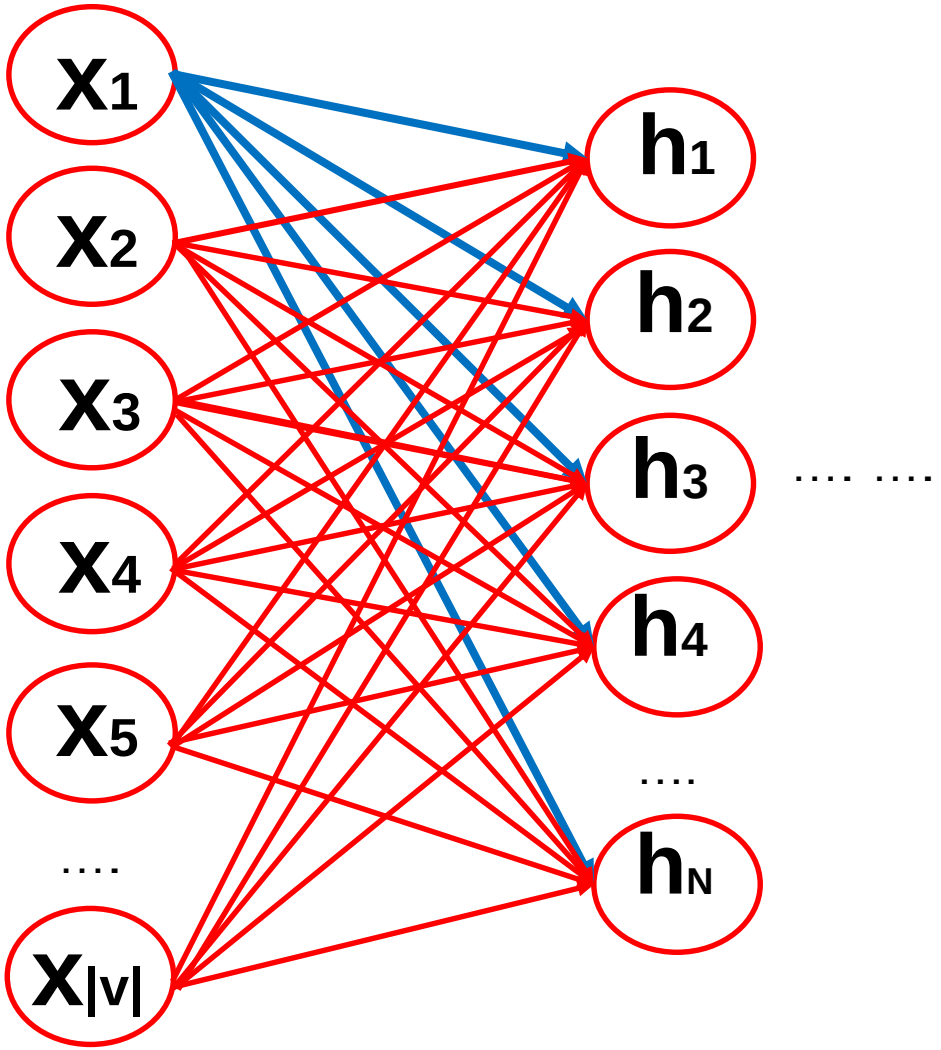
- **N nodes**
  - N = size of the embeddings
  - Parameter of the model

Weight of the network:

- **Randomly set (initially)**
- **Updated through the training**

Word2Vec

Structure of the network



Input Layer:

- **Vocabulary V**
  - |V| number of terms
  - |V| nodes
  - **One-hot representation**

Hidden Layer:

- **N nodes**
  - N = size of the embeddings
  - Parameter of the model

Weight of the network:

- **Randomly set (initially)**
- **Updated through the training**

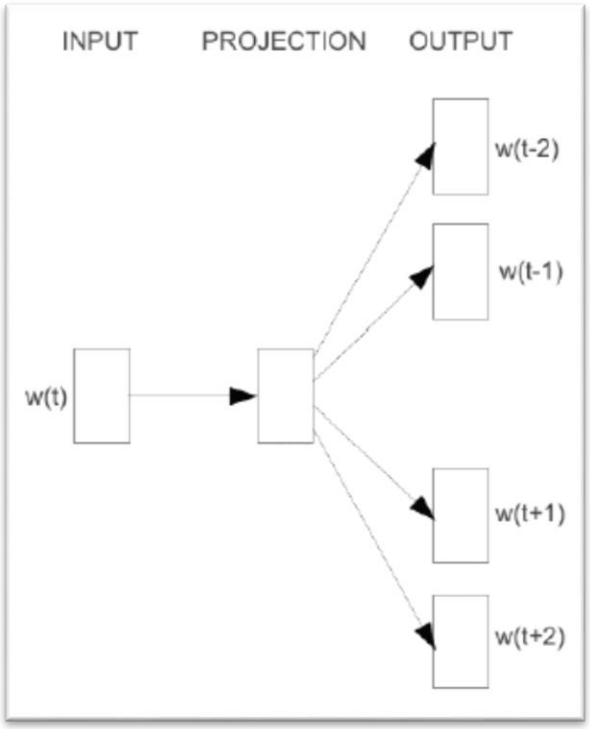
Final Representation for term t_k

- Weights Extracted from the network
- $t_k=[Wt_{kV1}, wt_{kV2} \ldots wt_{kVn}]$

Word2Vec

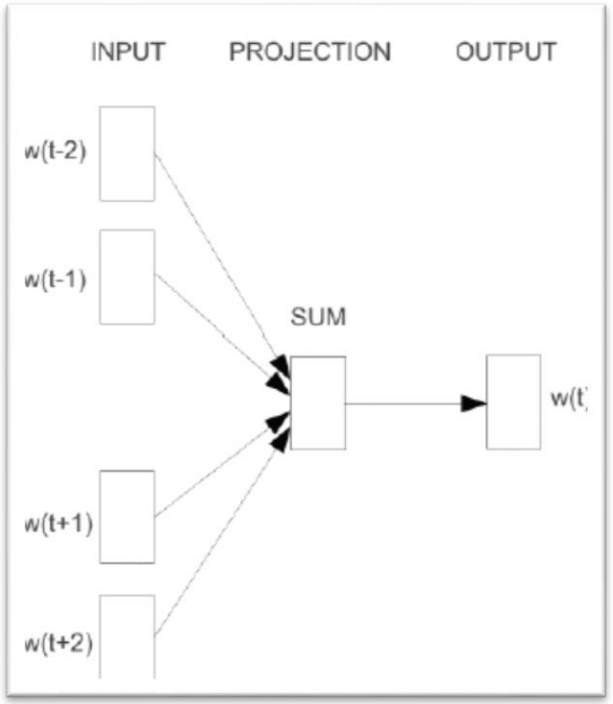
Training Procedure: how to create training examples?

Skip-Gram Methodology



Given **a word w(t),** predict its **context w(t-2), t(t-1).. w(t+1), w(t+2)**

Continuous Bag-of-Words Methodology

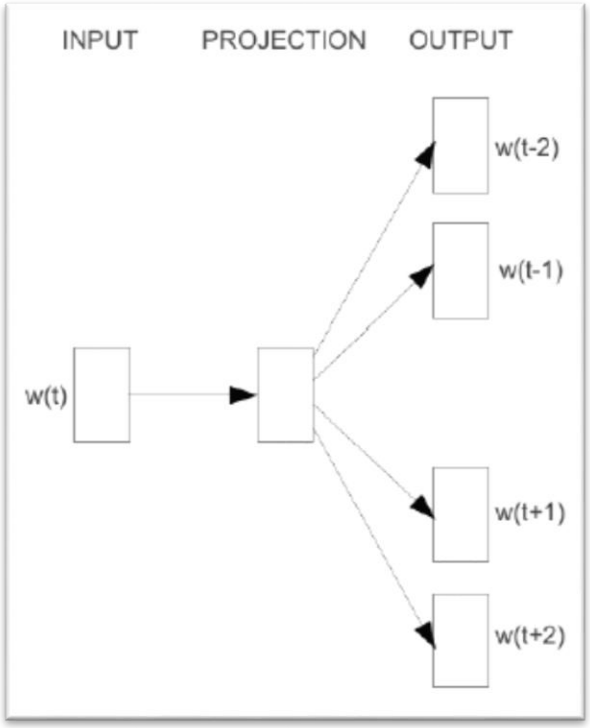


Given **a context w(t-2), t(t-1).. w(t+1), w(t+2)** predict **word w(t)**

Word2Vec

Training Procedure: how to create training examples?

Skip-Gram Methodology



Given **a word w(t),** predict its **context w(t-2), t(t-1).. w(t+1), w(t+2)**

Example

Input: ”the quick brown fox jumped over the lazy dog”

Window Size: 1

Contexts:

- ([the, brown], quick)
- ([quick, fox], brown)
- ([brown, jumped], fox) ...

Training Examples:

- (quick, the)
- (quick, brown)
- (brown, quick)
- (brown, fox) ...

Word2Vec

Training Procedure: how to evaluate the model?

Given a corpus, we create a **lot of training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context C given a word w

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

And probability is calculated **through soft-max**

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Word2Vec

Training Procedure: how to evaluate the model?

Given a corpus, we create a **lot of training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context c given a word w

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

And probability is calculated **through soft-max**

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Intuitively, probability is high when scalar product is close to 1 → when **vectors are similar!**

Word2Vec

Training Procedure: how to evaluate the model?

Given a corpus, we create a **lot of training examples through Skip-Gram**.

The model tries to maximize The probability of predicting a context C given a word w

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w)$$

And probability is calculated **through soft-max**

$$p(c|w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Intuitively, probability is high when scalar product is close to 1 → when **vectors are similar!**

Word2Vec is a distributional model since it learns a representation such that couples (word,context) appearing together have similar vectors

The error is collected and weights in the network are updated accordingly. **Typically is used Stochastic Gradient Descent or Mini-Batch (every 128 or 512 training examples)**

Word2Vec

Learning Word Embeddings through Neural Networks: it is not based on «counting» co-occurrences. It relies on «predict» the distribution

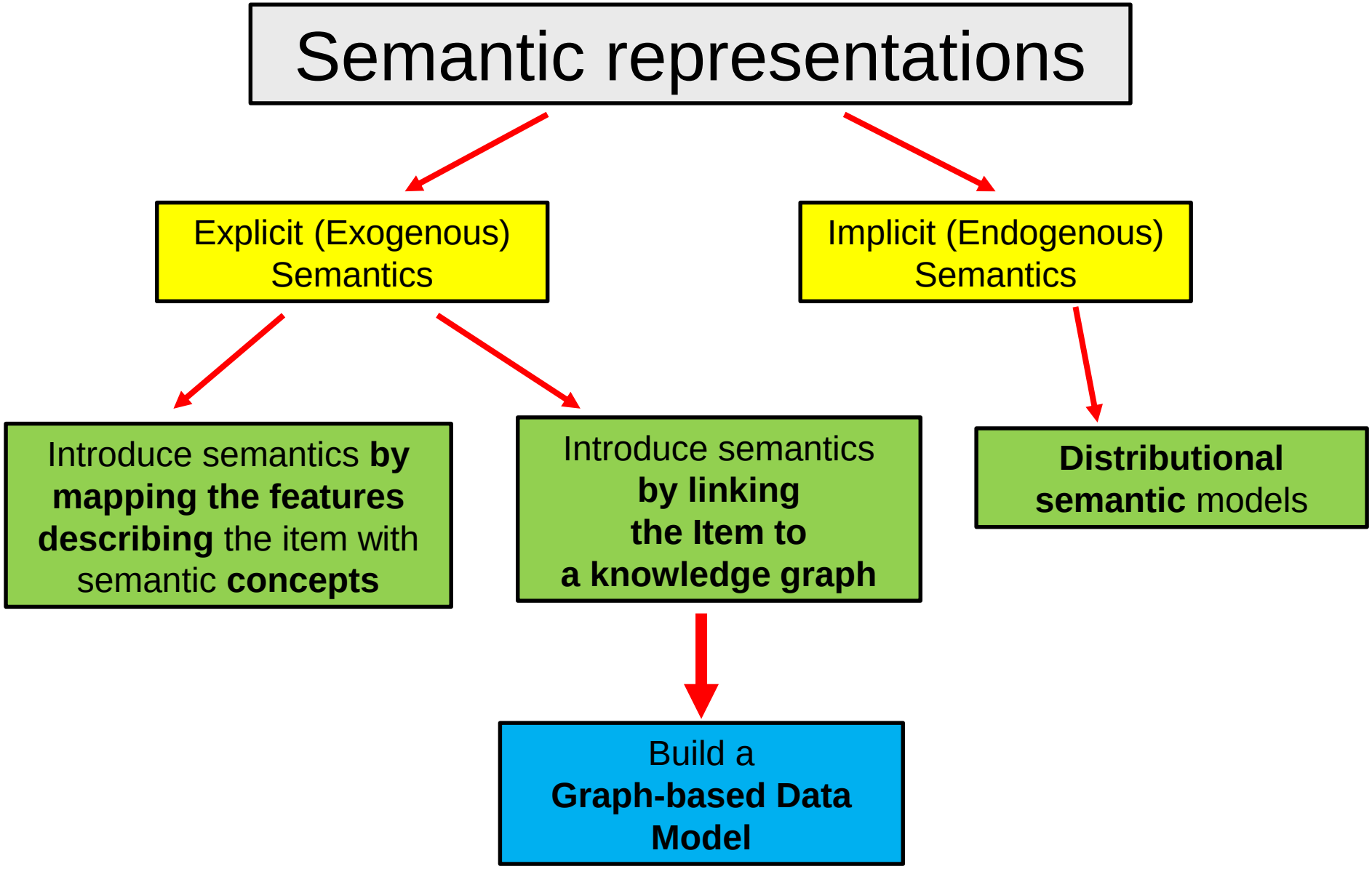
Representation can be really really small (size≤100, typically)

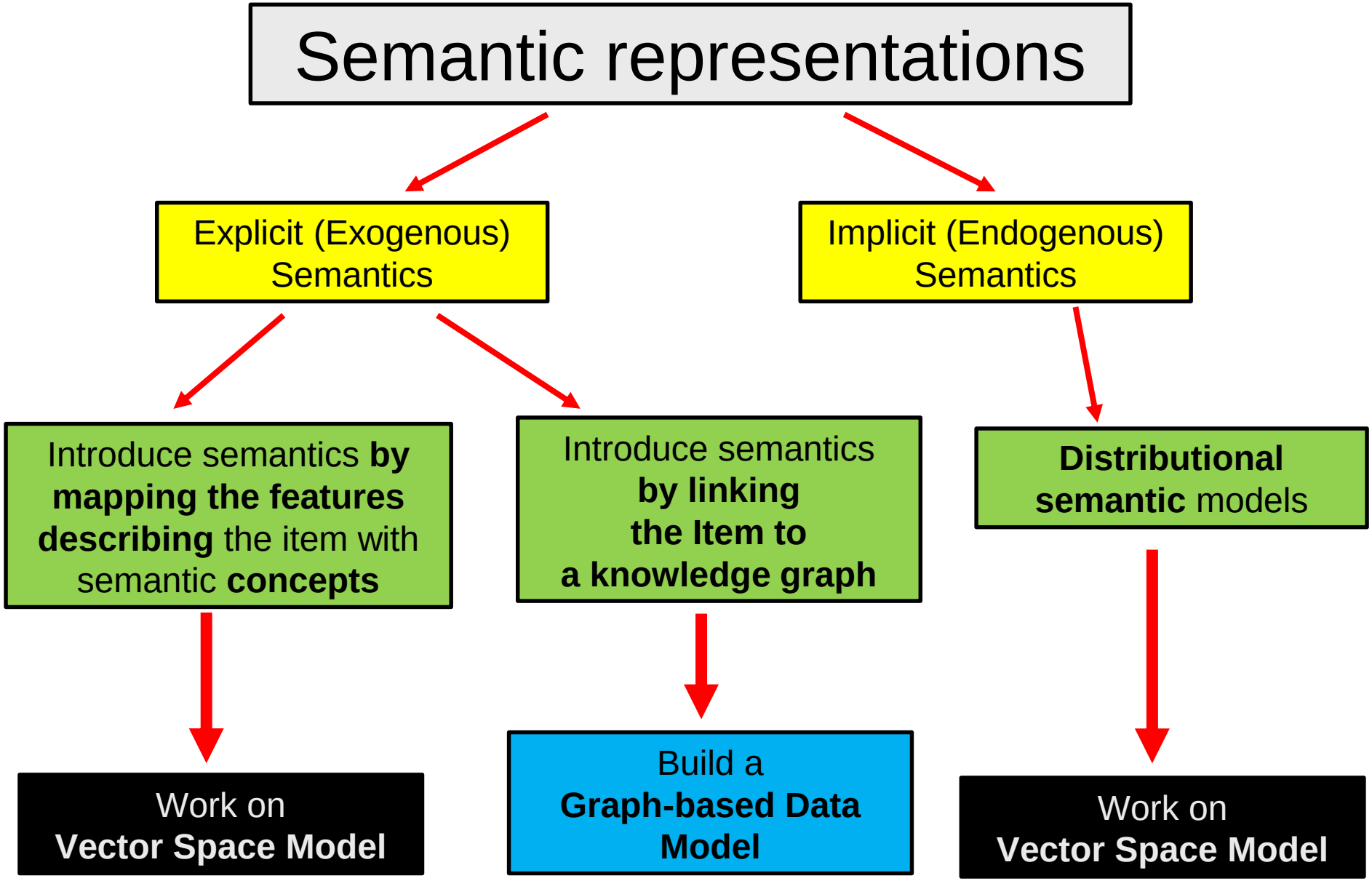
Trending ☺ - Recent and Very Hot technique

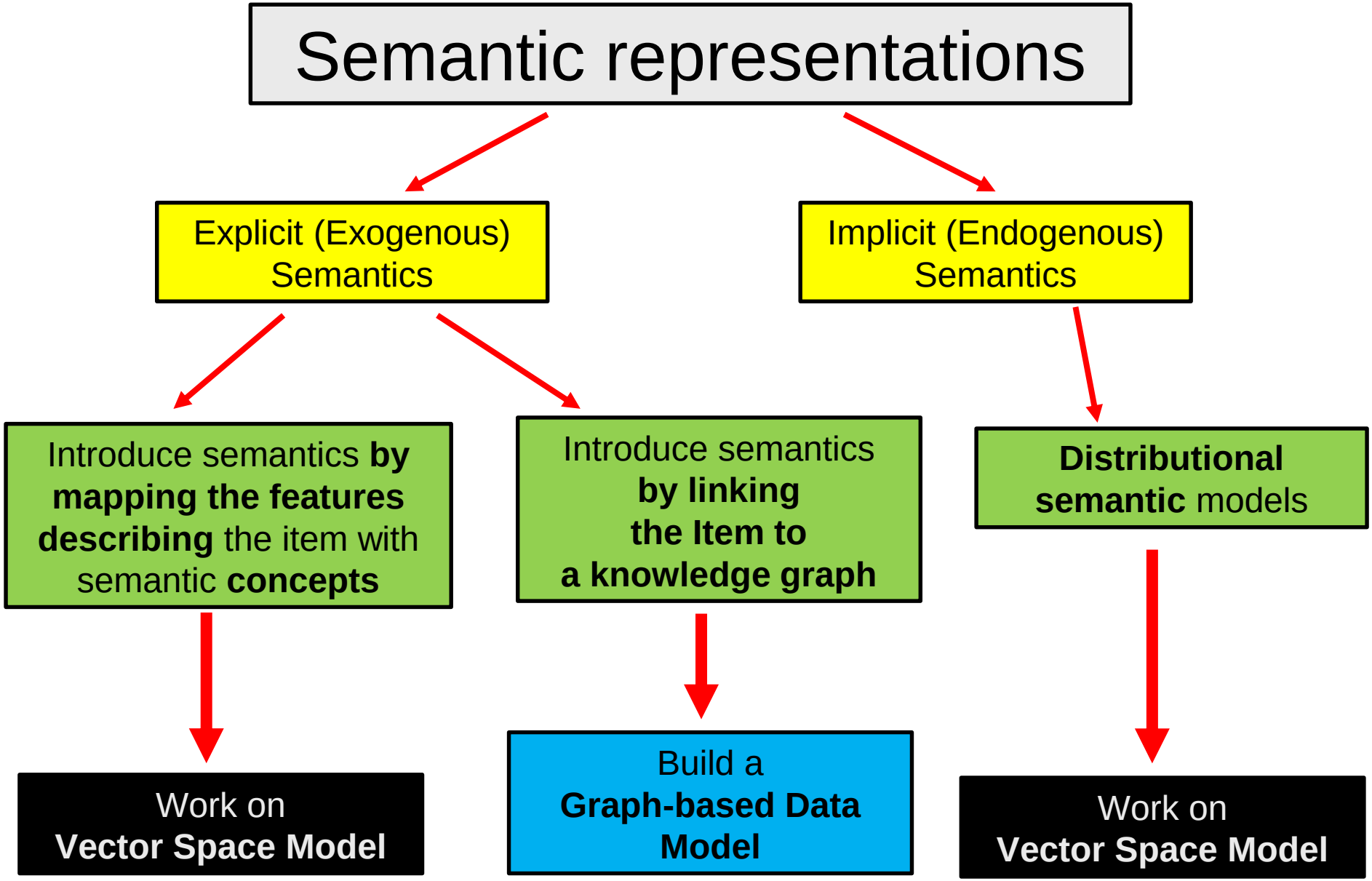
Not transparent anymore

Needs more computational resources

...Let's put everything together


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking the Item to a knowledge graph**
Distributional semantic models
Build a **Graph-based Data Model**


Semantic representations
Explicit (Exogenous) Semantics
Implicit (Endogenous) Semantics
Introduce semantics **by mapping the features describing** the item with semantic **concepts**
Introduce semantics **by linking the Item to a knowledge graph**
Distributional semantic models
Work on **Vector Space Model**
Build a **Graph-based Data Model**
Work on **Vector Space Model**



Can Exogenous and Endogenous approaches be combined?

...Let's put everything together

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✓	✓			✓	✓		
concept2		✓	✓			✓	✓	✓	
concept3	✓			✓				✓	✓
concept4	✓	✓	✓		✓				✓

Exogenous Approaches as Entity Linking and WSD
work on the row of the matrix

...Let's put everything together

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✓	✓			✓	✓		
concept2		✓	✓			✓	✓	✓	
concept3	✓			✓				✓	✓
concept4	✓	✓	✓		✓				✓

Exogenous Approaches as Entity Linking and WSD
work on the row of the matrix

Endogenous Approaches as ESA or Word2Vec
work on the columns of the matrix

...Let's put everything together

	c1	c2	c3	c4	c5	c6	c7	c8	c9
concept1		✓	✓			✓	✓		
concept2		✓	✓			✓	✓	✓	
concept3	✓			✓				✓	✓
concept4	✓	✓	✓		✓				✓

Exogenous Approaches as Entity Linking and WSD
work on the row of the matrix

Endogenous Approaches as ESA or Word2Vec
work on the columns of the matrix

Both approaches can be combined to obtain richer and more precise semantic representations
(e.g. Word2Vec over textual description processed with WSD)

What?

semantics-aware recommender systems

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

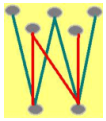
Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Word Sense Disambiguation

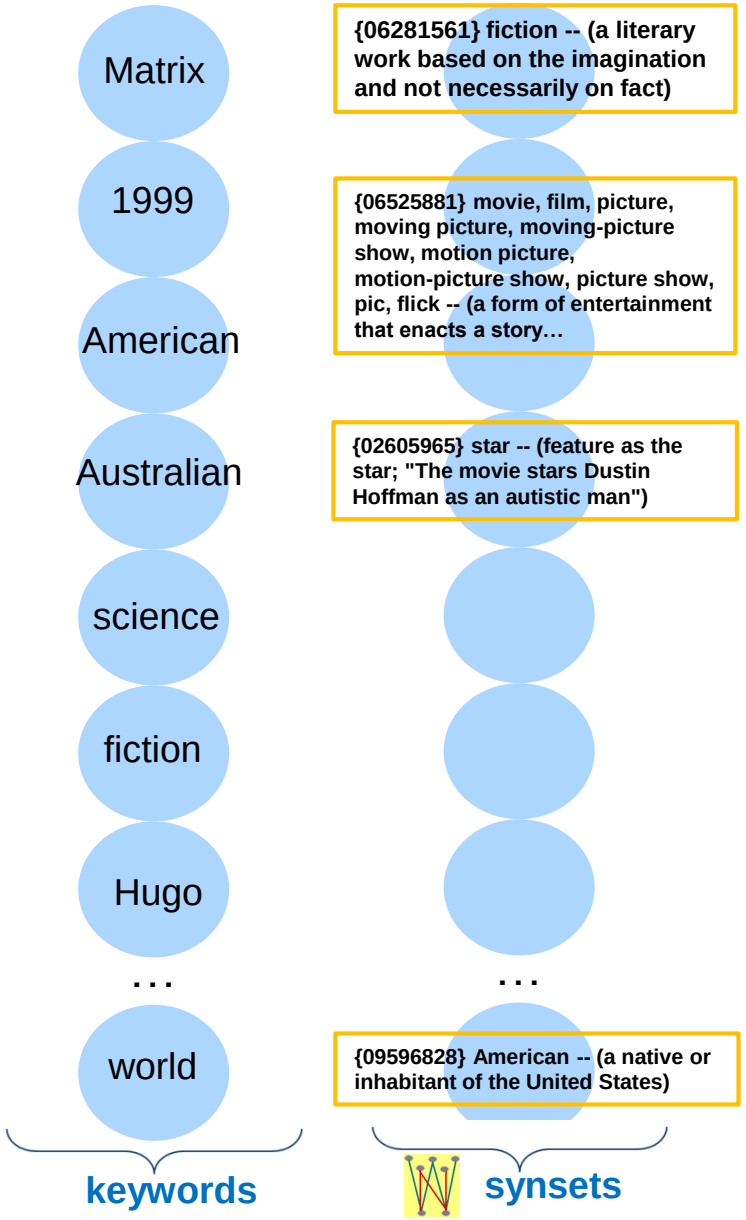
Entity Linking

.....



Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Giovanni Semeraro. **Semantics-Aware Content-Based Recommender Systems**. Recommender Systems Handbook 2015: 119-159

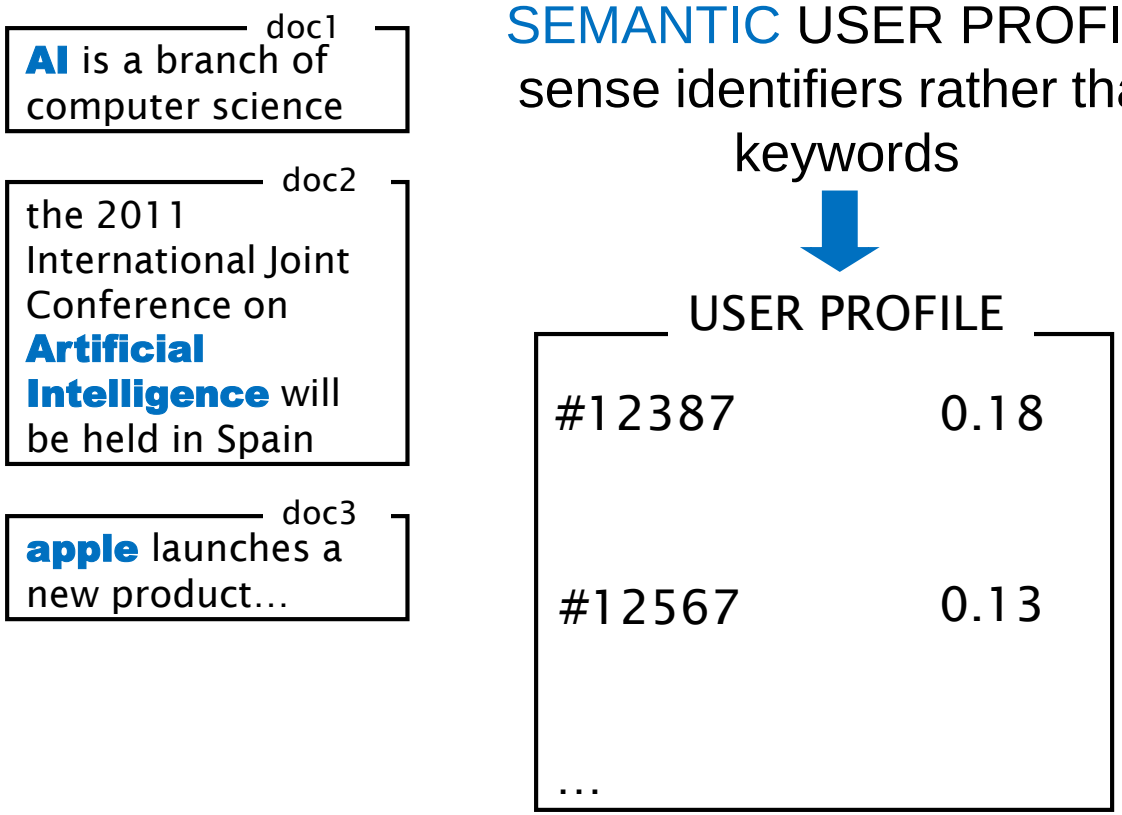
The Matrix representation



through WSD we process the textual description of the item **and we obtain a semantics-aware representation of the item as output.**

In this case, keyword-based features **are replaced with the concepts (in this case, a WordNet synset) they refer to.**

Synset-based representation



M. Degemmis, P. Lops, and G. Semeraro. A Content-collaborative Recommender that Exploits WordNet-based User Profiles for Neighborhood Formation. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI)*, 17(3):217–255, Springer Science + Business Media B.V., 2007.

G. Semeraro, M. Degemmis, P. Lops, and P. Basile. Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. In M. M. Veloso, editor, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007* , pages 2856–2861. Morgan Kaufmann, 2007.

M. de Gemmis, P. Lops, G. Semeraro, Pierpaolo Basile: Integrating tags in a semantic content-based recommender ACM Conference on Recommender Systems, RecSys 2008: 163-170

Keywords- vs synsets-based profiles

EachMovie dataset

- 1,628 movies grouped into 10 categories
- Users who rated between 30 and 100 movies
- Movie content crawled from IMDb

Algorithm	Precision	Recall	F1
Rocchio	Kwd 0.74 Syn 0.76	Kwd 0.81 Syn 0.84	Kwd 0.76 Syn 0.78
Naive Bayes	Kwd 0.67 Syn 0.75	Kwd 0.78 Syn 0.88	Kwd 0.73 Syn 0.81

Results in a cultural heritage scenario

***C**ultural **H**eritage fruition & e-learning applications of new **A**dvanced (multimodal) **T**echnologies*

In the context of cultural heritage personalization, does the integration of UGC and textual description of artwork collections cause an increase of the prediction accuracy in the process of recommending artifacts to users?

M. de Gemmis, P. Lops, G. Semeraro, and P. Basile. **Integrating Tags in a Semantic Content-based Recommender**. *In RecSys '08, Proceed. of the 2nd ACM Conference on Recommender Systems, pages 163–170, October 23-25, 2008, Lausanne, Switzerland, ACM, 2008.*

Results in a cultural heritage scenario

27) Caravaggio - Deposition from the Cross

Textual description of items (static content)

Descrizione dell'opera

The Deposition, considered one of Caravaggio's greatest masterpieces, was commissioned by Girolamo Vittrice for his famil chapel in S. Maria in Vallicella (Chiesa Nuova) in Rome. In 1797 it was included in the group of works transferred to Paris i execution of the Treaty of Tolentino. After its return in 1817 it became part of Pius VII's Pinacoteca. Caravaggio did not rea portray the Burial or the Deposition in the traditional way, inasmuch as Christ is not shown at the moment when he is laid in t tomb, but rather when, in the presence of the holy women, he is laid by Nicodemus and John on the Anointing Stone, that is stone with which the sepulchre will be closed. Around the body of Christ are the Virgin, Mary Magdalene, John, Nicodemus and Mary of Cleophas, who raises her arms and eyes to heaven in a gesture of high dramatic tension. Caravaggio, who arriv in Rome towards 1592-93, was the protagonist of a real artistic revolution as regards the way of treating subjects and the us of colour and light, and was certainly the most important personage of the "realist" trend of seventeenth century painting.

Social Tags

Social Tags (from other users): caravaggio, deposition, christ, cross, suffering, religion

Inserisci il tuo voto e dei tag descrittivi (separati da una VIRGOLA, senza spazi)

5-point rating scale

1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Personal Tags

passion

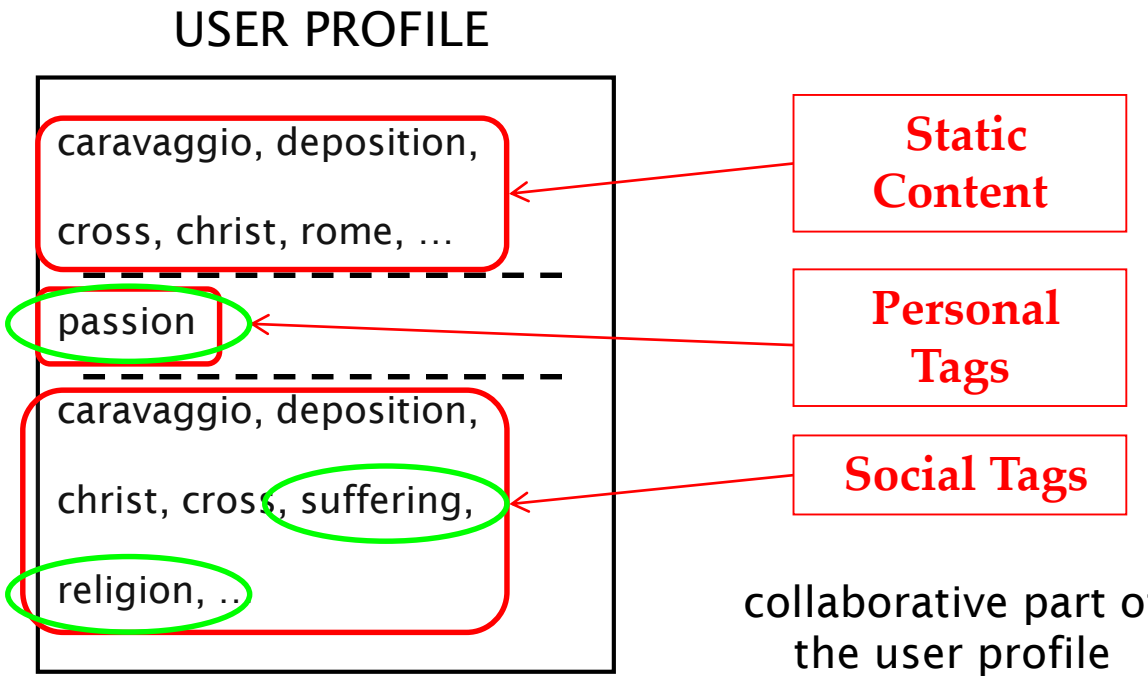
Inserisci i voti e prosegui

Results in a cultural heritage scenario

27) Caravaggio - Deposition from the Cross

Descrizione dell'opera

The Deposition, considered one of Caravaggio's greatest masterpieces, was commissioned by Girolamo Vittrice for his famil chapel in S. Maria in Vallicella (Chiesa Nuova) in Rome. In 1797 it was included in the group of works transferred to Paris i execution of the Treaty of Tolentino. After its return in 1817 it became part of Pius VII's Pinacoteca. Caravaggio did not rea portray the Burial or the Deposition in the traditional way, inasmuch as Christ is not shown at the moment when he is laid in t tomb, but rather when, in the presence of the holy women, he is laid by Nicodemus and John on the Anointing Stone, that is stone with which the sepulchre will be closed. Around the body of Christ are the Virgin, Mary Magdalene, John, Nicodemus and Mary of Cleophas, who raises her arms and eyes to heaven in a gesture of high dramatic tension. Caravaggio, who arriv in Rome towards 1592-93, was the protagonist of a real artistic revolution as regards the way of treating subjects and the us of colour and light, and was certainly the most important personage of the "realist" trend of seventeenth century painting.



Results in a cultural heritage scenario

- Artwork representation
  - Artist
  - Title
  - Description
  - Tags

- change of text representation from vectors of words (BOW) into vectors of WordNet synsets (BOS)
  - From tags to semantic tags
- supervised Learning
  - Bayesian Classifier learned from artworks labeled with user ratings and tags

Results in a cultural heritage scenario

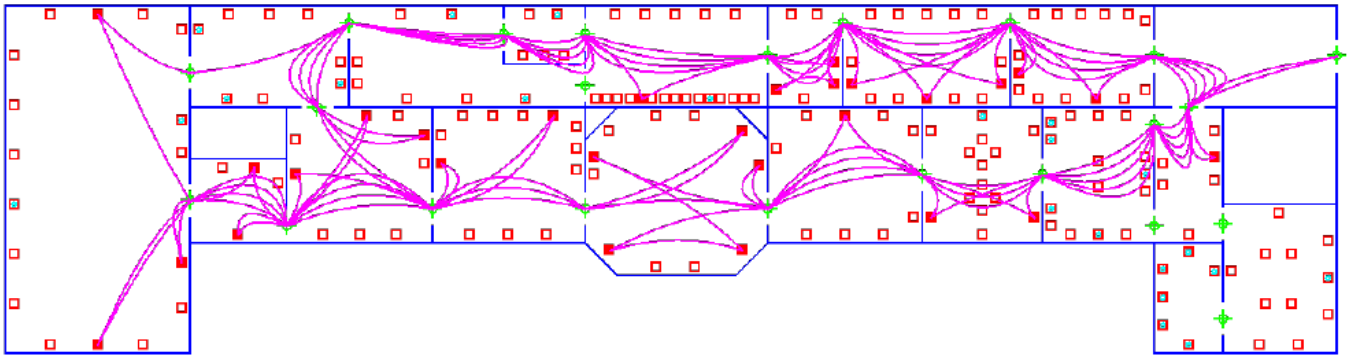
	Type of Content	Precision*	Recall*	F1*
Content-based Profiles	EXP#1: Static Content	75.86	94.27	84.07
Tag-based Profiles	EXP#2: Personal Tags	75.96	92.65	83.48
	EXP#3: Social Tags	75.59	90.50	82.37
Augmented Profiles	EXP#4: Static Content + Personal Tags	78.04	93.60	**85.11**
	EXP#5: Static Content + Social Tags	78.01	93.19	**84.93**

* Results averaged over the 30 study subjects

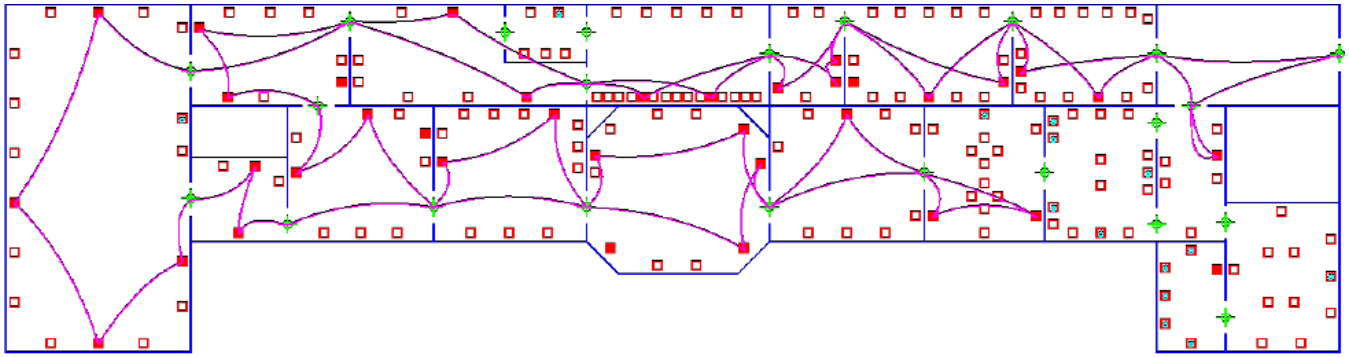
Overall accuracy F1 ≈ 85%

Results in a cultural heritage scenario

personalized museum tours by arranging the most interesting items for the active user



step forward to take into account **spatial layout** & **time constraint**



L. Iaquinta, M. de Gemmis, P. Lops, G. Semeraro: **Recommendations toward Serendipitous Diversions**. *ISDA 2009: 1049-1054*

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Ontologies

Linked Open Data

DBpedia

.....

Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Giovanni Semeraro. **Semantics-Aware Content-Based Recommender Systems**. Recommender Systems Handbook 2015: 119-159

Semantic Analysis using Ontologies

Quickstep & Foxtrot

- on-line academic **research papers** recommenders
- **items** and **user profiles** represented through a **research topic ontology**
- **is-a relationships** exploited to infer general interests when specific topics are observed
- match based on the **correlation** between the topics in the user profile and topics in papers

S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. **Ontological User Profiling in Recommender Systems**. *ACM Transactions on Information Systems, 22(1):54–88, 2004*

Semantic Analysis using Ontologies

News@hand

- **news** descriptions as **vectors of TF-IDF** scores in the space of **ontology concepts**
- user profiles represented in the same space
- news-profile matching performed using **cosine-based vector similarity**

I. Cantador, A. Bellogin, and P. Castells. **News@hand: A Semantic Web Approach to Recommending News**. *In W. Nejdl, J. Kay, P. Pu, and E. Herder, editors, Adaptive Hypermedia and AdaptiveWeb-Based Systems, volume 5149 of Lecture Notes in Computer Science, pages 279–283. Springer, 2008.*

Semantic Analysis using Ontologies

- user interests propagation from concepts which received the user feedback to others related ones though **spreading activation**
- **contextualized propagation strategies** of user interests
  - ✓ horizontal propagation among siblings
  - ✓ **anisotropic** vertical propagation, i.e. user interests propagated differently upward and downward

F. Cena, S. Likavec, and F. Osborne. **Anisotropic Propagation of User Interests in Ontology-based User Models**. *Inf. Sci., 250:40–60, 2013.*

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Ontologies

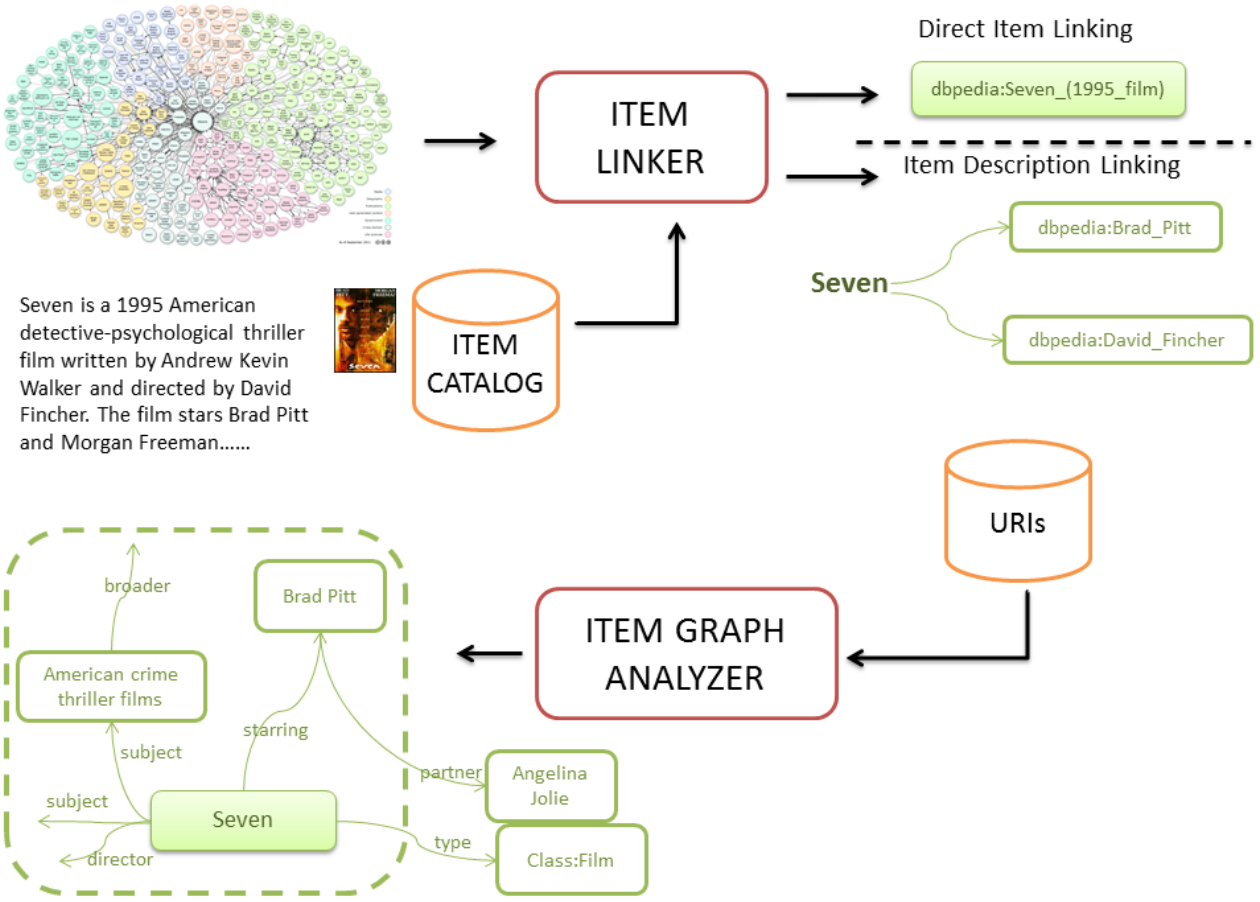
Linked Open Data

DBpedia

.....

Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Giovanni Semeraro. **Semantics-Aware Content-Based Recommender Systems**. Recommender Systems Handbook 2015: 119-159

A high level architecture



Tommaso Di Noia, Vito Claudio Ostuni, Paolo Tomeo, Eugenio Di Sciascio: **SPrank: Semantic Path-Based Ranking for Top-N Recommendations Using Linked Open Data**. ACM TIST 8(1): 9:1-9:34 (2016)

Direct Item Linking

Datasets

Subset of **Movielens** mapped to DBpedia

Subset of **Last.fm** mapped to DBpedia

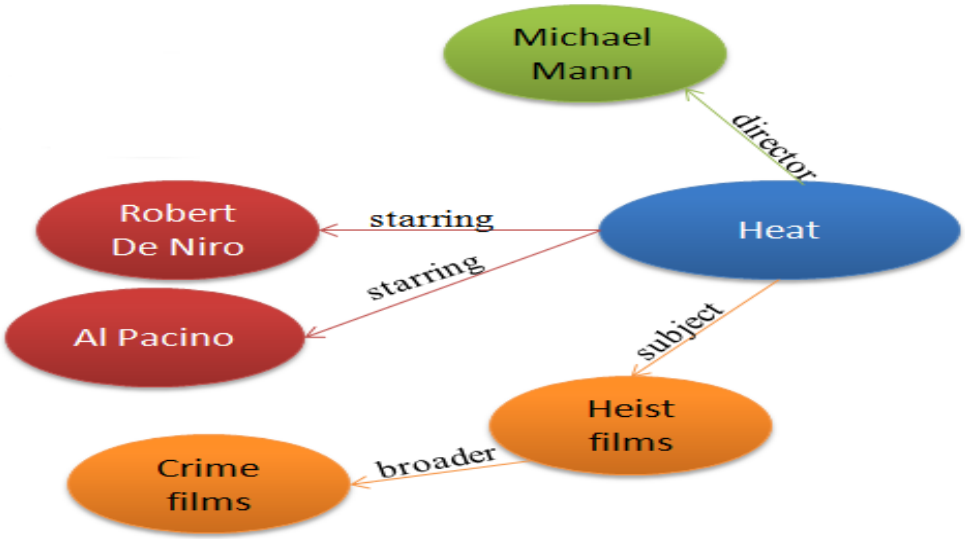
Subset of **The Library Thing** mapped to DBpedia

Mappings

https://github.com/sisinflab/LODrecsys-datasets

Linked Open Data & RecSys

structured information source for item descriptions

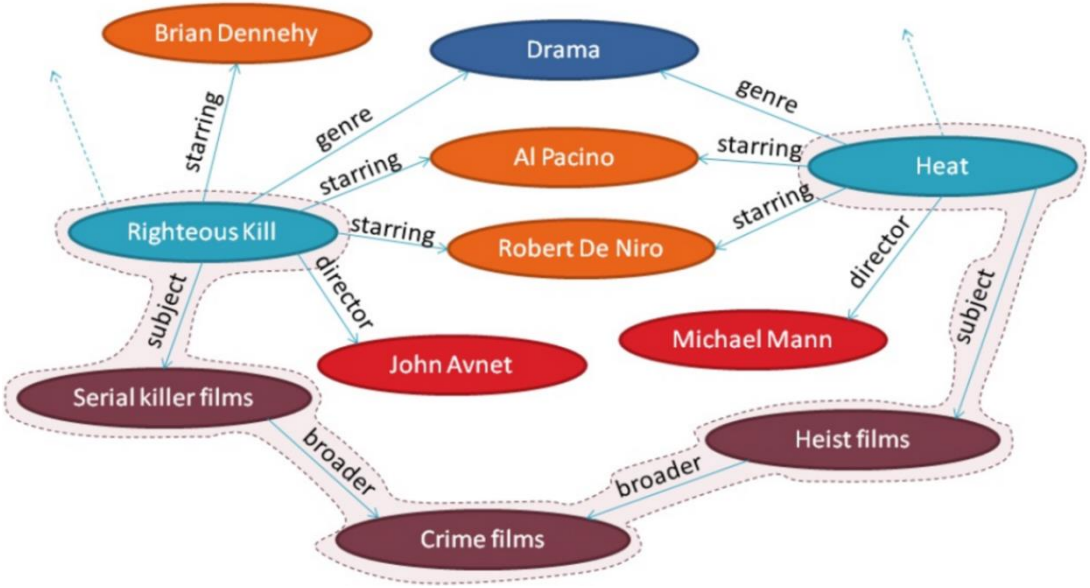


Rich item representation

	STARRING		DIRECTOR	SUBJECT+BROADER	
Heat	Robert DeNiro	Al Pacino	Michael Mann	Heist films	Crime films

LOD & Recommender Systems

Simplest RecSys: common features



$$sim_{jaccard}(x_i, x_j) = \frac{|N_d(x_i) \cap N_d(x_j)|}{|N_d(x_i) \cup N_d(x_j)|}$$

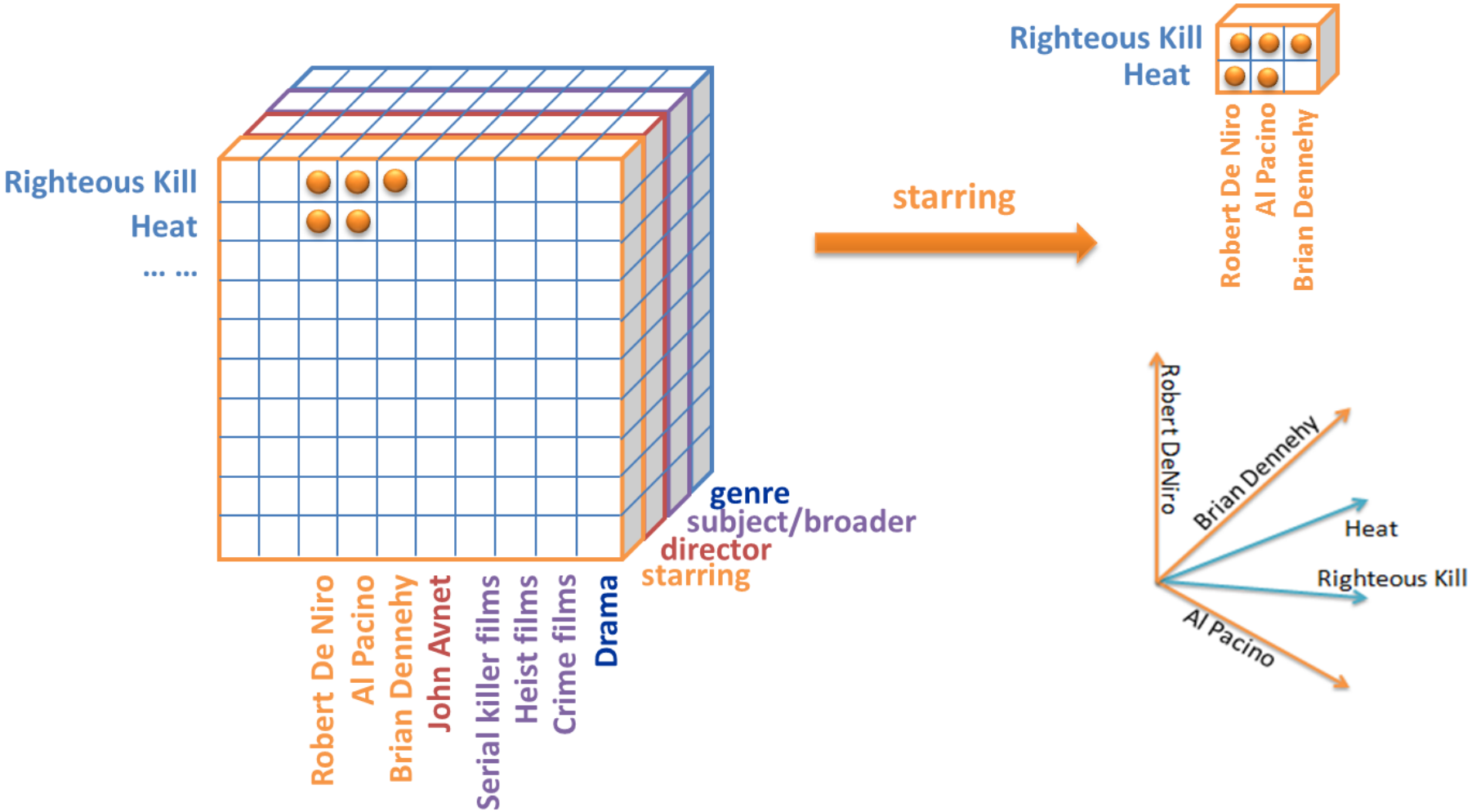
similarity between items

$$\tilde{r}(u, x_j) = \frac{\sum_{x_i \in N \cap Profile(u)} r(u, x_i) \cdot sim(x_i, x_j)}{\sum_{x_i \in N \cap Profile(u)} sim(x_i, x_j)}$$

content-based prediction

LOD & Recommender Systems

Vector Space Model for LOD



Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD & Recommender Systems

Vector Space Model for LOD

Righteous Kill

STARRING	Al Pacino (v1)	Robert De Niro (v2)	Brian Dennehy (v3)
Righteous Kill (m1)	X	X	X
Heat (m2)	X	X	

Heat

$$w_{AlPacino,Heat} = tf_{AlPacino,Heat} * idf_{AlPacino}$$

Righteous Kill (x1)	$w_{v1,x1}$	$w_{v2,x1}$	$w_{v3,x1}$
Heat (x2)	$w_{v1,x2}$	$w_{v2,x2}$	0

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD & Recommender Systems

VSM content-based recommender

$$\tilde{r}(u, x_j) = \frac{\sum_{x_i \in Profile(u)} r(u, x_i) \cdot \frac{\sum_{p \in P} \alpha_p \cdot sim_p(x_i, x_j)}{|P|}}{|profile(u)|}$$

Predict the rating using a **Nearest Neighbor Classifier** wherein the similarity measure is a linear combination of **local property similarities**

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. Linked Open Data to support Content-based Recommender Systems. 8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)

LOD & Recommender Systems

Vector Space Model for LOD

$$sim_{starring}(\overrightarrow{x_i}, \overrightarrow{x_j}) = \frac{w_{v_1,x_i} * w_{v_1,x_j} + w_{v_2,x_i} * w_{v_2,x_j} + w_{v_3,x_i} * w_{v_3,x_j}}{\sqrt{w_{v_1,x_i}^2 + w_{v_2,x_i}^2 + w_{v_3,x_i}^2} * \sqrt{w_{v_1,x_j}^2 + w_{v_2,x_j}^2 + w_{v_3,x_j}^2}}$$

$$\alpha_{starring} * sim_{starring}(\overrightarrow{x_i}, \overrightarrow{x_j}) +$$

$$\alpha_{director} * sim_{director}(\overrightarrow{x_i}, \overrightarrow{x_j}) +$$

$$\alpha_{subject} * sim_{subject}(\overrightarrow{x_i}, \overrightarrow{x_j}) +$$

$$\ldots =$$

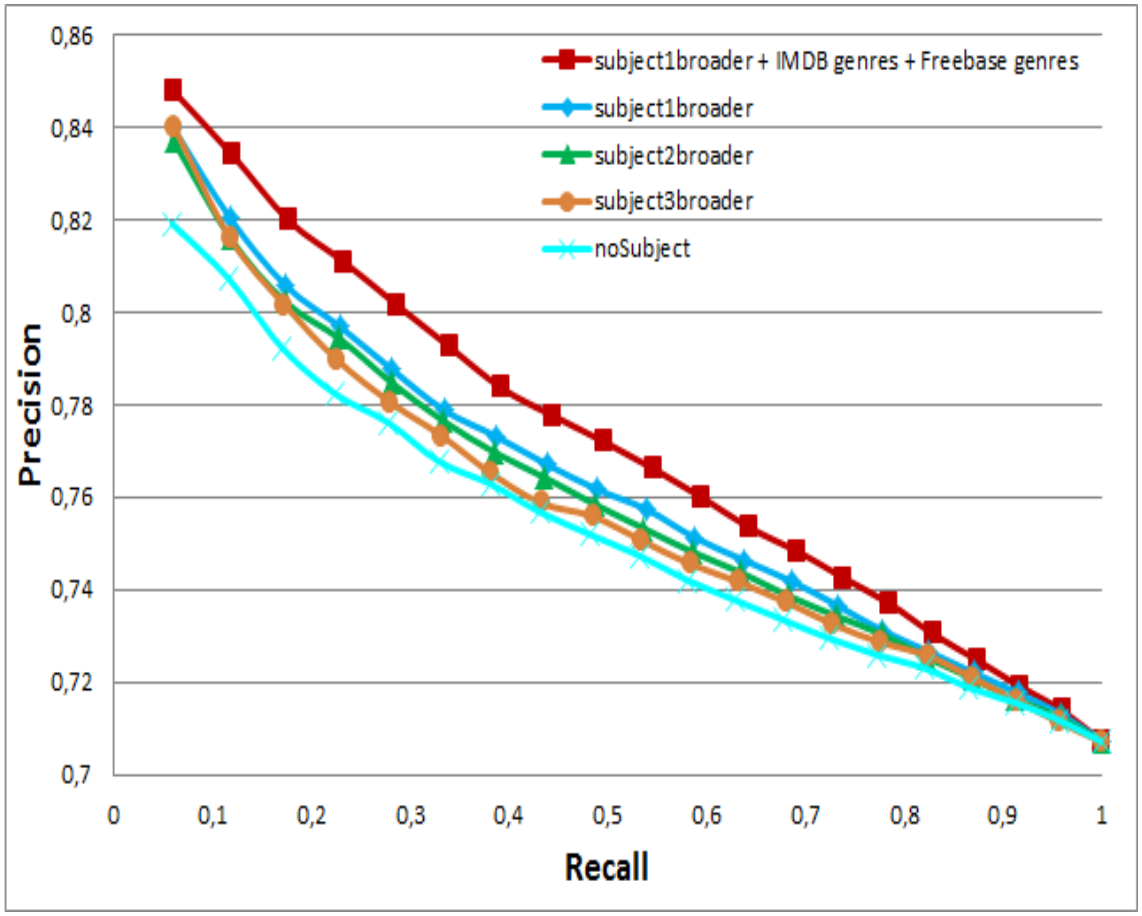
$$sim\ (\overrightarrow{x_i}, \overrightarrow{x_j})$$

Similarity between items as linear combination of the similarity **among Dbpedia facets** *(starring, directors, subject, etc.)*

Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD & Recommender Systems

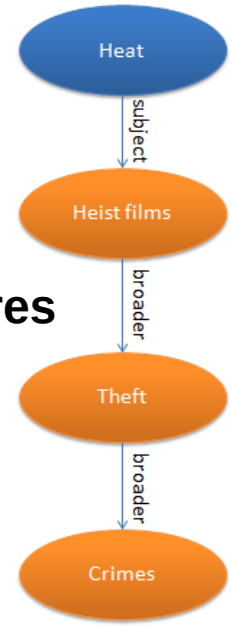
Property subset evaluation



subject+broader solution better than only subject or subject+more broaders

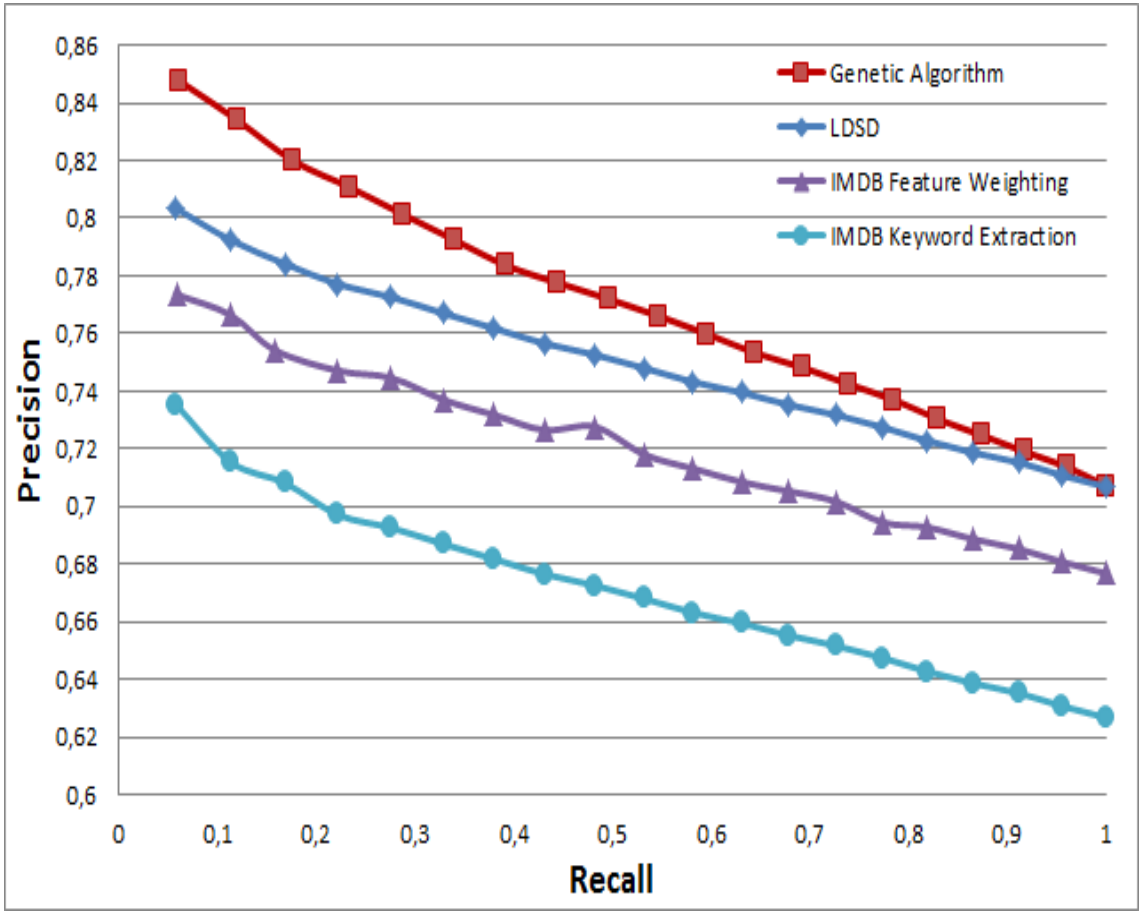
too many broaders introduce noise

best solution achieved with **subject+broader+genres**



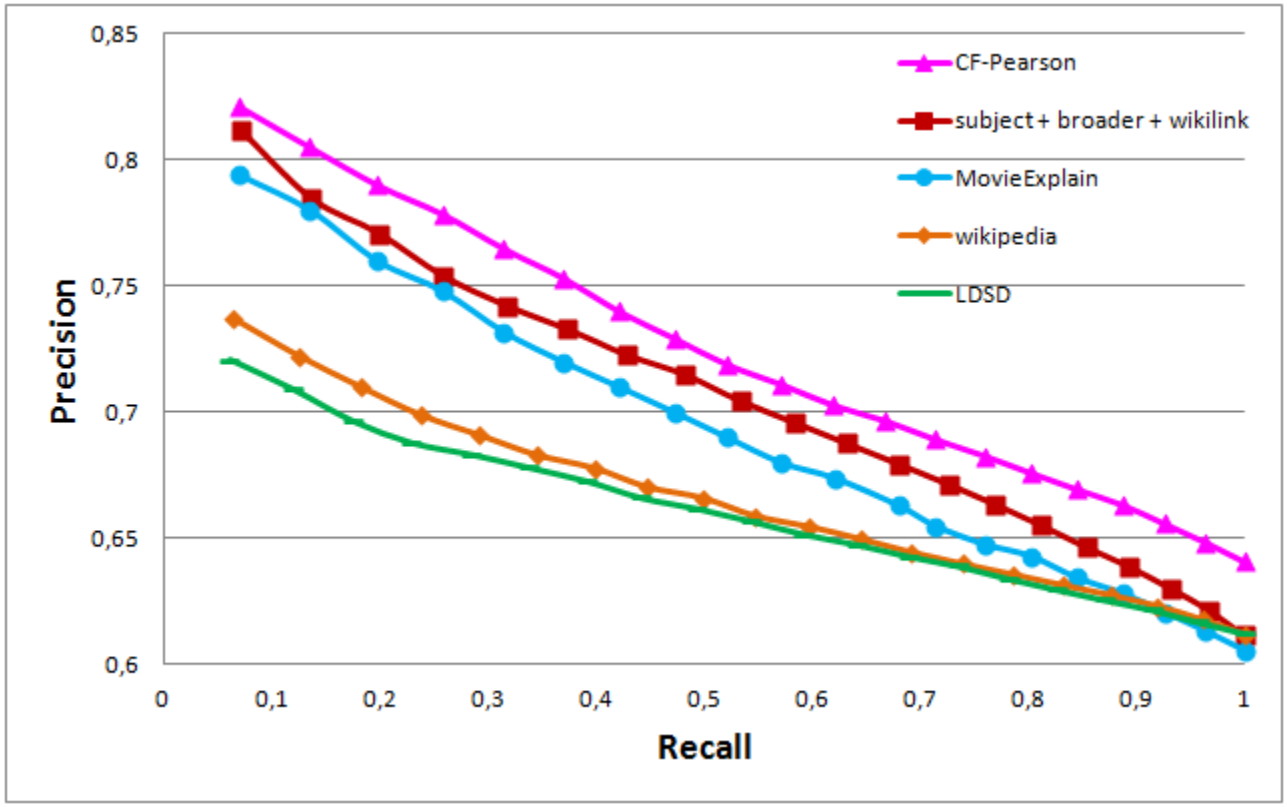
Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD & Recommender Systems



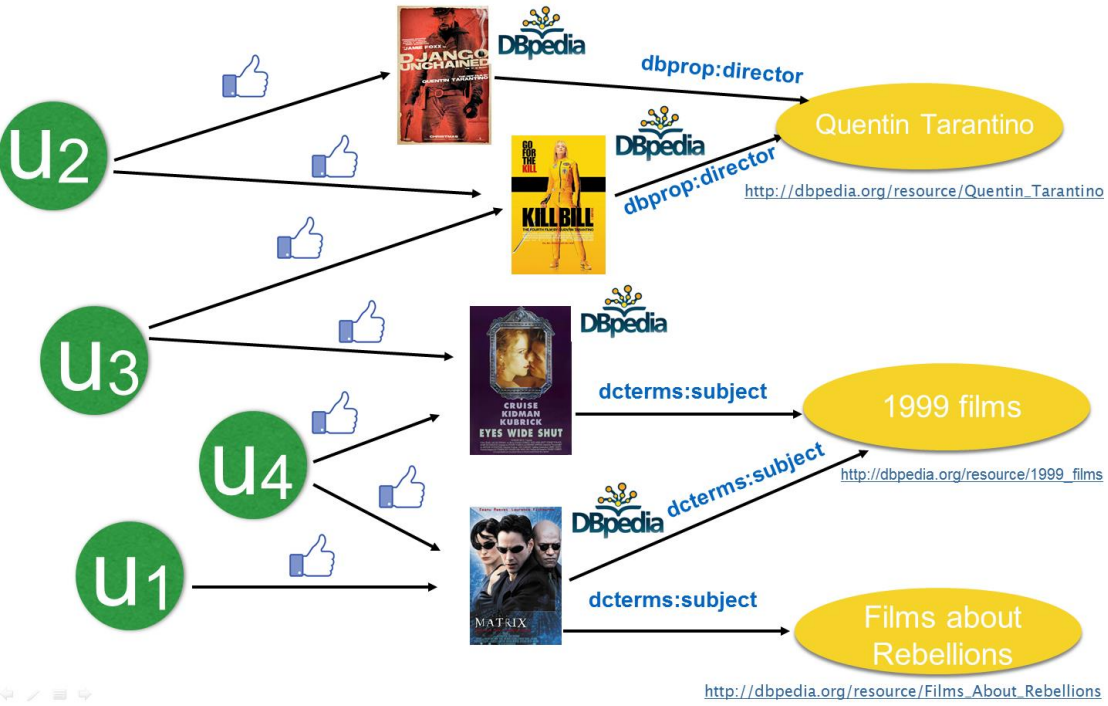
Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

LOD & Recommender Systems



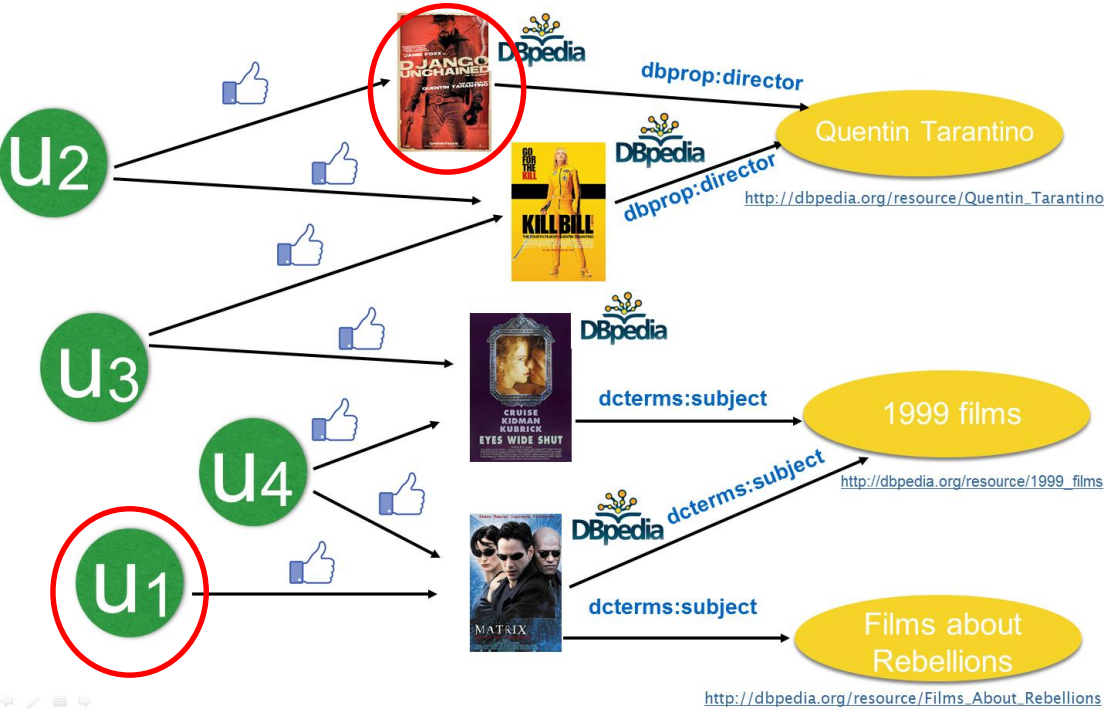
Tommaso Di Noia, Roberto Mirizzi, Vito Claudio Ostuni, Davide Romito, Markus Zanker. **Linked Open Data to support Content-based Recommender Systems.** *8th International Conference on Semantic Systems (I-SEMANTICS) - 2012 (Best Paper Award)*

Graph-based RecSys



Recommendations obtained by mining the graph

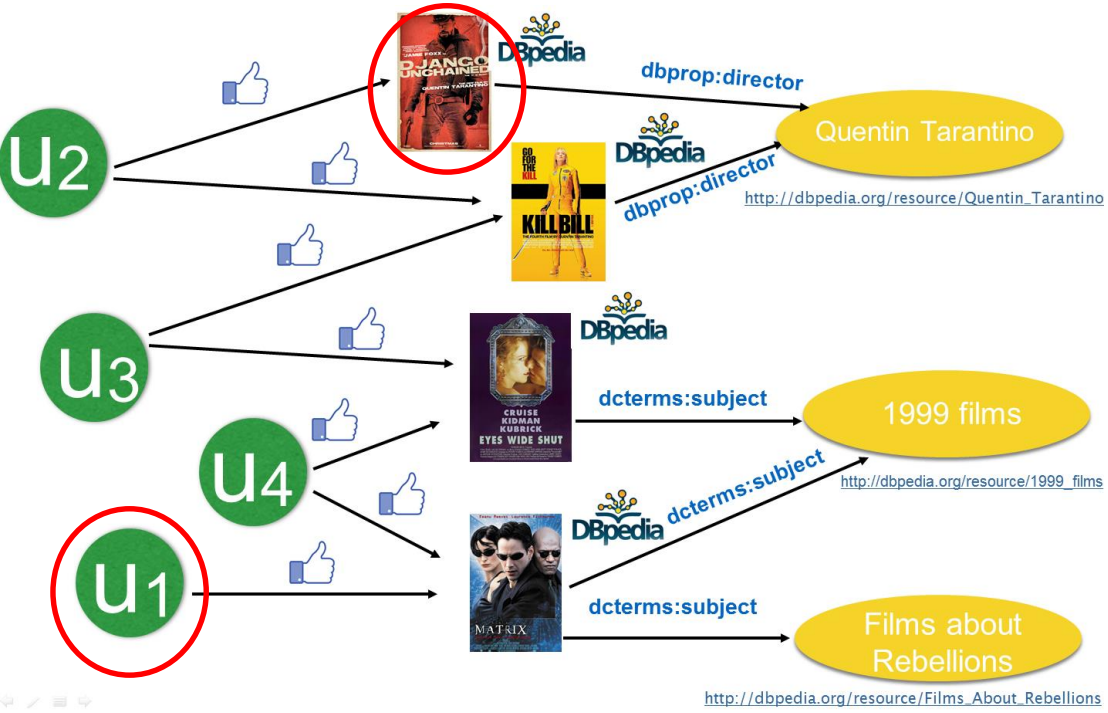
Graph-based RecSys



Recommendations obtained by mining the graph

Identification of the most relevant (target) nodes, according to the recommendation scenario

Graph-based RecSys



Recommendations obtained by mining the graph

Identification of the most relevant (target) nodes, according to the recommendation scenario

PageRank
Spreading Activation
Personalized PageRank
…

Cataldo Musto, Pasquale Lops, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro. **Semantics-aware Graph-based Recommender Systems Exploiting Linked Open Data**. UMAP 2016

Phuong Nguyen, Paolo Tomeo, Tommaso Di Noia, Eugenio Di Sciascio. **An evaluation of SimRank and Personalized PageRank to build a recommender system for the Web of Data**. 7th International Workshop on Web Intelligence & Communities @ WWW 2015

Personalized PageRank

A node receives an amount of rank from every node which points to it and in turn transfers an amount of its rank to the nodes it refers to.

The similarity between two items α and β represented by vectors $\alpha = \{a_i\}$ i=1,..,n and $\beta = \{b_i\}$ i=1,..,n is computed as the inner product space between the two vectors

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

T. H. Haveliwala. Topic-Sensitive PageRank: **A Context-Sensitive Ranking Algorithm for Web Search.** IEEE Trans. Knowl. Data Eng., 15(4):784–796, 2003.

SimRank

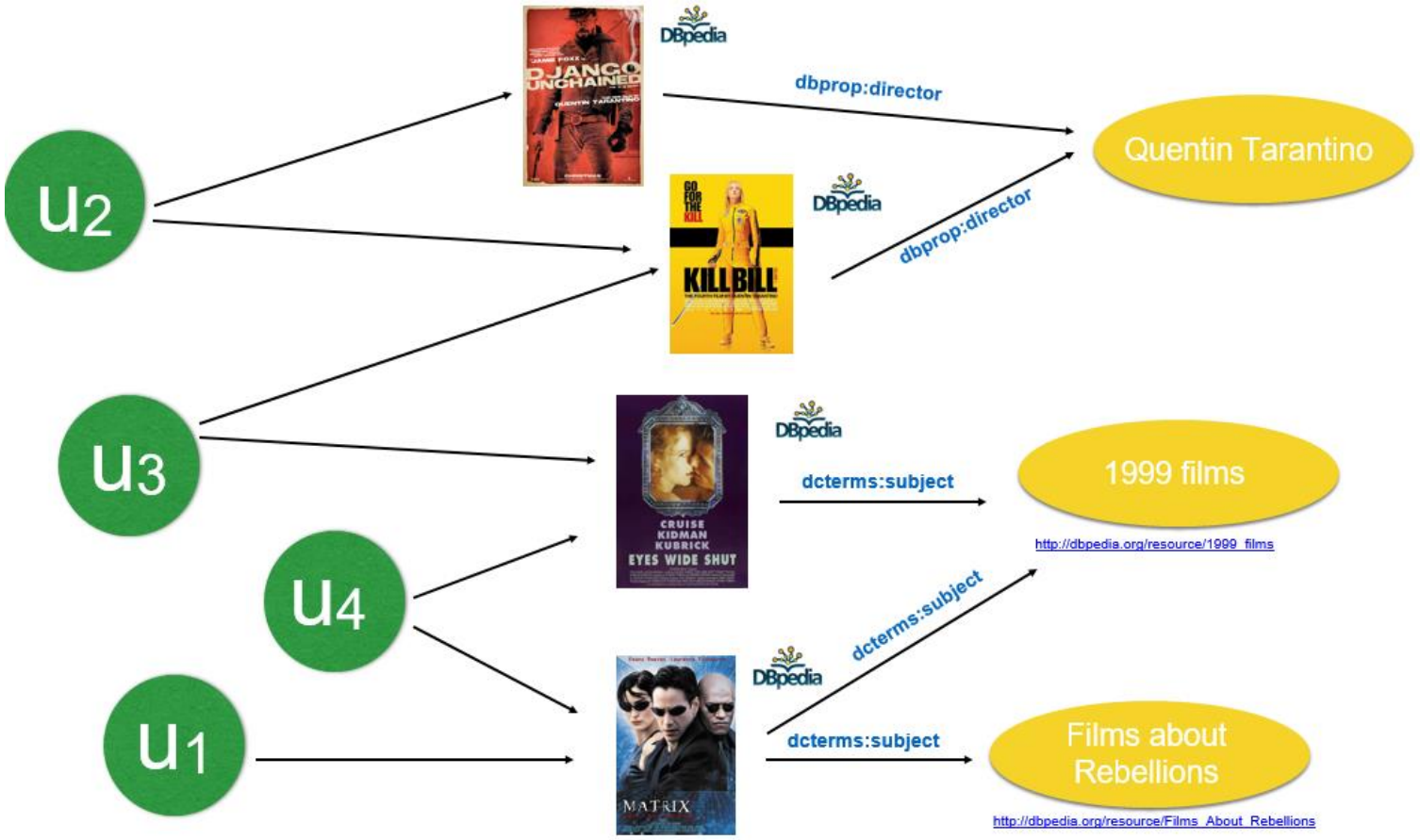
SimRank computes similarity between nodes in a graph using the structural context: two nodes are similar if they are referenced by similar nodes.

Given $k \geq 0$, $R^{(k)}(\alpha, \beta) = 1$ with $\alpha = \beta$. $R^{(k)}(\alpha, \beta) = 0$ with $k = 0$ and $\alpha = \beta$.
Otherwise, the general formula is

$$R^{(k+1)}(\alpha, \beta) = \frac{d}{|I(\alpha)| \cdot |I(\beta)|} \sum_{i=1}^{|I(\alpha)|} \sum_{j=1}^{|I(\beta)|} R^{(k)}(I_i(\alpha), I_j(\beta))$$

G. Jeh and J. Widom. Simrank: **A measure of structural-context similarity**. In ACM KDD '02, pages 538–543, 2002.

Graph-based RecSys

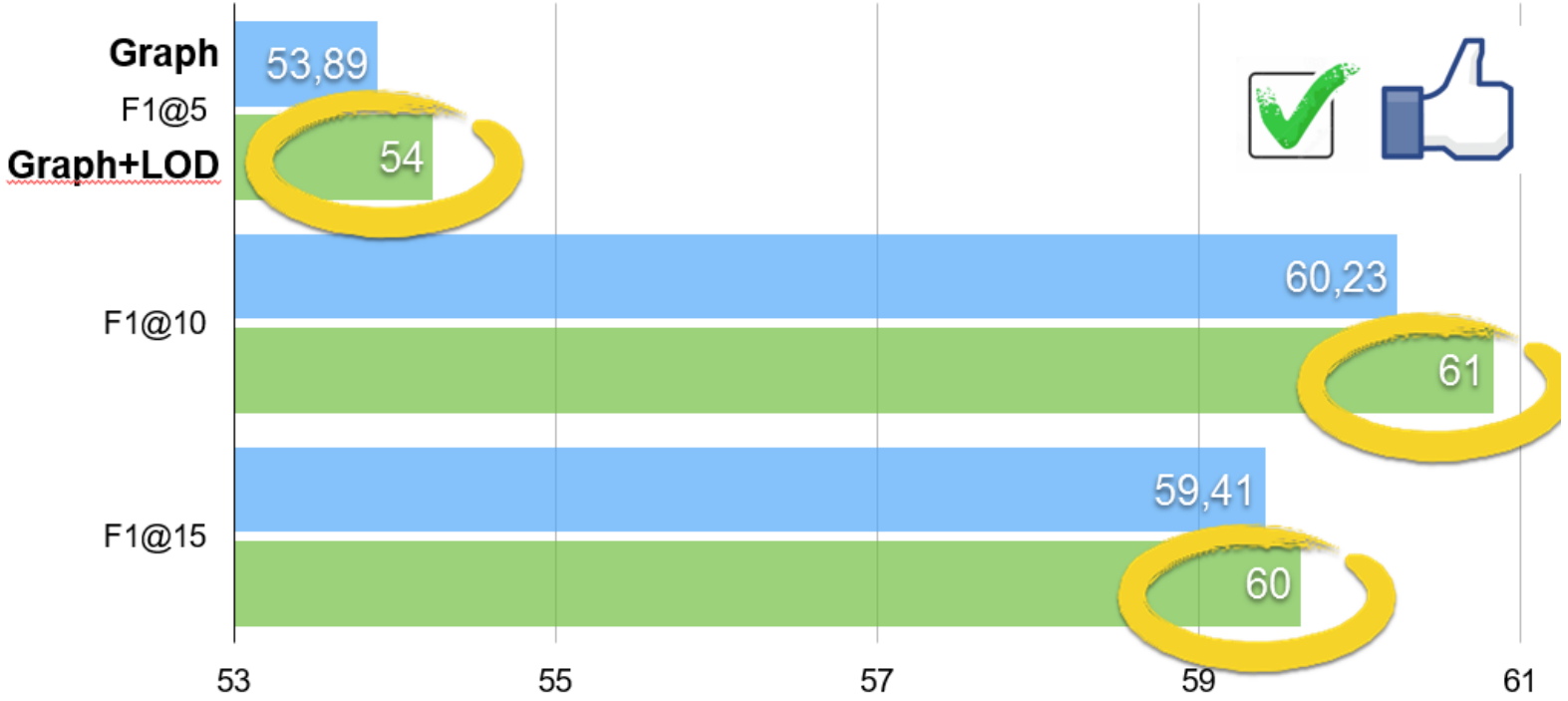


Personalized PageRank to identify the most relevant nodes in the graph

Graph-based RecSys

MovieLens 100K dataset

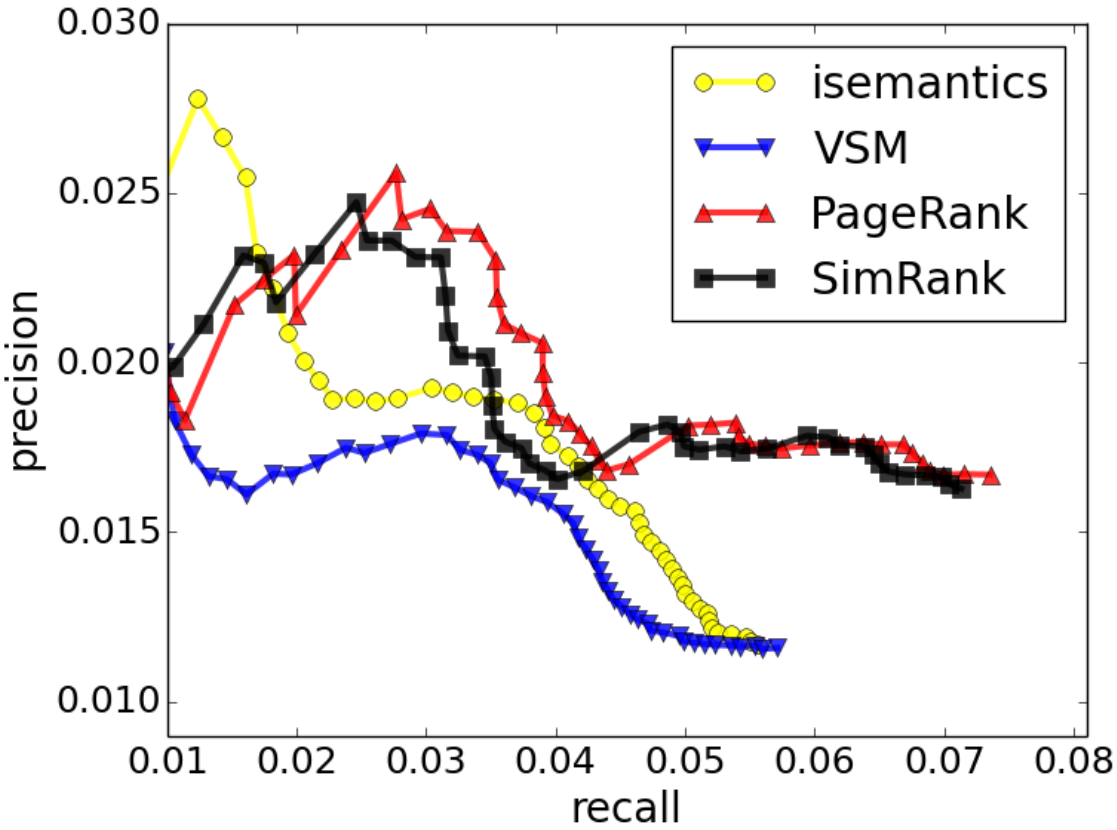
Graph = Personalized PageRank on Bipartite User-Item Graph
Graph+LOD = Tripartite Graph modeling also Linked Open Data



Significant improvement in all the metrics (Wilcoxon test)

Cataldo Musto, Pasquale Lops, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro. **Semantics-aware Graph-based Recommender Systems Exploiting Linked Open Data**. UMAP 2016

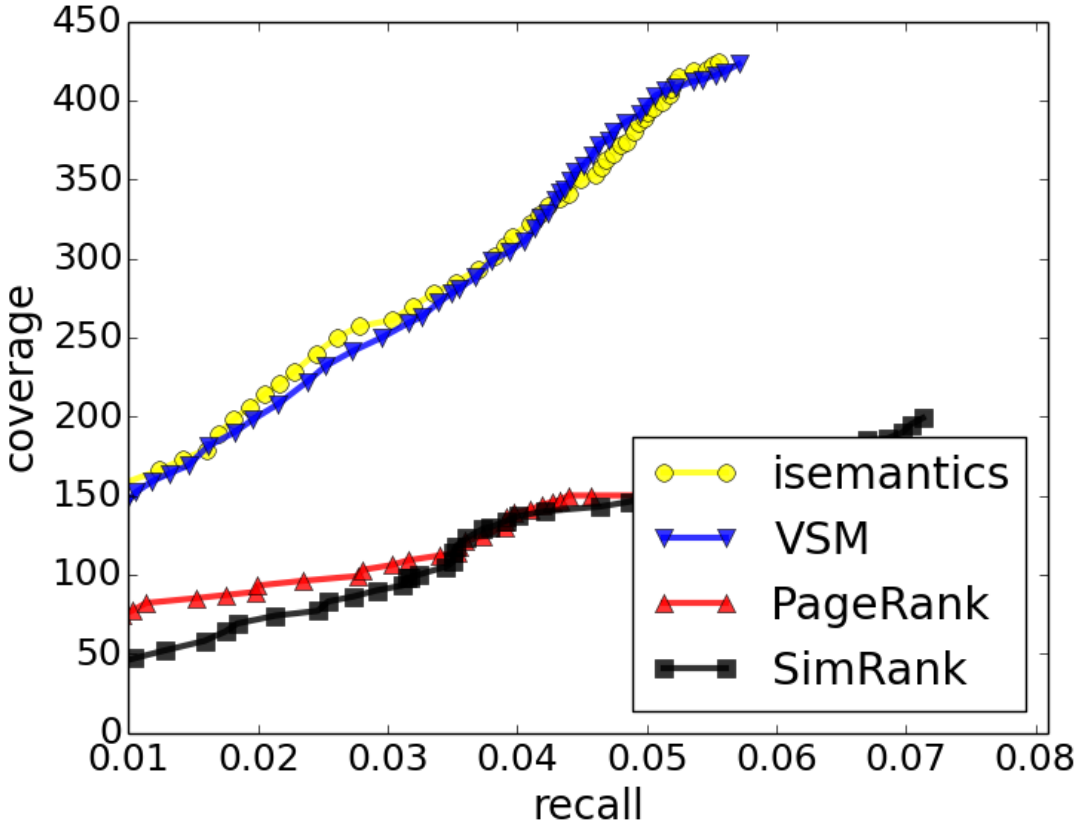
Accuracy with 40 neighbors



PageRank and SimRank more accurate

Phuong Nguyen, Paolo Tomeo, Tommaso Di Noia, Eugenio Di Sciascio. **An evaluation of SimRank and Personalized PageRank to build a recommender system for the Web of Data**. 7th International Workshop on Web Intelligence & Communities @ WWW 2015

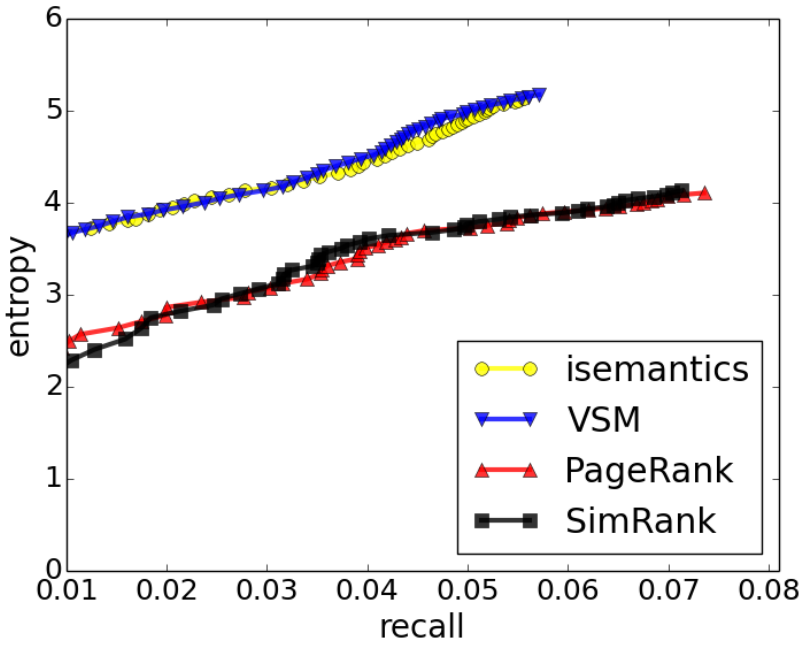
Catalog coverage with 40 neighbors

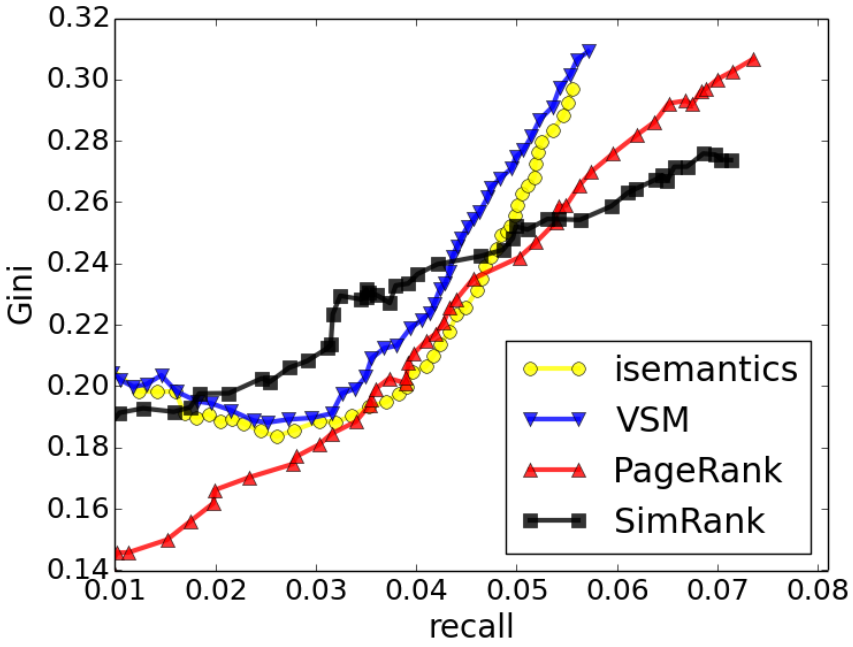


Worse coverage

Phuong Nguyen, Paolo Tomeo, Tommaso Di Noia, Eugenio Di Sciascio. **An evaluation of SimRank and Personalized PageRank to build a recommender system for the Web of Data**. 7th International Workshop on Web Intelligence & Communities @ WWW 2015

Distribution with 40 neighbors





Recommendations concentred on a few items

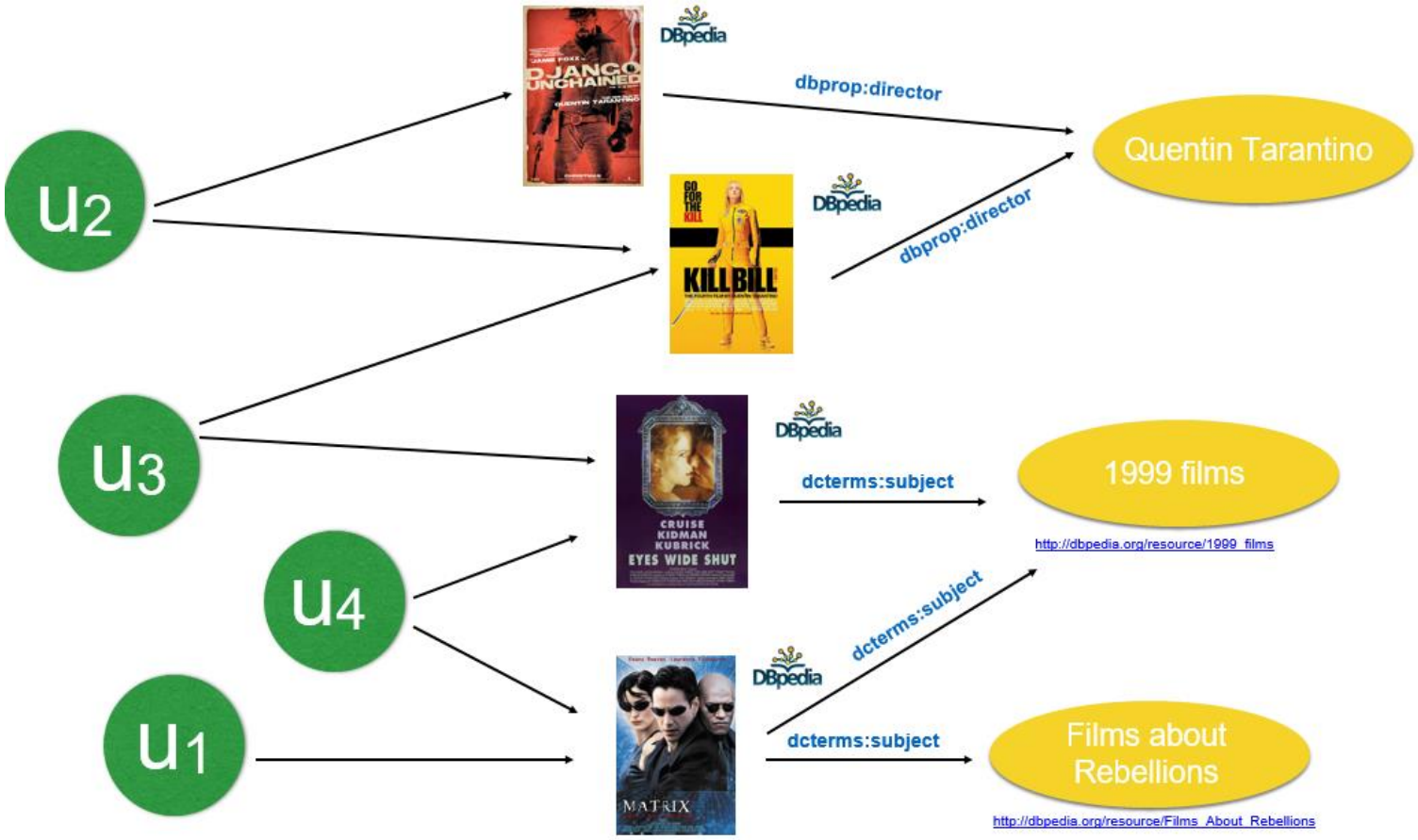
Phuong Nguyen, Paolo Tomeo, Tommaso Di Noia, Eugenio Di Sciascio. **An evaluation of SimRank and Personalized PageRank to build a recommender system for the Web of Data**. 7th International Workshop on Web Intelligence & Communities @ WWW 2015

Novelty Results

	EPC@20	Long-tail%@20	EPC@30	Long-tail%@30	EPC@40	Long-tail%@40	EPC@50	Long-tail%@50
isemantics	0.844	0.648	0.841	0.657	0.839	0.658	0.836	0.643
VSM	0.848	0.657	0.847	0.675	0.848	0.673	0.843	0.658
PageRank	**0.913**	**0.823**	**0.893**	**0.767**	0.867	**0.743**	0.860	0.710
SimRank	0.906	0.793	0.881	0.747	**0.874**	0.742	**0.867**	**0.717**

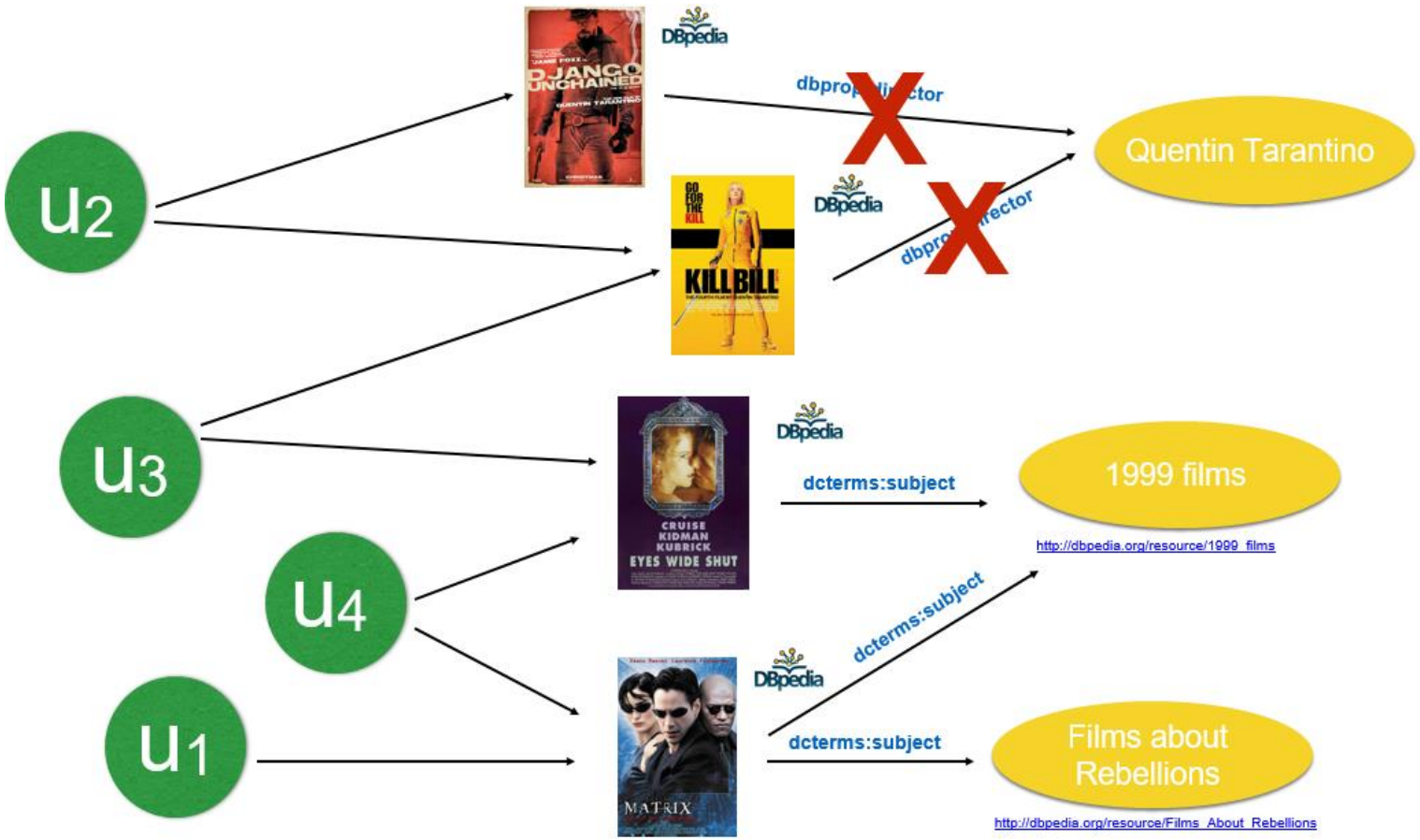
Phuong Nguyen, Paolo Tomeo, Tommaso Di Noia, Eugenio Di Sciascio. **An evaluation of SimRank and Personalized PageRank to build a recommender system for the Web of Data**. 7th International Workshop on Web Intelligence & Communities @ WWW 2015

Graph-based RecSys



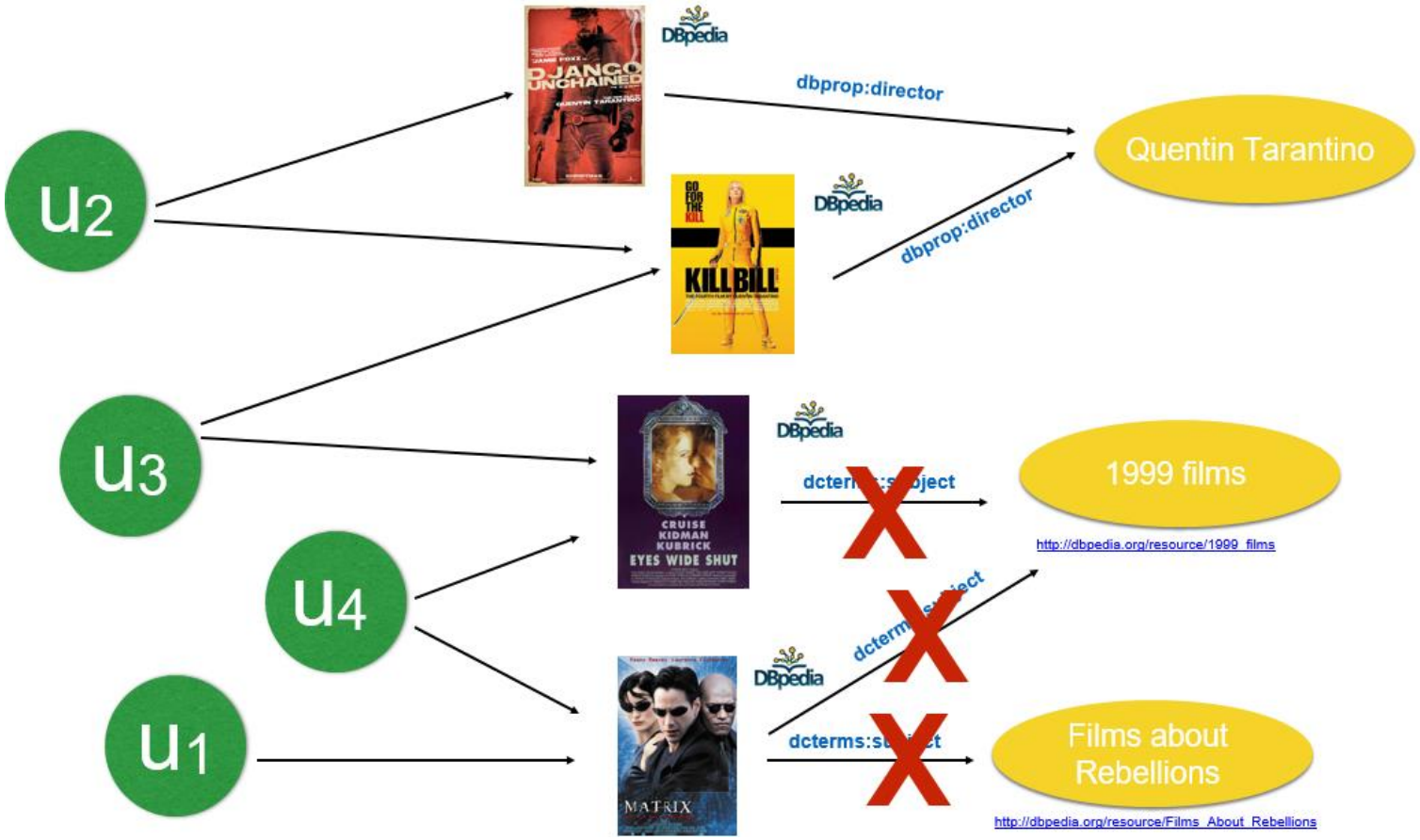
is it necessary to inject all the properties available in LOD cloud?

Graph-based RecSys



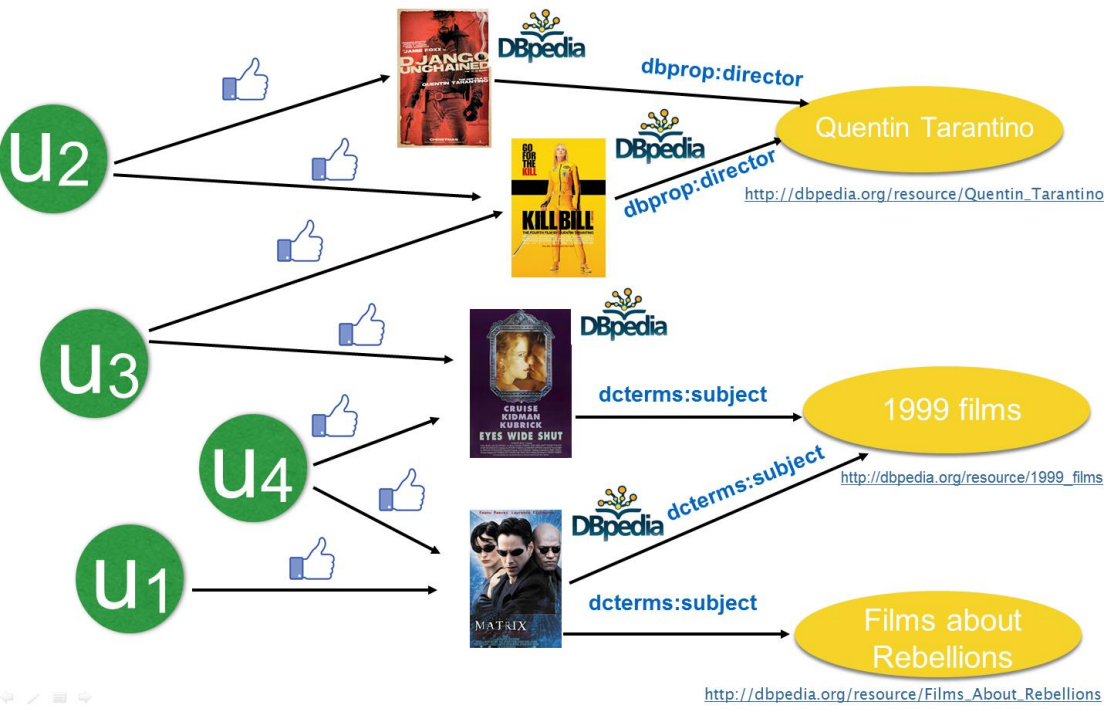
is it necessary to inject all the properties available in LOD cloud?

Graph-based RecSys



is it necessary to inject all the properties available in LOD cloud?

Graph-based RecSys



what are the most promising properties to include?

manual selection

- domain-specific properties
- most frequent properties
- …

automatic selection

- **more difficult** to implement

Cataldo Musto, Pasquale Lops, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro. **Semantics-aware Graph-based Recommender Systems Exploiting Linked Open Data**. UMAP 2016

Azzurra Ragone, Paolo Tomeo, Corrado Magarelli, Tommaso Di Noia, Matteo Palmonari, Andrea Maurino, and Eugenio Di Sciascio. **Schema-summarization in linked-data-based feature selection for recommender systems**. SAC 2017

Feature selection

selecting the **most promising subset** of LOD-based properties

possible techniques

Statistical

- PageRank
- Principal Component Analysis
- Information Gain
- Information Gain Ratio
- Mininum Redundancy Maximum Relevance

Cataldo Musto, Pasquale Lops, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro. **Semantics-aware Graph-based Recommender Systems Exploiting Linked Open Data**. UMAP 2016

Schema-aware

- Data-driven Schema Summarization

Azzurra Ragone, Paolo Tomeo, Corrado Magarelli, Tommaso Di Noia, Matteo Palmonari, Andrea Maurino, and Eugenio Di Sciascio. **Schema-summarization in linked-data-based feature selection for recommender systems**. SAC 2017

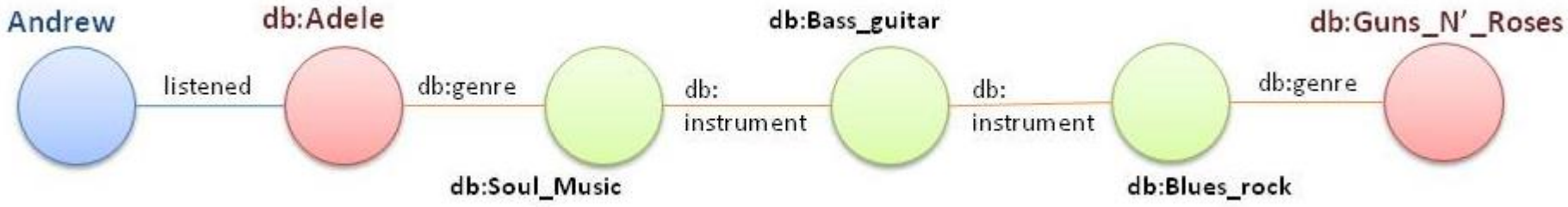
LOD & Recommender Systems

path-based features

analysis of **complex relations** between **user preferences** and the **target item**

extraction of **path-based features**

SPrank system



Tommaso Di Noia, Vito Claudio Ostuni, Paolo Tomeo, Eugenio Di Sciascio: **SPrank: Semantic Path-Based Ranking for Top-N Recommendations Using Linked Open Data**. ACM TIST 8(1): 9:1-9:34 (2016)

LOD & Recommender Systems

path-based features

Implicit Feedback Matrix

$\hat{S} =$

	I_1	i_2	i_3	i_4
u1	1	1	0	0
u2	1	0	1	0
u3	0	1	1	0
u4	0	1	0	1

Knowledge Graph

LOD & Recommender Systems

path-based features

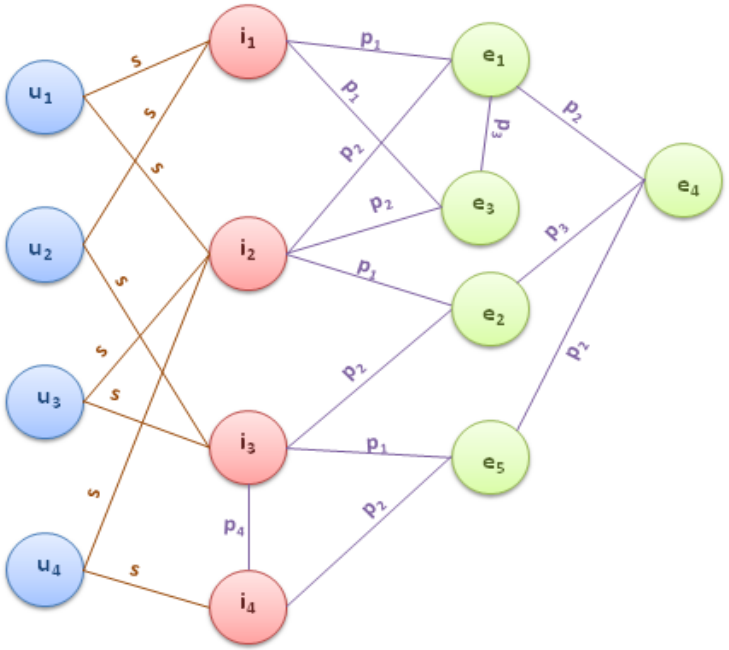
Implicit Feedback Matrix

$\hat{S} =$

	I_1	i_2	i_3	i_4
u1	1	1	0	0
u2	1	0	1	0
u3	0	1	1	0
u4	0	1	0	1

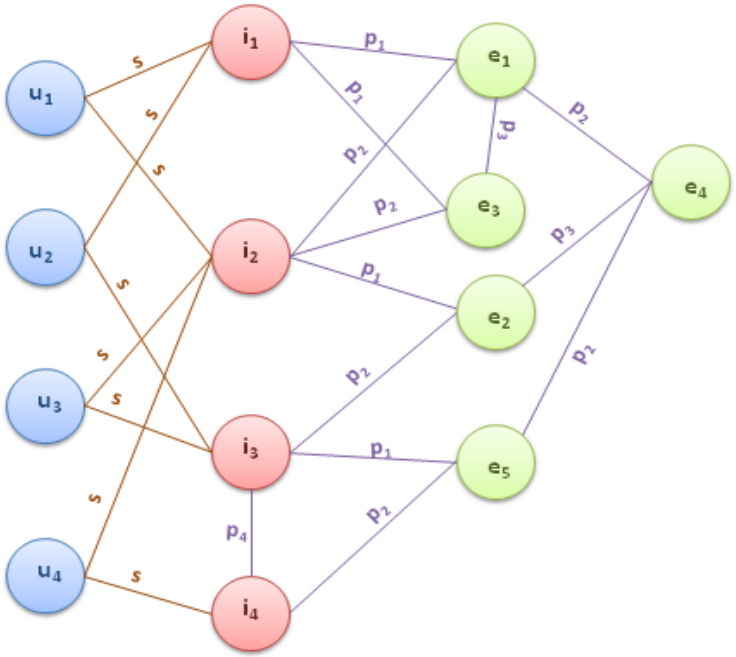
Knowledge Graph

DBpedia



Semantic Graph-based Data Model
(Path Based Features)

we can further build **some extra features** by **mining the paths** occurring in the graph



path acyclic sequence of relations (s , .. r_l , .. r_L)

$u_3\ s\ i_2\ p_2\ e_1\ p_1\ i_1 \rightarrow (s, p_2, p_1)$

- The more the paths, the more the relevance of the item.
- Different paths have different meaning.
- Not all types of paths are relevant.

$$w_{ux}(j) = \frac{\#path_{ux}(j)}{\sum_j \#path_{ux}(j)}$$

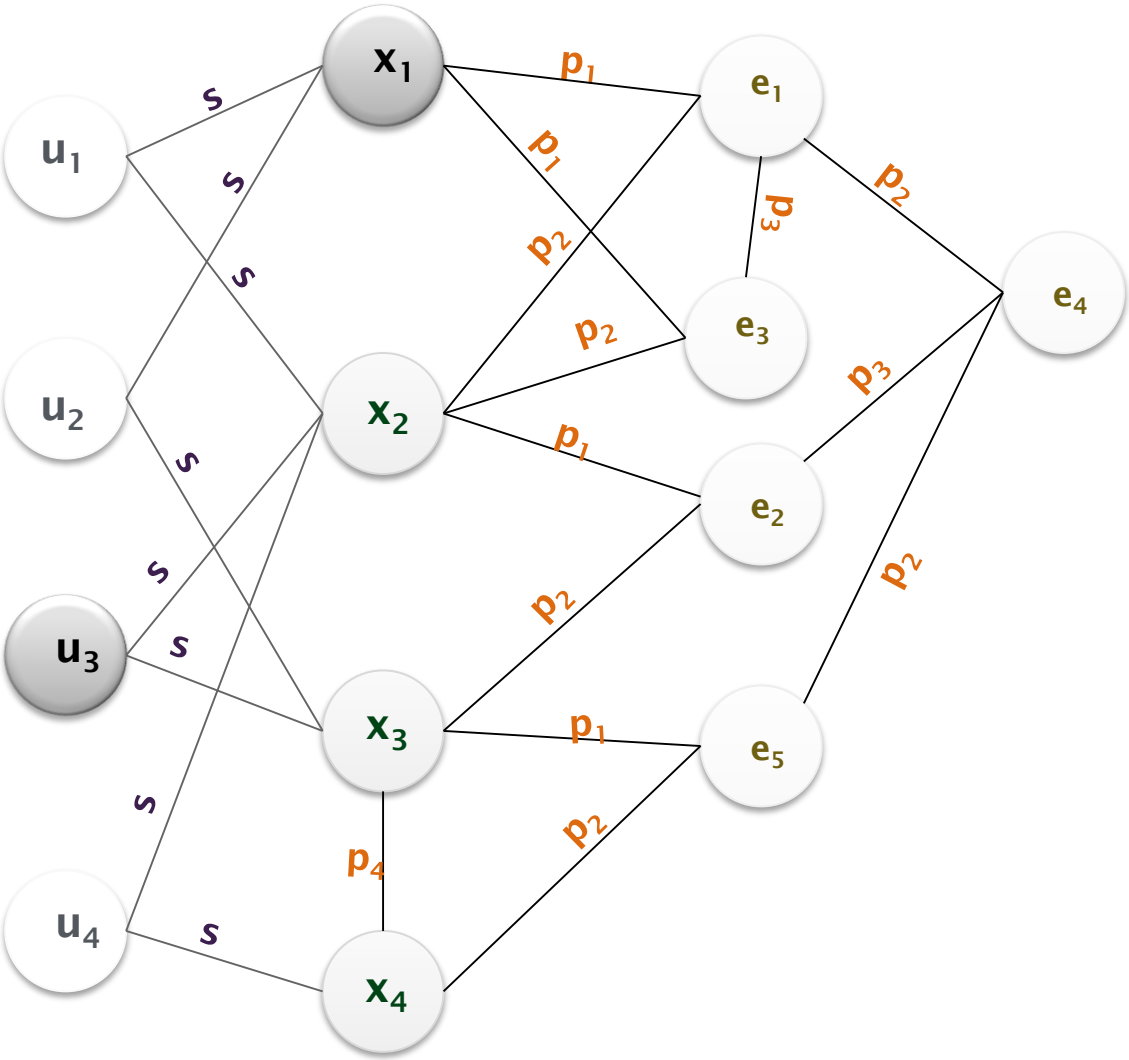
frequency of $path_j$ in the sub-graph related to u ad x

Tommaso Di Noia, Vito Claudio Ostuni, Paolo Tomeo, Eugenio Di Sciascio: **SPrank: Semantic Path-Based Ranking for Top-N Recommendations Using Linked Open Data**. ACM TIST 8(1): 9:1-9:34 (2016)

Semantic Graph-based Data Model

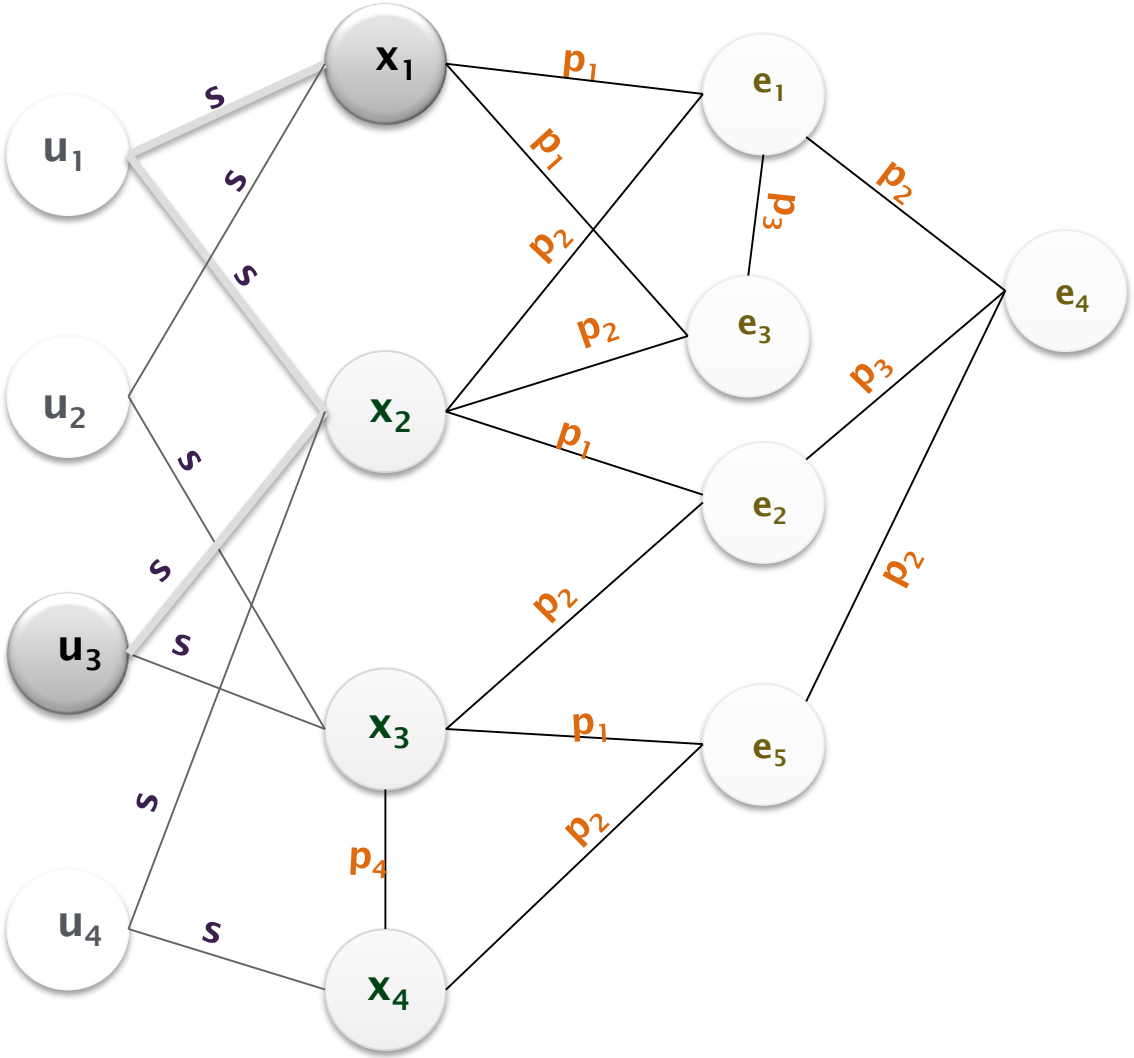
(Path Based Features)

$w_{u_3 x_1}$?



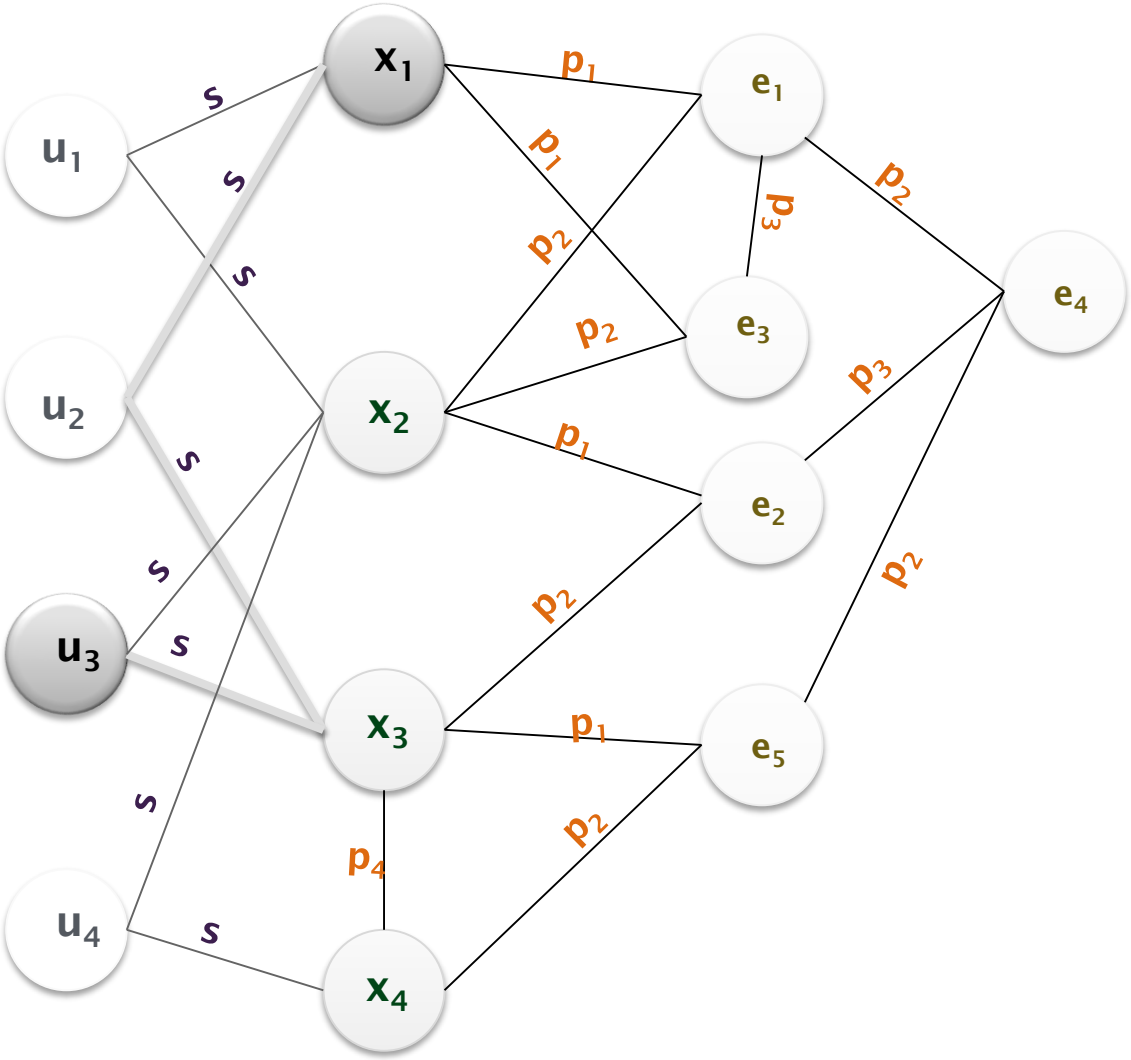
Semantic Graph-based Data Model
(Path Based Features)

$path_1$ $(s, s, s) : 1$



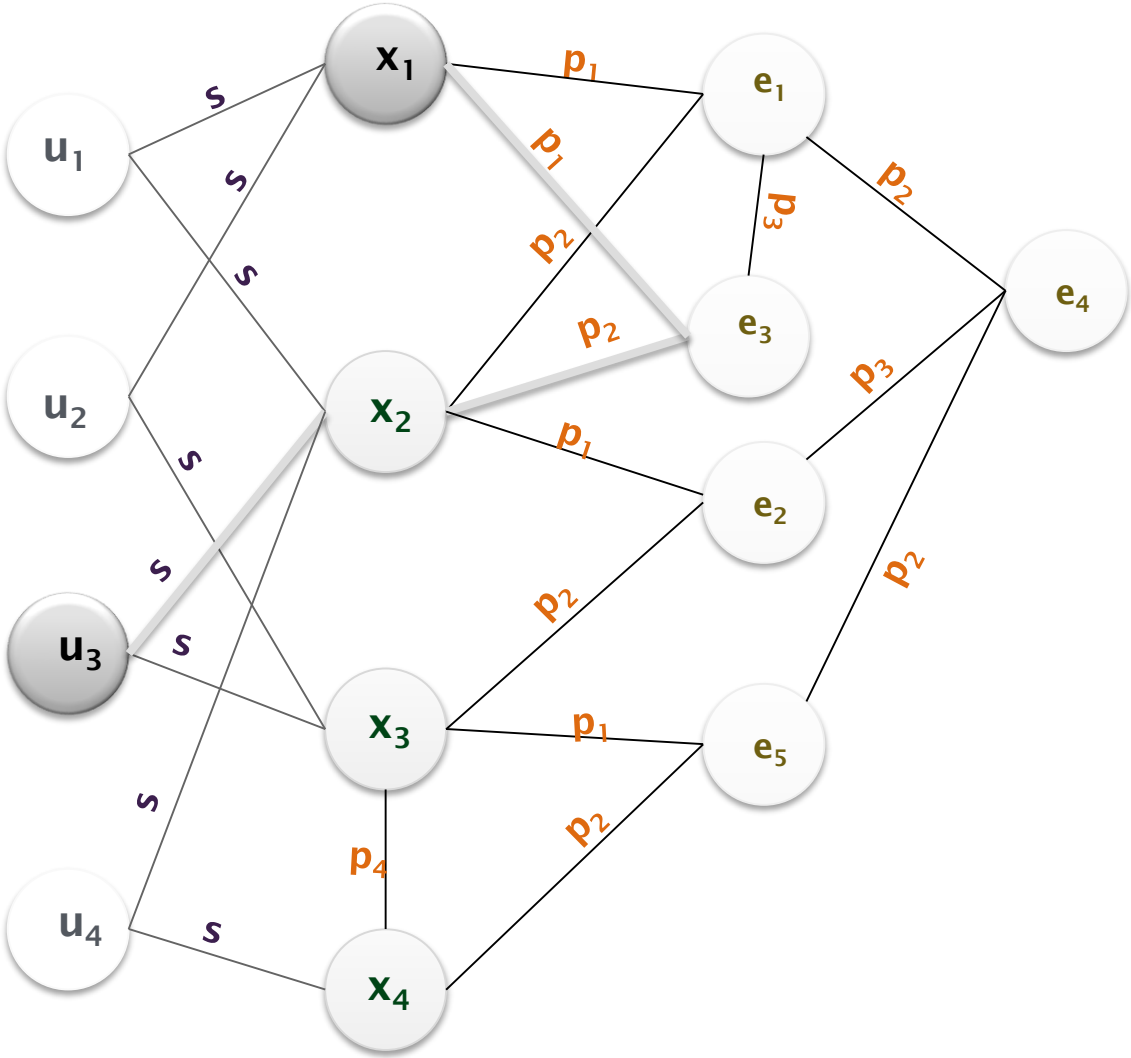
Semantic Graph-based Data Model
(Path Based Features)

$path_1$ (s, s, s) : 2



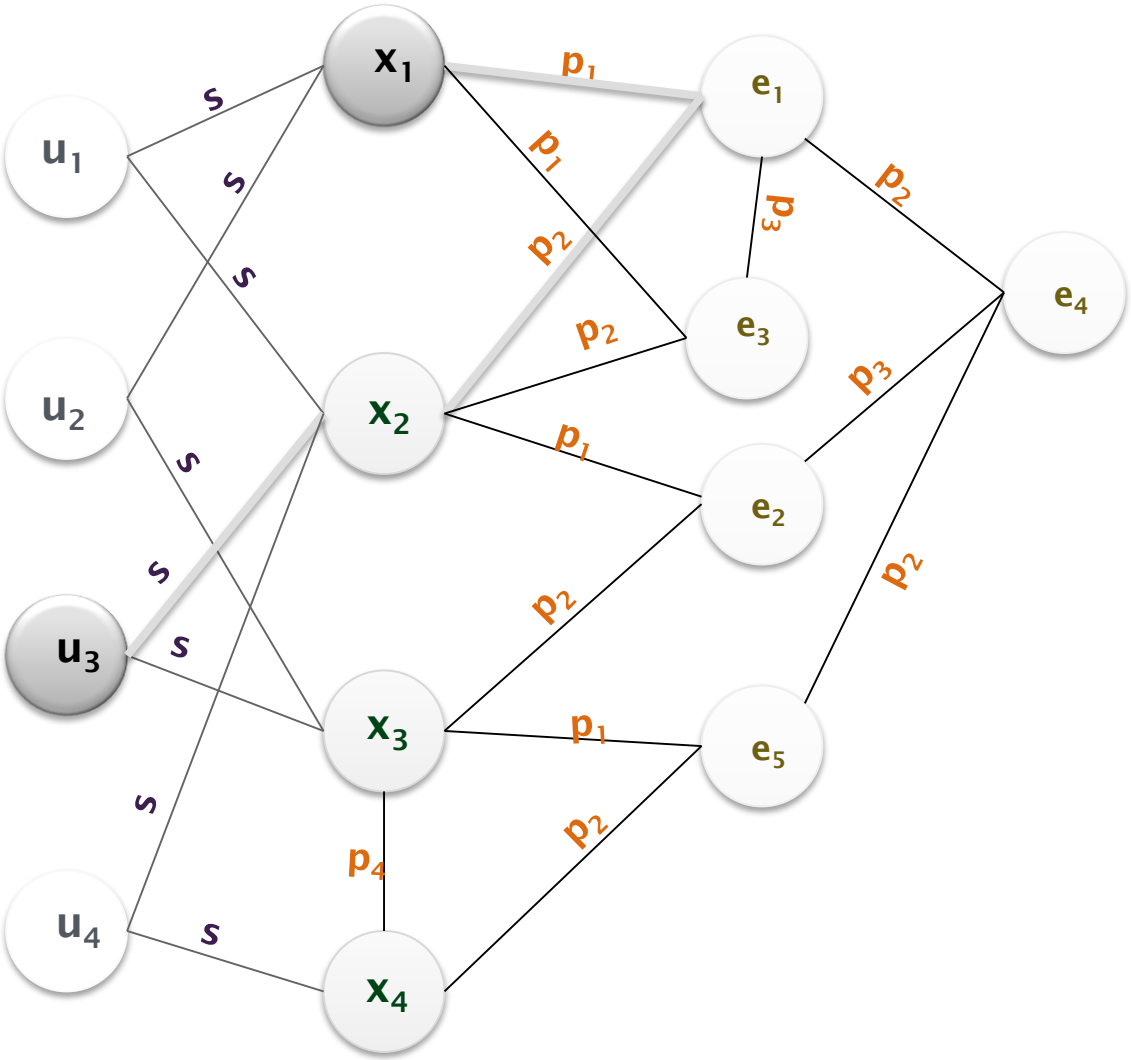
Semantic Graph-based Data Model
(Path Based Features)

$path_1$ $(s, s, s) : 2$
$path_2$ $(s, p_2, p_1) : 1$



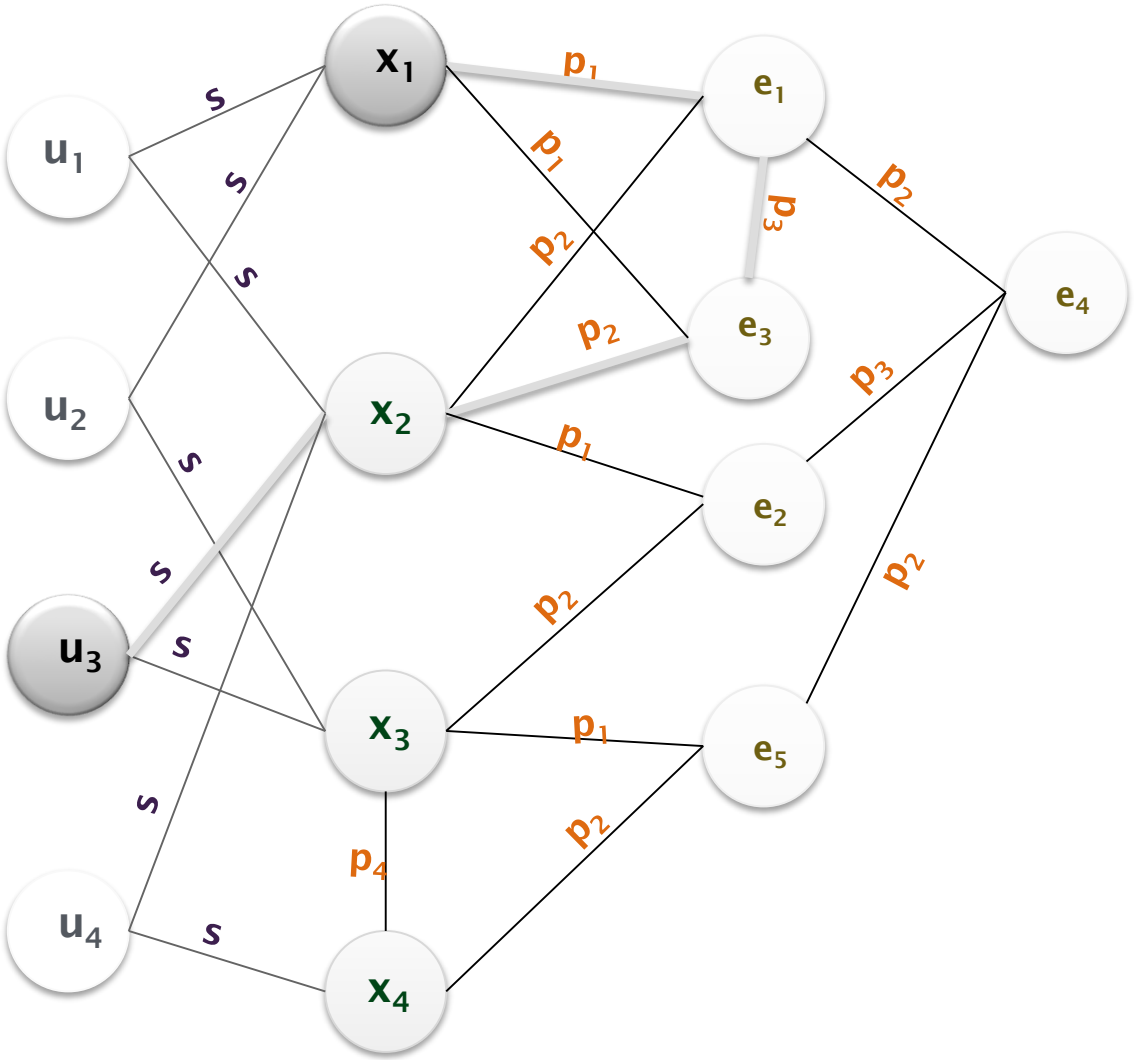
Semantic Graph-based Data Model
(Path Based Features)

$path_1$ $(s, s, s) : 2$
$path_2$ $(s, p_2, p_1) : 2$



Semantic Graph-based Data Model
(Path Based Features)

$path_1$ $(s, s, s) : 2$
$path_2$ $(s, p_2, p_1) : 2$
$path_3$ $(s, p_2, p_3, p_1) : 1$



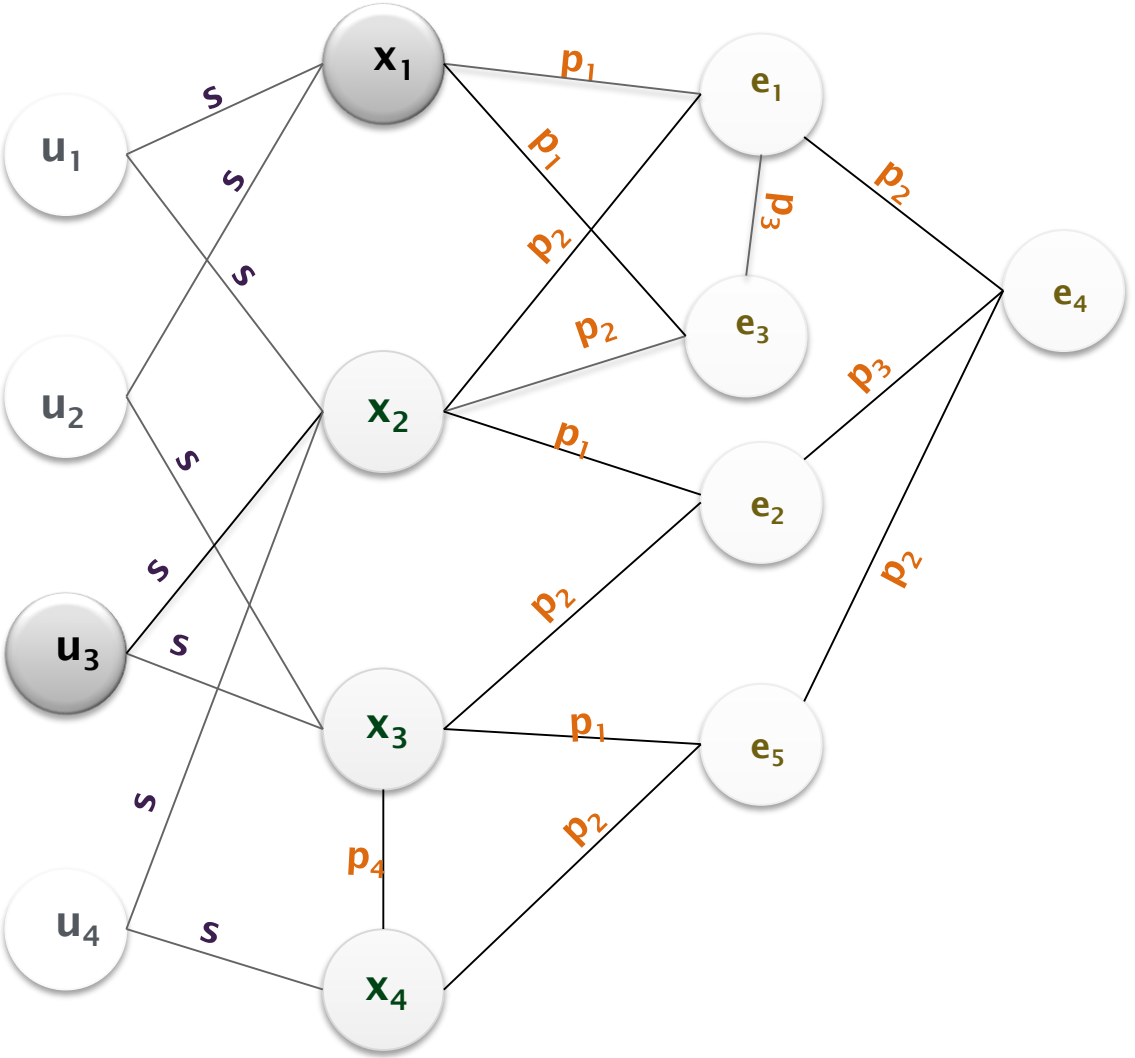
Semantic Graph-based Data Model
(Path Based Features)

path$_1$ (s, s, s) : 2
path$_2$ (s, p_2, p_1) : 2
path$_3$ (s, p_2, p_3, p_1) : 1

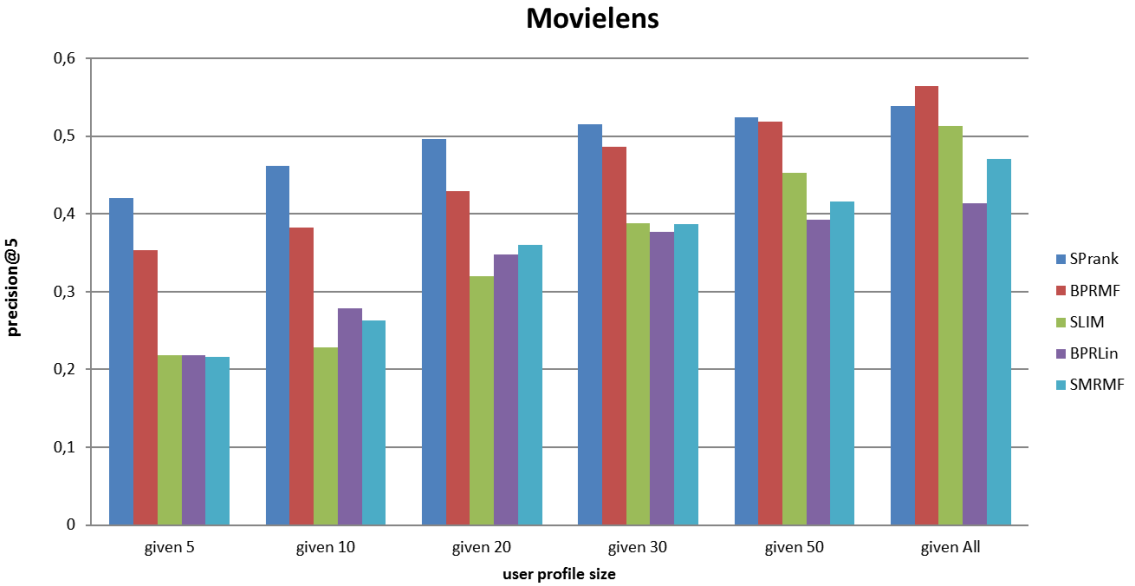
$$w_{u_3x_1}(1) = \frac{2}{5}$$

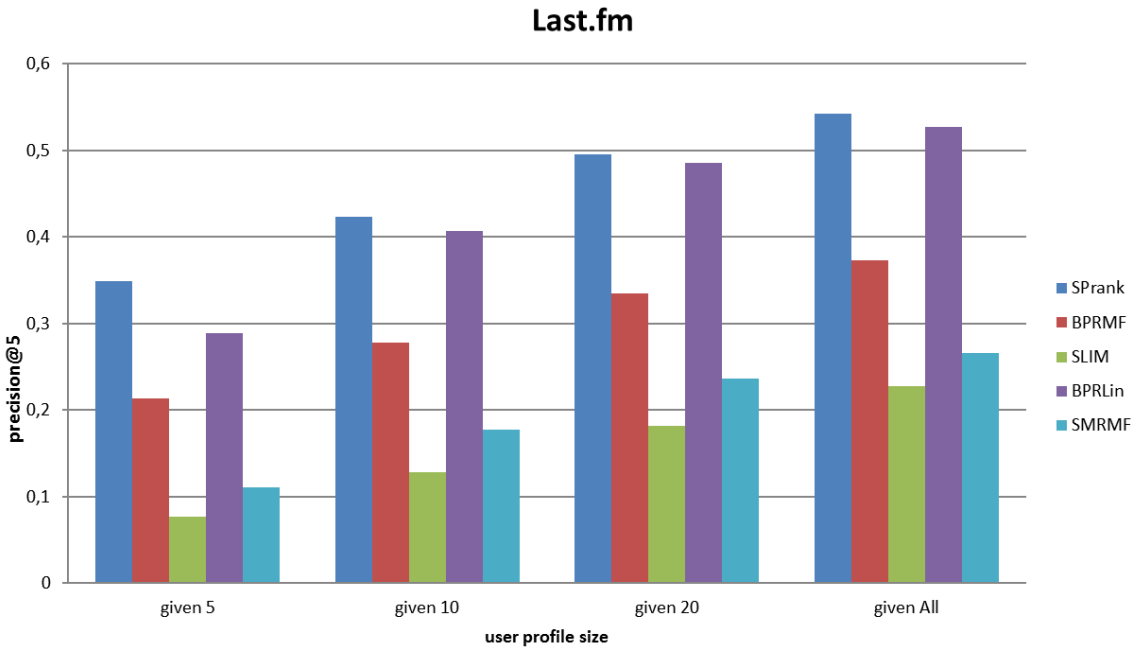
$$w_{u_3x_1}(2) = \frac{2}{5}$$

$$w_{u_3x_1}(3) = \frac{1}{5}$$



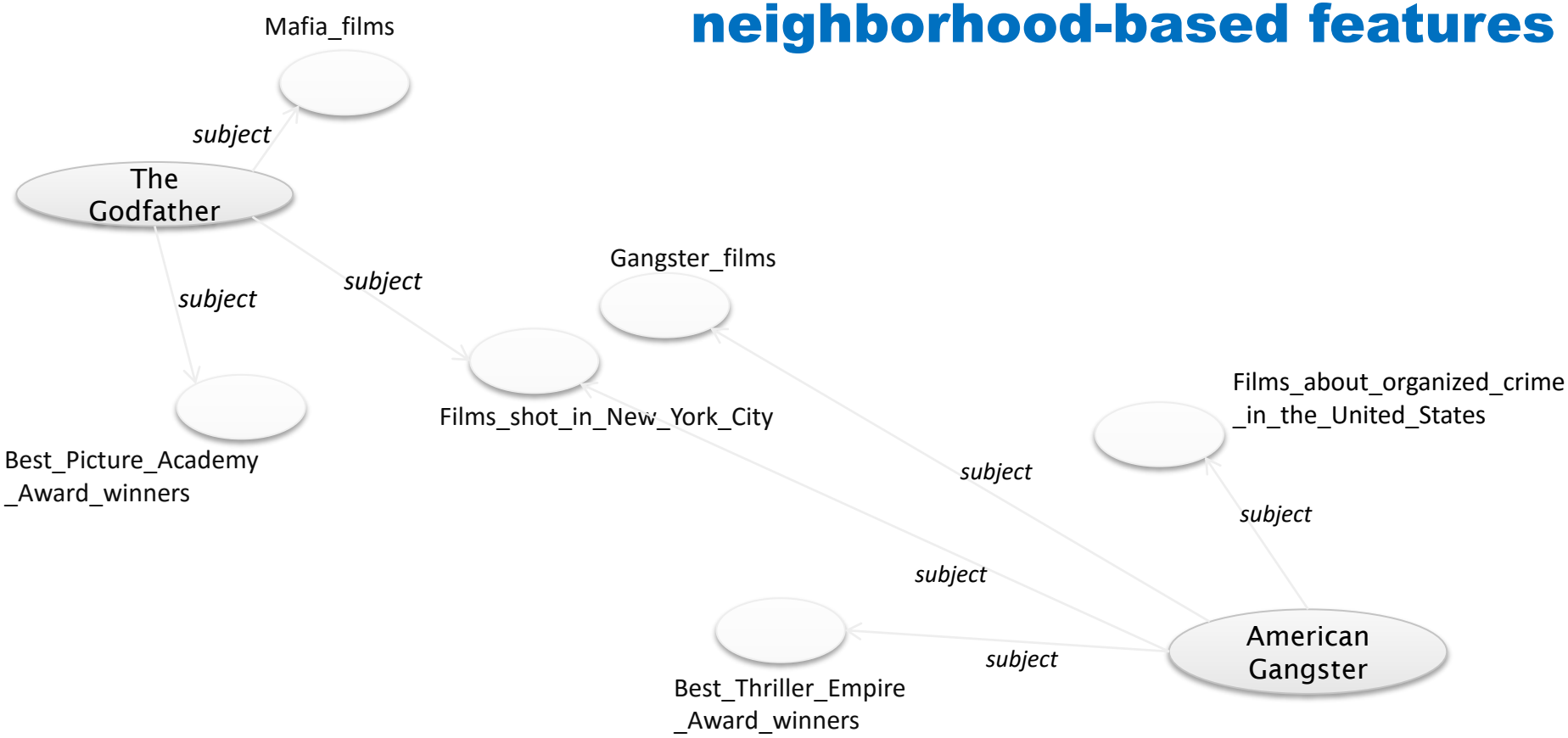
Evaluation of different ranking functions





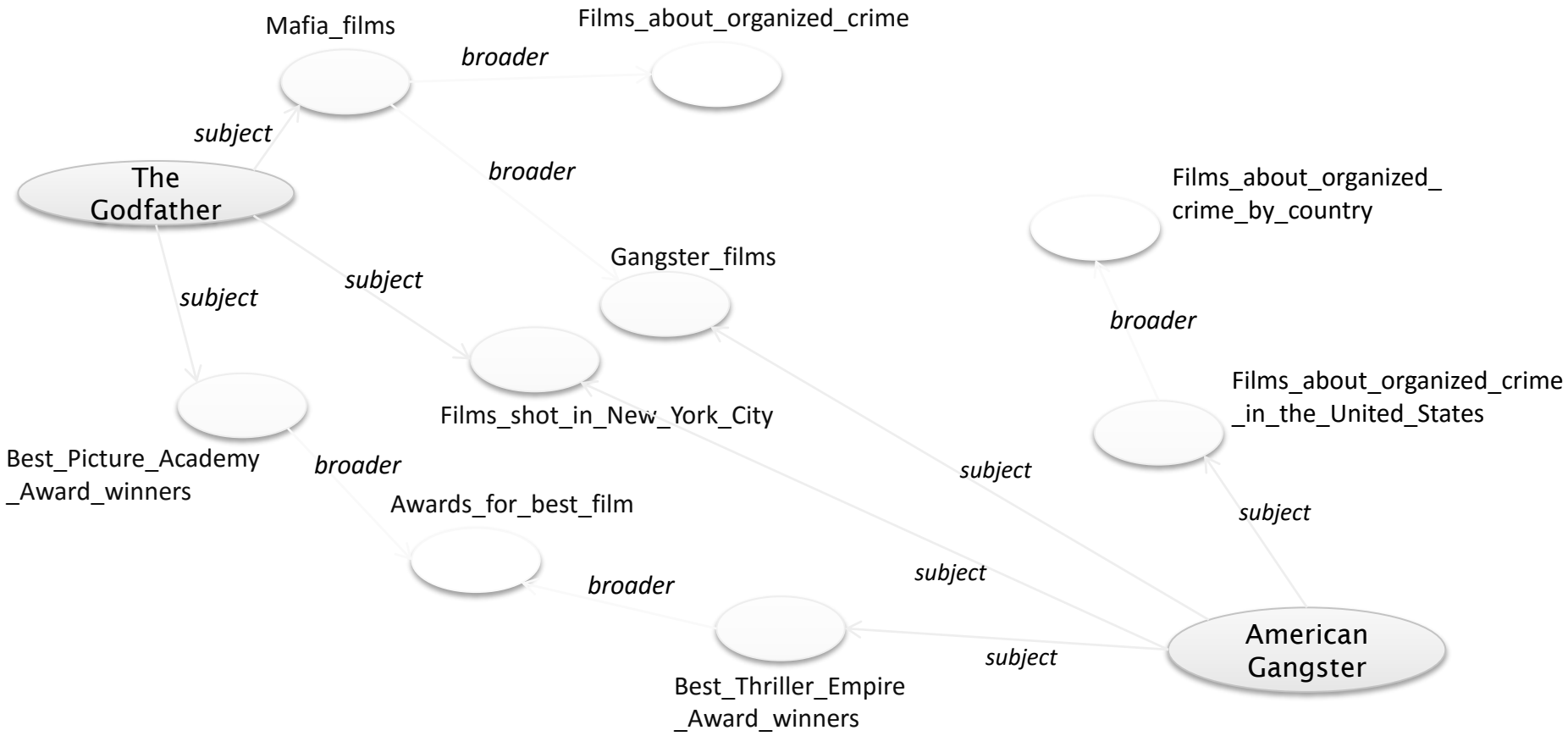
LOD & Recommender Systems

neighborhood-based features



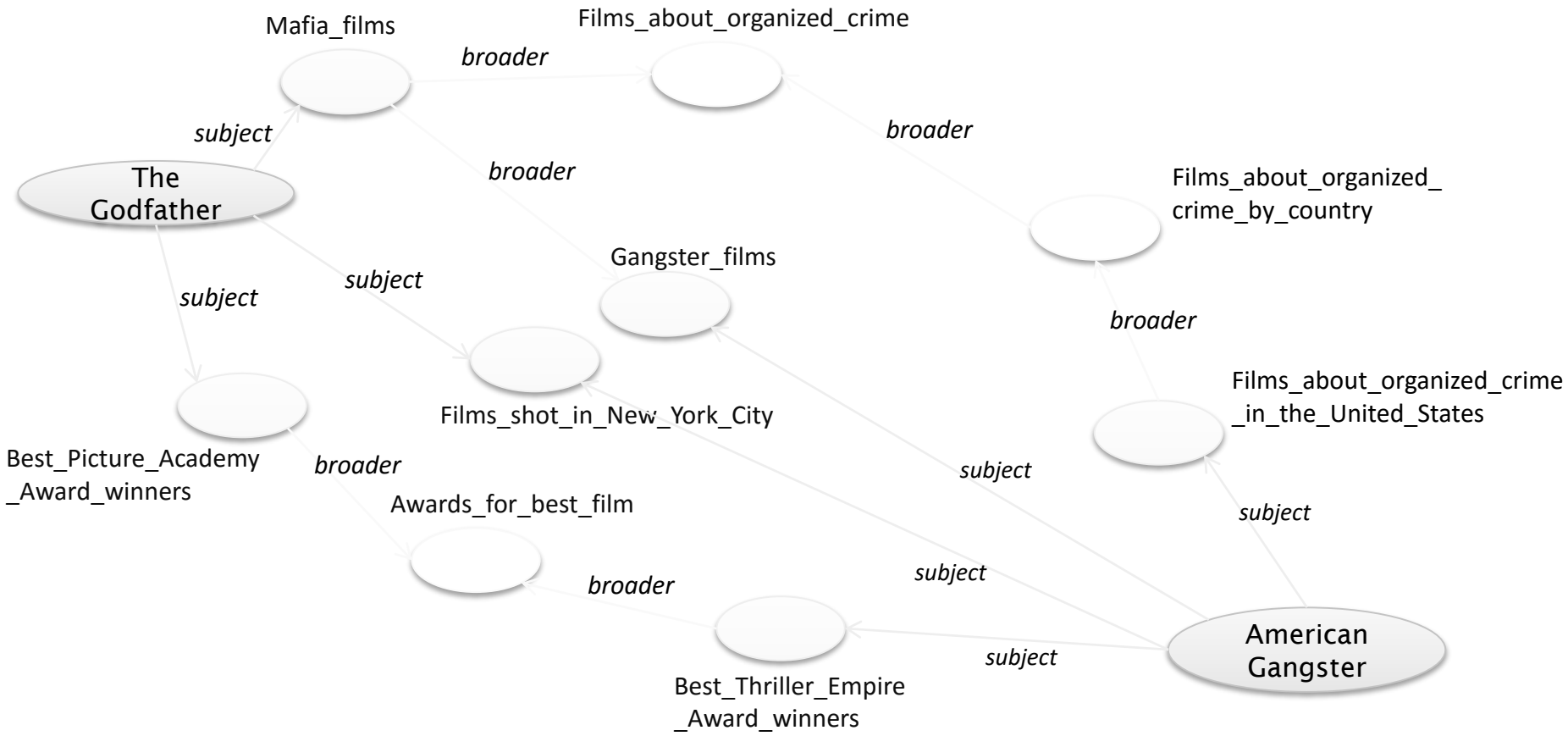
Sergio Oramas, Vito Claudio Ostuni, Tommaso Di Noia, Xavier Serra, Eugenio Di Sciascio: **Sound and Music Recommendation with Knowledge Graphs**. ACM TIST 8(2): 21:1-21:21 (2017)

Graph-based Item Representation



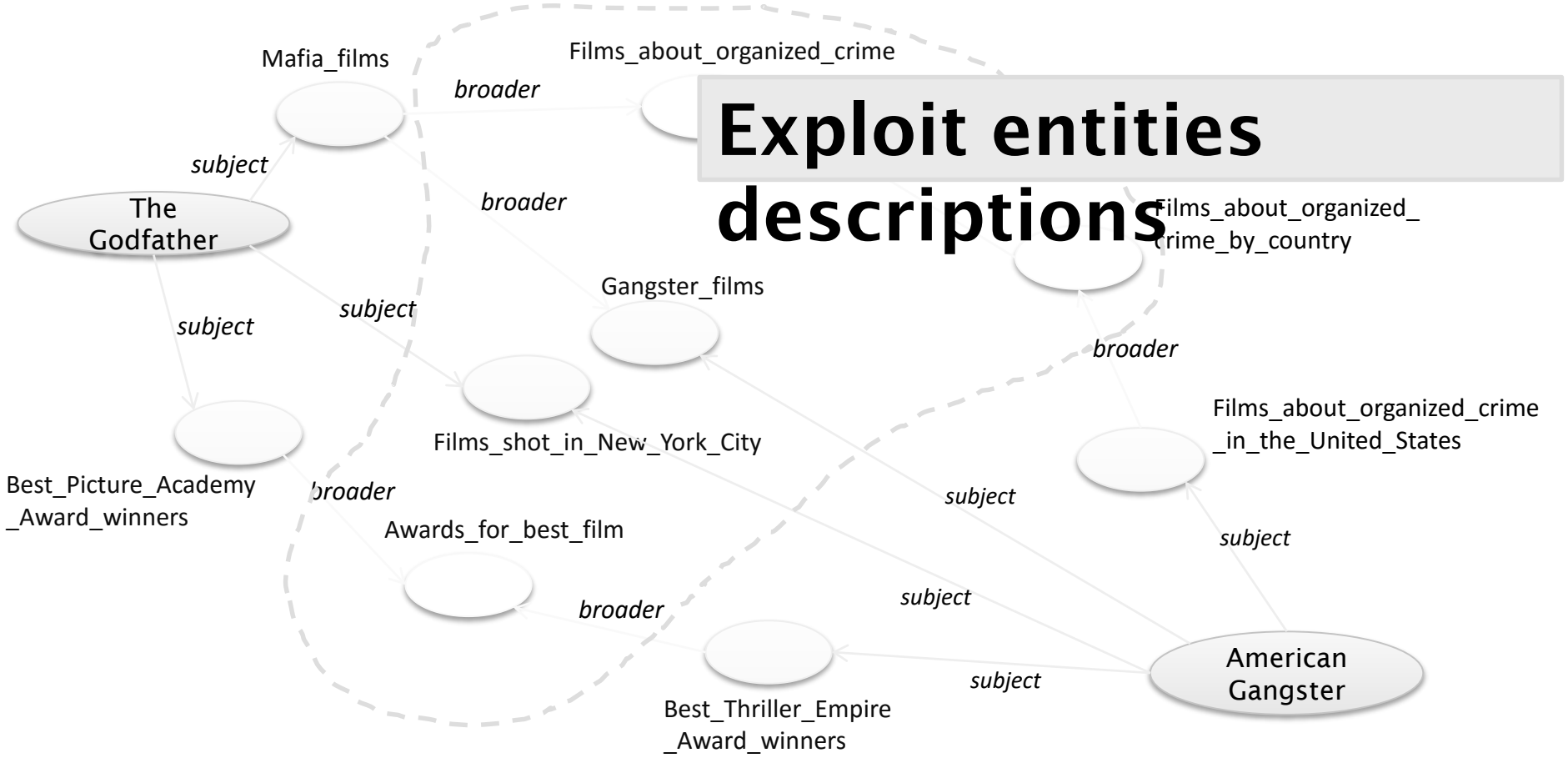
Sergio Oramas, Vito Claudio Ostuni, Tommaso Di Noia, Xavier Serra, Eugenio Di Sciascio: **Sound and Music Recommendation with Knowledge Graphs**. ACM TIST 8(2): 21:1-21:21 (2017)

Graph-based Item Representation



Sergio Oramas, Vito Claudio Ostuni, Tommaso Di Noia, Xavier Serra, Eugenio Di Sciascio: **Sound and Music Recommendation with Knowledge Graphs**. ACM TIST 8(2): 21:1-21:21 (2017)

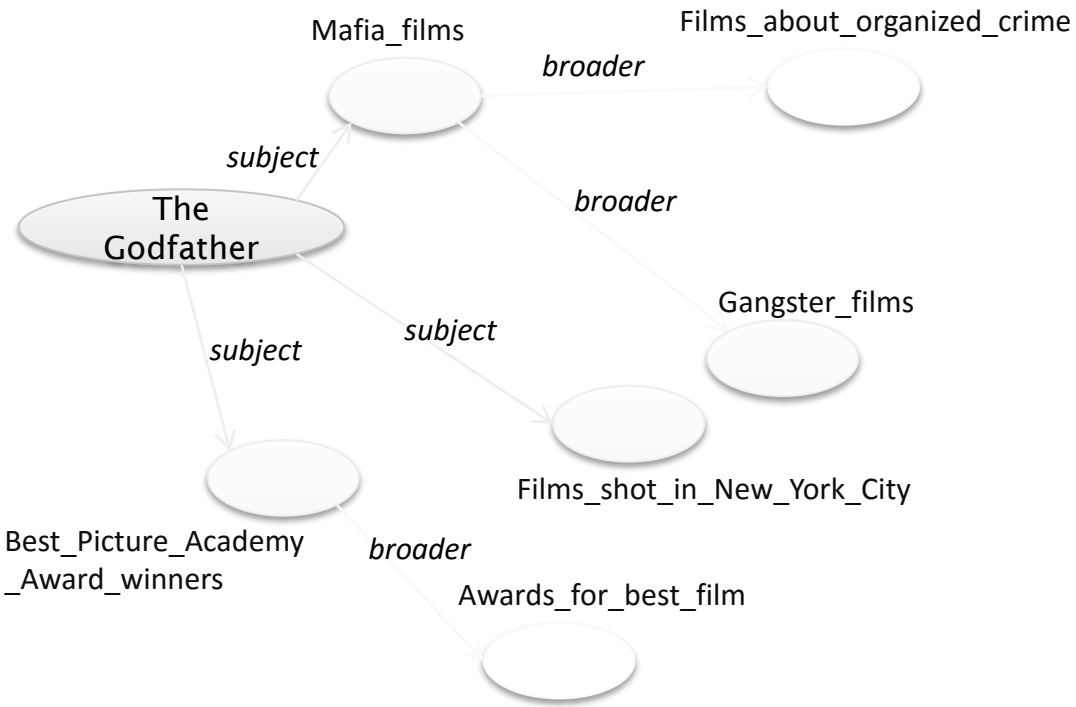
Graph-based Item Representation



Exploit entities descriptions

Sergio Oramas, Vito Claudio Ostuni, Tommaso Di Noia, Xavier Serra, Eugenio Di Sciascio: **Sound and Music Recommendation with Knowledge Graphs**. ACM TIST 8(2): 21:1-21:21 (2017)

h-hop Item Neighborhood Graph

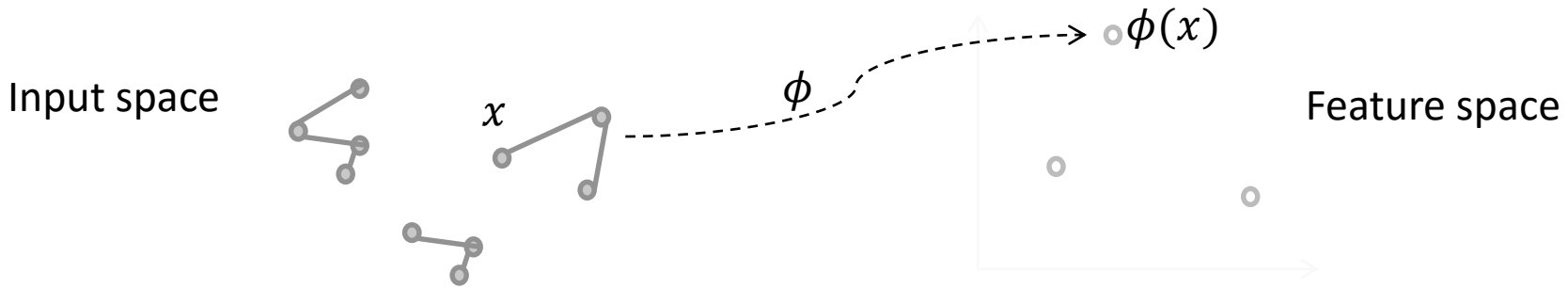


Sergio Oramas, Vito Claudio Ostuni, Tommaso Di Noia, Xavier Serra, Eugenio Di Sciascio: **Sound and Music Recommendation with Knowledge Graphs**. ACM TIST 8(2): 21:1-21:21 (2017)

Kernel Methods

Work by ***embedding data*** in a vector space and looking for linear patterns in such space

$$x \rightarrow \phi(x)$$



We can work in the new space *F* by specifying an inner product function between points in it

$$k(x_i, x_j) = < \phi(xi), \phi(xj) >$$

[Kernel Methods for General Pattern Analysis. Nello Cristianini . *http://www.kernel-methods.net/tutorials/KMtalk.pdf*]

Neighborhood Graph Kernel

$$k_{G^h}(x_i, x_j) = \langle \phi_{G^h}(x_i), \phi_{G^h}(x_j) \rangle$$

Explicit computation of the feature map

$$\phi_{G^h}(x_i) = (w_{x_i,e_1}, w_{x_i,e_2}, \ldots, w_{x_i,e_m}, \ldots, w_{x_i,e_t})$$

Importance of the entity e_m in the neighborhood graph for the item x_i

h-hop Item Entity-based Neighborhood Graph Kernel

$$k_{G^h}(x_i, x_j) = \langle \phi_{G^h}(x_i), \phi_{G^h}(x_j) \rangle$$

Explicit computation of the feature map

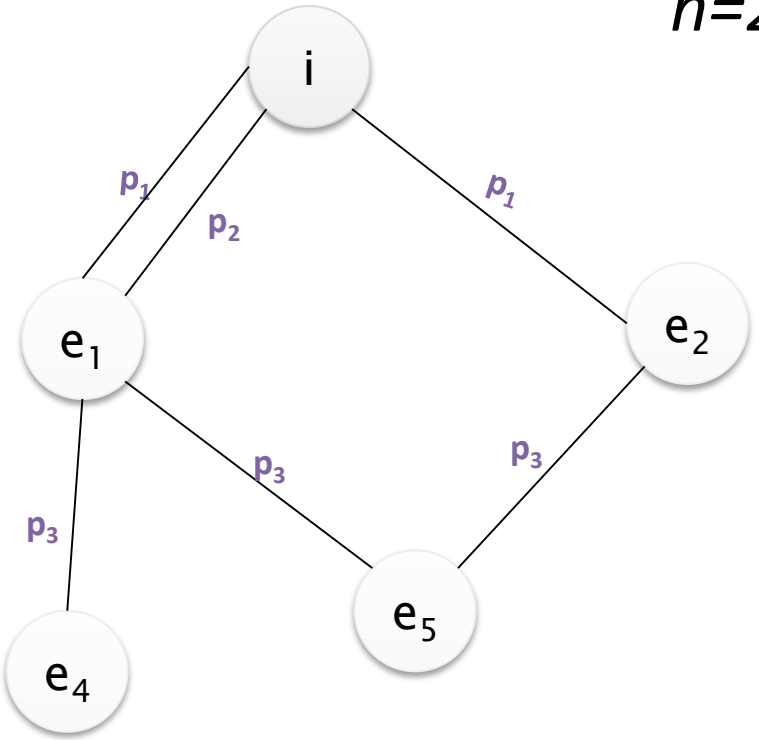
$$\phi_{G^h}(x_i) = (w_{x_i,e_1}, w_{x_i,e_2}, \ldots, w_{x_i,e_m}, \ldots, w_{x_i,e_t})$$

$$w_{x_i,e_m} = \sum_{l=1}^{h} \alpha_l \cdot c_{\hat{P}^l(x_i),e_m}$$

← # edges involving e_m at l hops from x_i a.k.a. frequency of the entity in the item neighborhood graph

↑ factor taking into account at which hop the entity appears

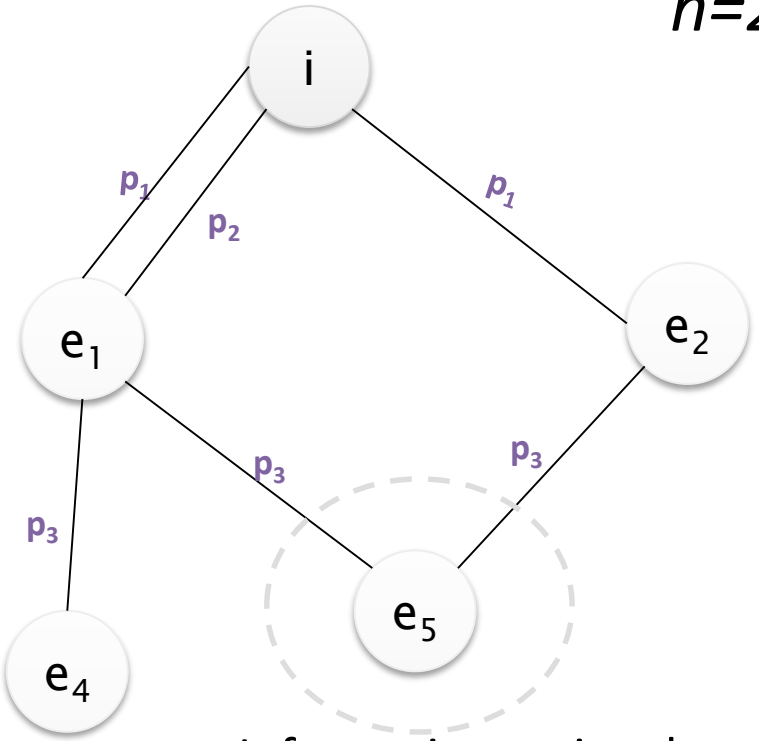
Weights computation



$$c_{\hat{P}^1(x_i),e_1} = 2$$
$$c_{\hat{P}^1(x_i),e_2} = 1$$
$$c_{\hat{P}^2(x_i),e_4} = 1$$
$$c_{\hat{P}^2(x_i),e_5} = 2$$

Weights computation example

h=2



$$c_{\hat{P}^1(x_i),e_1} = 2$$
$$c_{\hat{P}^1(x_i),e_2} = 1$$
$$c_{\hat{P}^2(x_i),e_4} = 1$$
$$c_{\hat{P}^2(x_i),e_5} = 2$$

Informative entity about the item even if not directly related to it

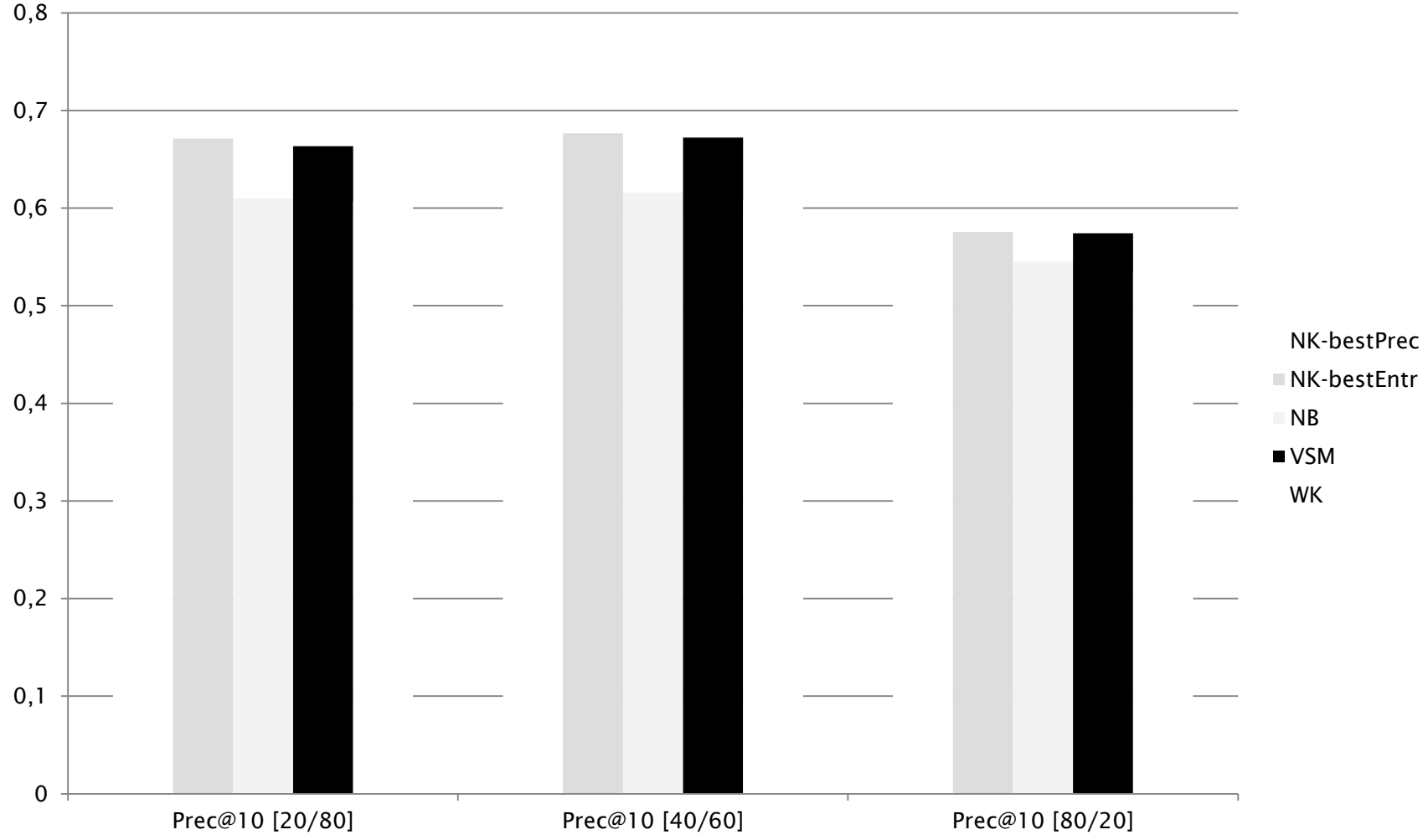
Experimental Settings

- Trained a SVM Regression model for each user
- Accuracy Evaluation: *Precision, Recall*
- Novelty Evaluation: *Entropy-based Novelty* (All Items protocol) *[the lower the better]*

Comparative approaches

- NB: 1-hop item neigh. + Naive Bayes classifier
- VSM: 1-hop item neigh. Vector Space Model (tf-idf) + SVM regr
- WK: 2-hop item neigh. Walk-based kernel + SVM regr

Comparison with other approaches (i)



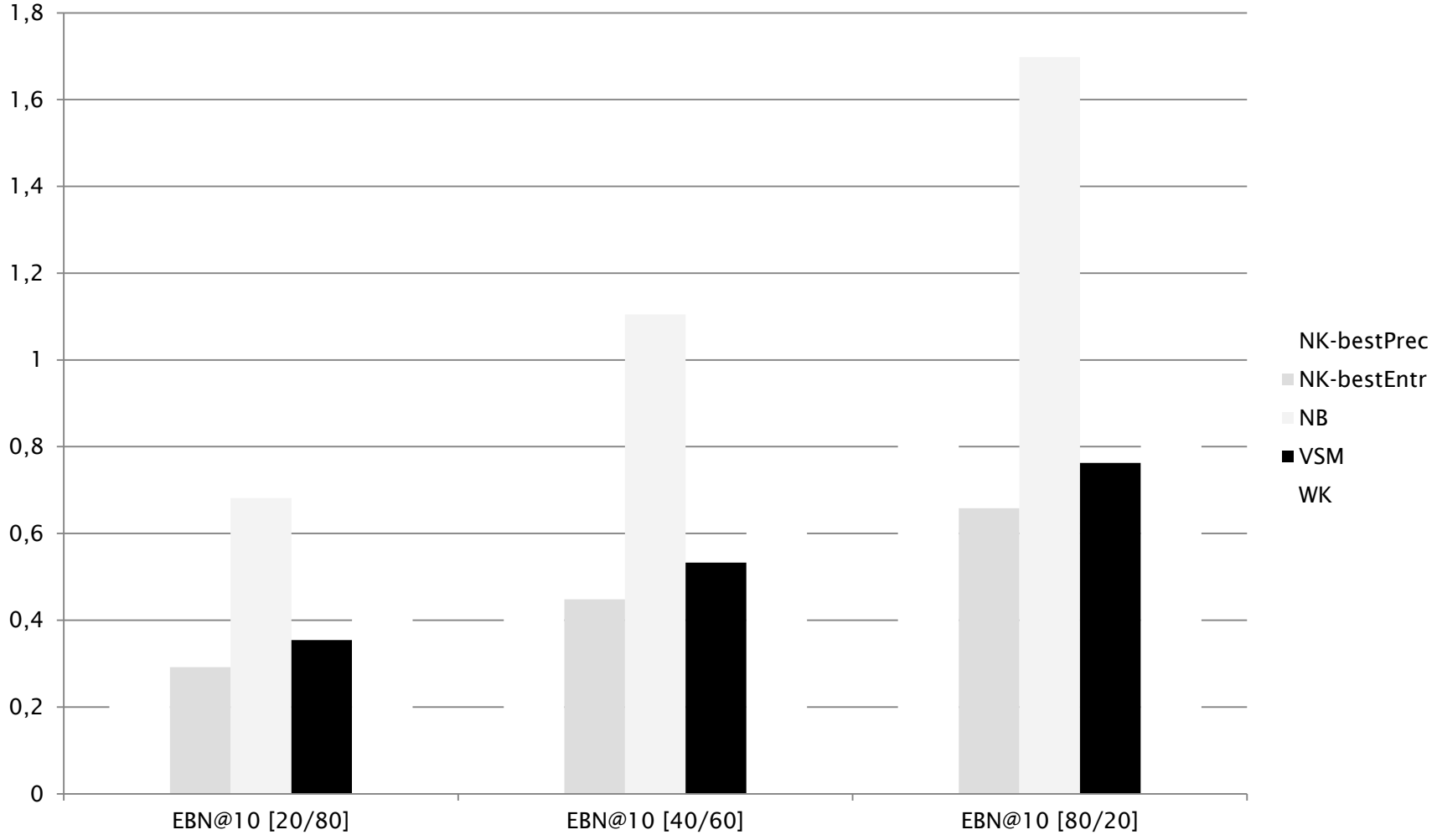
Rated test items protocol

NB: 1-hop item neigh. + Naive Bayes classifier

VSM: 1-hop item neigh. Vector Space Model (tf-idf) + SVM regr

WK: 2-hop item neigh. Walk-based kernel + SVM regr

Comparison with other approaches (ii)



Neighborhood Graph Kernel

$$k_{G^h}(x_i, x_j) = \langle \phi_{G^h}(x_i), \phi_{G^h}(x_j) \rangle$$

Explicit computation of the feature map

$$\phi_{G^h}(x_i) = (w_{x_i,p*_1}, \ldots, w_{x_i,p*_m}, \ldots, w_{x_i,p*_t})$$

$$w_{x_i,p*_m} = \frac{\#p *_m (x_i)}{|p_m| - |p *_m|}$$

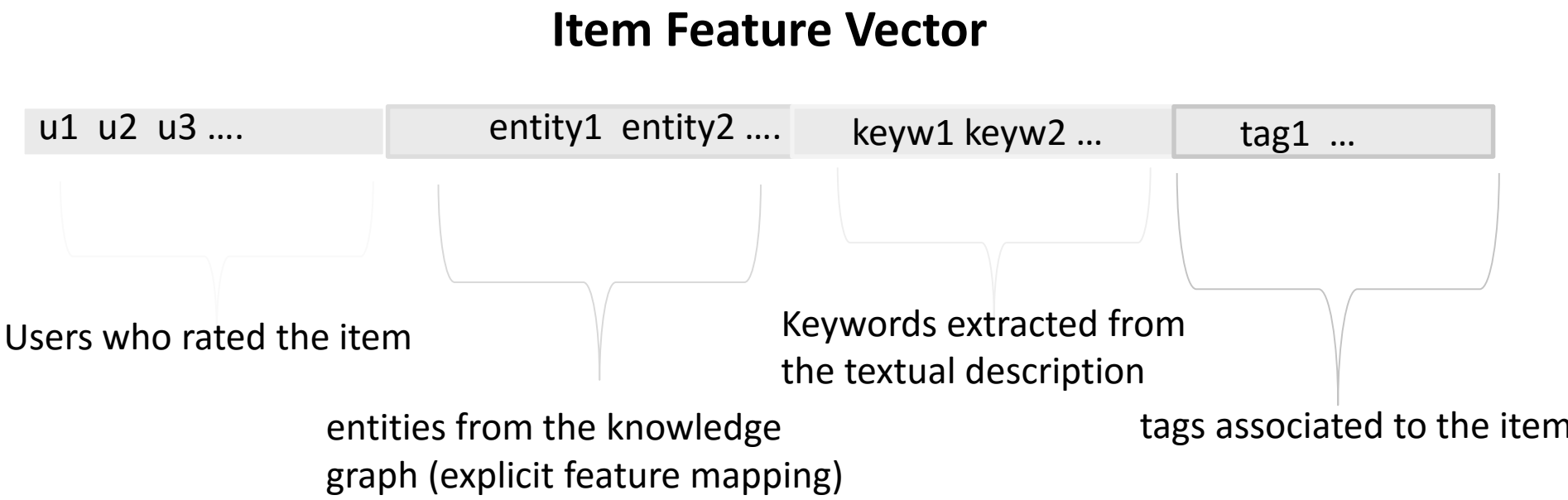
← # sequences and subsequences of nodes from x_i to e_m

↑ Normalization factor

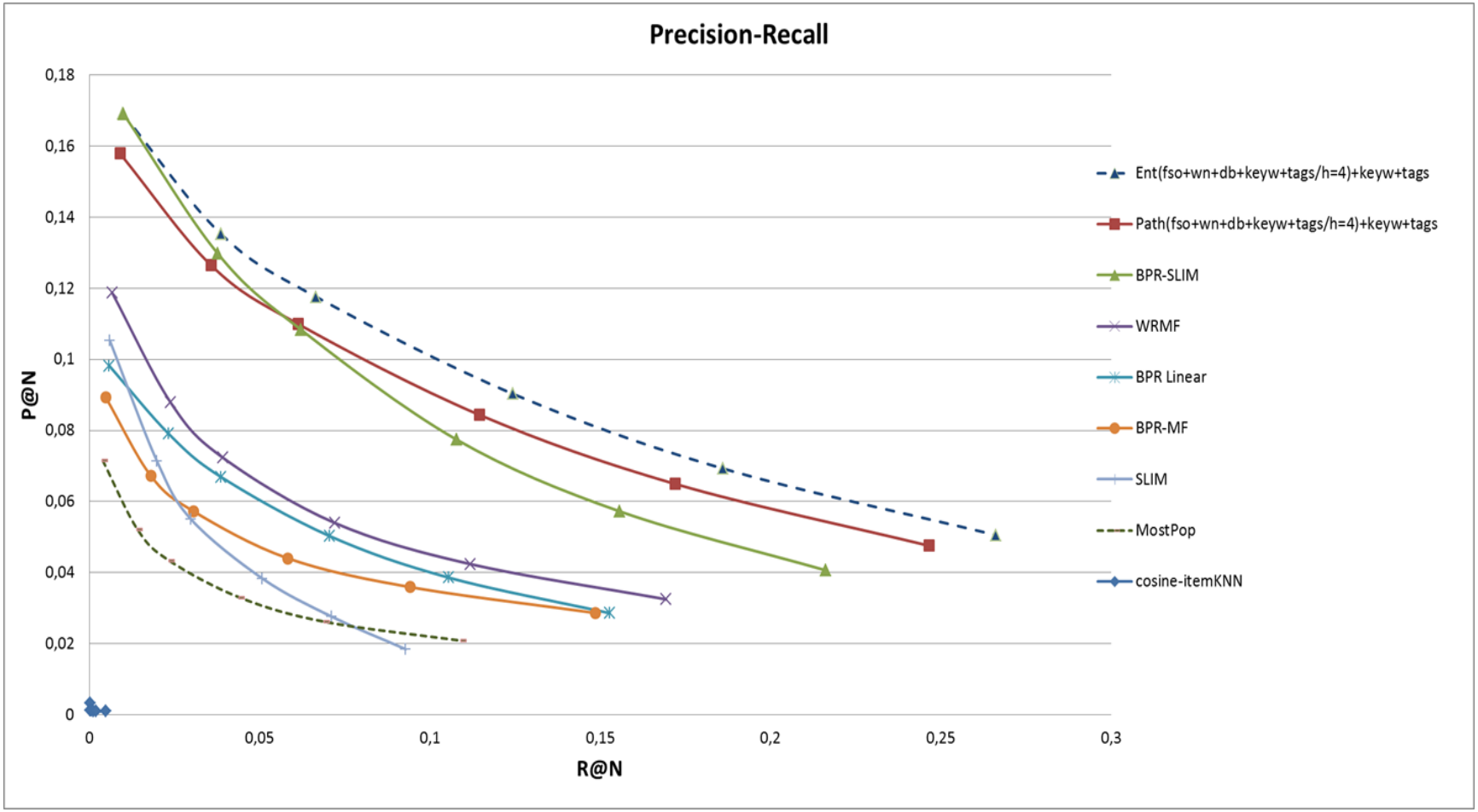
via
Feature Combination

The hybridizations is based on the combination of different data sources

Final approach: collaborative + LOD + textual description + tags

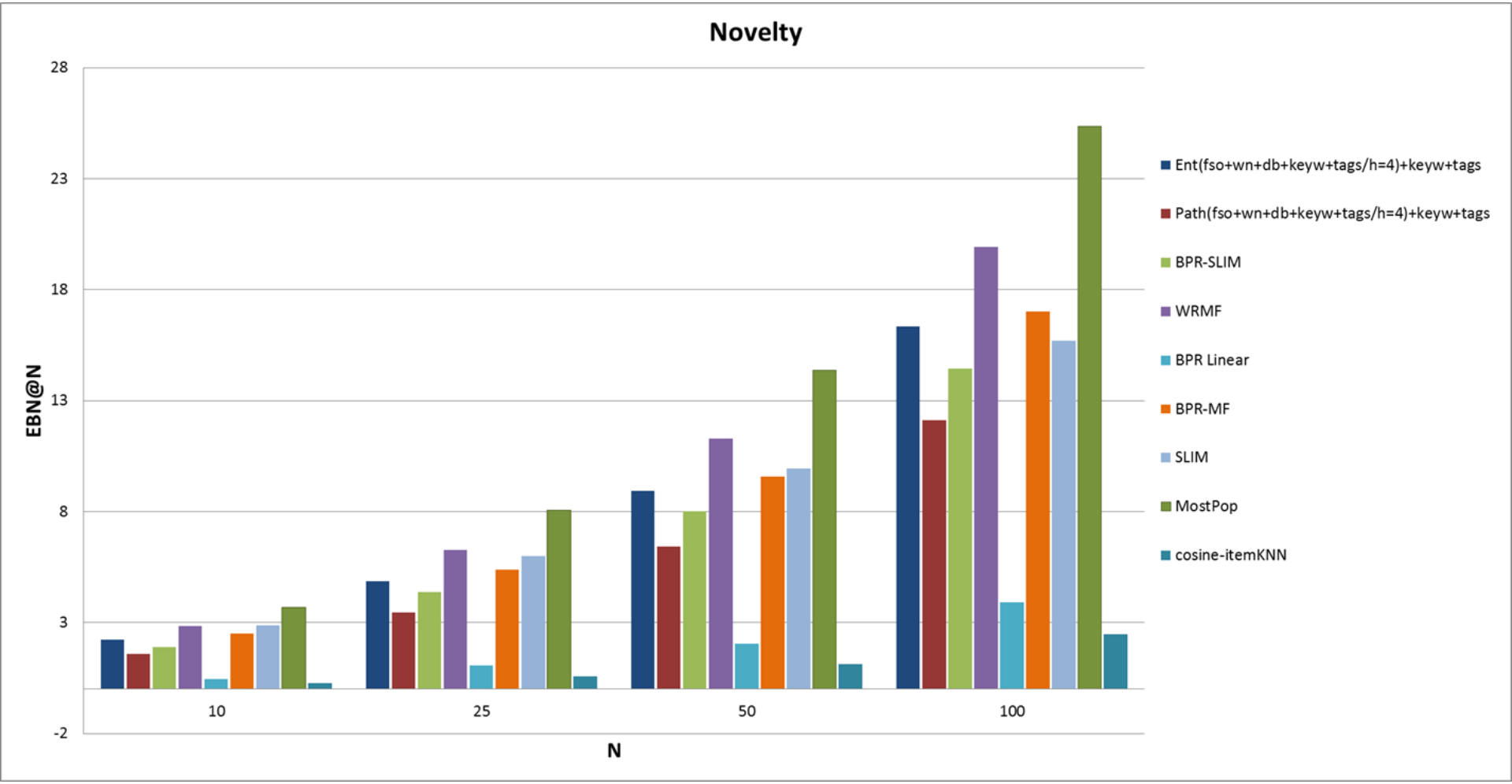


Accuracy

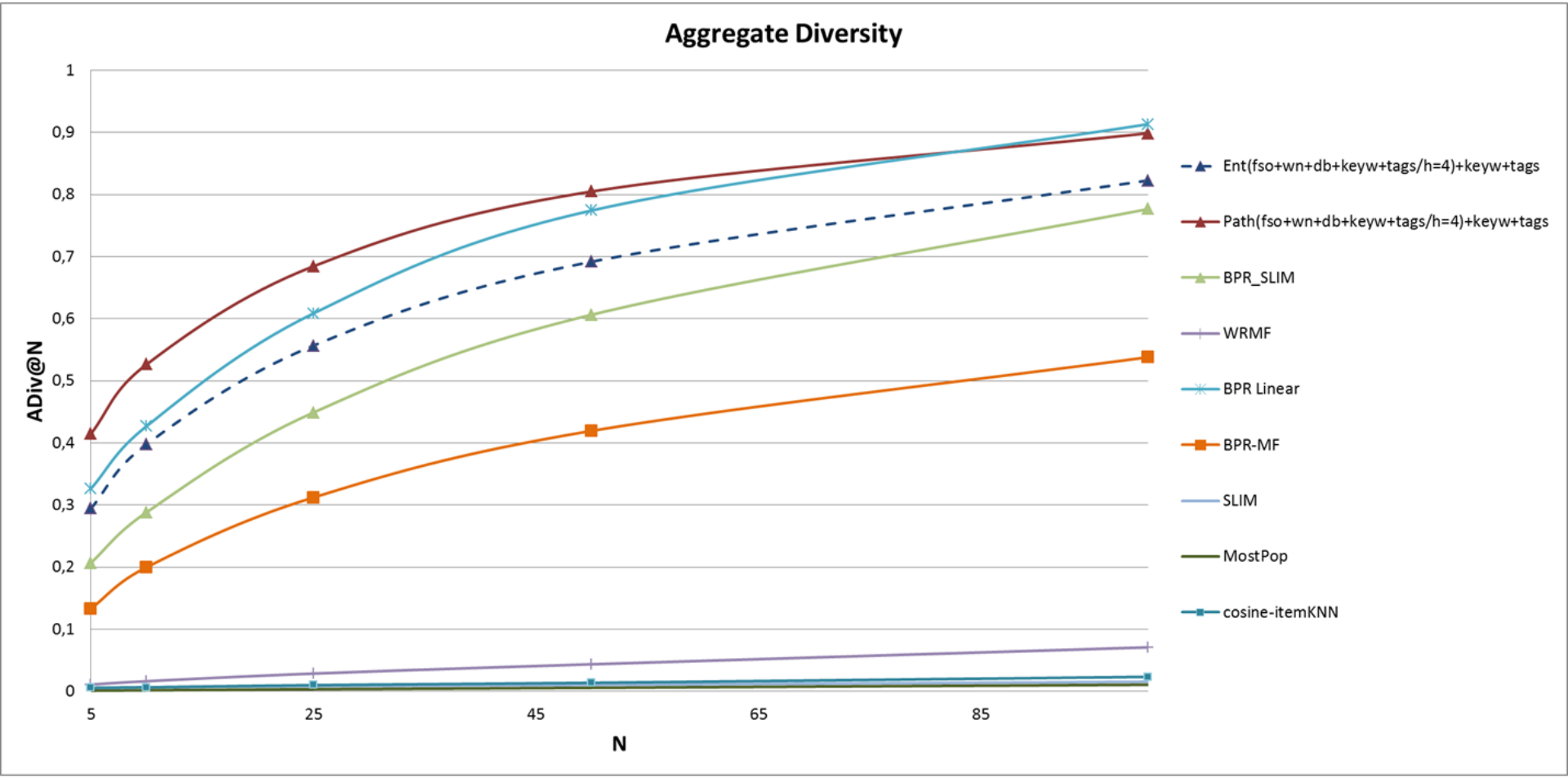


All items protocol

Long Tail


Novelty
EBN@N
28
23
18
13
8
3
-2
10
25
50
100
N
Ent(fso+wn+db+keyw+tags/h=4)+keyw+tags
Path(fso+wn+db+keyw+tags/h=4)+keyw+tags
BPR-SLIM
WRMF
BPR Linear
BPR-MF
SLIM
MostPop
cosine-itemKNN

Aggregate Diversity



Implementation

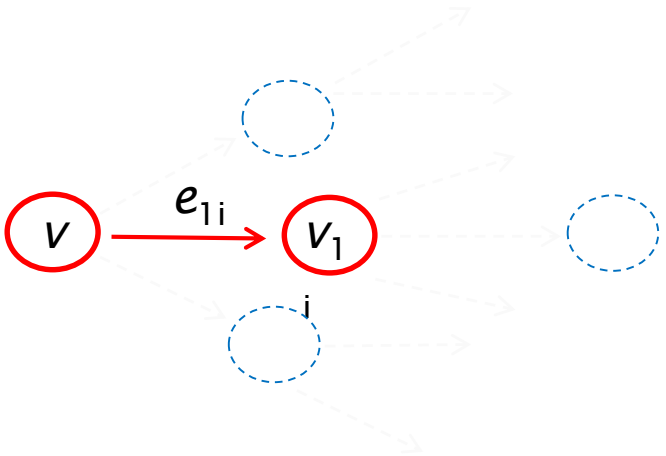
- LODreclib – a Java library to build a LOD based recommender system

https://github.com/sisinflab/lodreclib

RDF Graph Embeddings

Given G = (**V**, **E**), for each vertex $v \in \mathbf{V}$, we generate all graph walks P_V of depth *d* rooted in *v*.

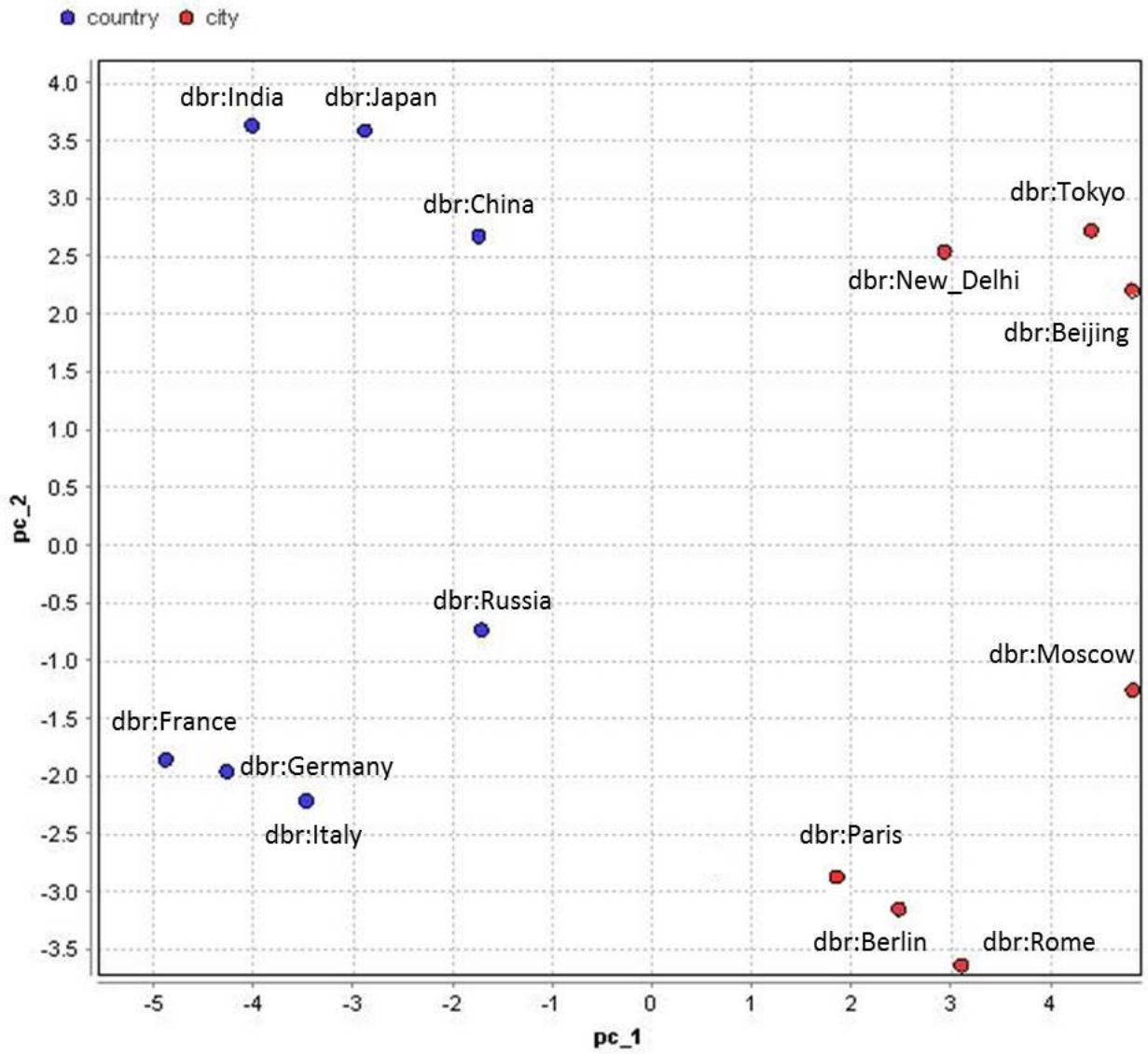
1. $v \rightarrow e_{1i,}$ $i \in \mathbf{E}(v)$
2. $v \rightarrow e_{1i} \rightarrow v_{1i}$
3. Continue until *d* iterations are reached



Following the approach proposed in "*DeepWalk: Online Learning of Social Representations*" by Bryan Perozzi et al., only a **limited** number of **random** walks per each entity may be considered.

The stream of random walks becomes the input of neural models.

Jessica Rosati, Petar Ristoski, Tommaso Di Noia, Renato De Leone, Heiko Paulheim: **RDF Graph Embeddings for Content-based Recommender Systems**. CBRecSys@RecSys 2016: 23-30



Using **DBpedia** as RDF graph, semantic similar entities appear close in the latent space and in its 2-dimensional PCA projection .

Item description for CBRSs

Cosine similarity between the feature vectors has been adopted to estimate the **closeness**/similarity between to items.

In our experiments we use two RDF graphs to define the item content

- *Dbpedia* ~4 million of resources: extracted from structured data in Wikipedia (e.g., infoboxes);

- *Wikidata* ~17 million of resources : a collaboratively edited knowledge graph that also hosts various edition of Wikipedia.

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

Random Indexing

Word2Vec

……

Word2Vec

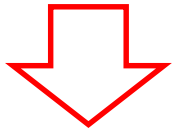
- **Empirical Comparison of Word Embedding Techniques for Content-based Recommender Systems [*]**

- **Methodology**
  - **Build a WordSpace using different Word Embedding techniques (and different sizes)**
  - Build a DocSpace as the centroid vectors of term vectors
  - **Build User Profiles as centroid** of the items they liked
  - Provide **Users with Recommendations**
  - **Compare the approaches**

Musto, Cataldo, et al. **"Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems."** *European Conference on Information Retrieval*. ECIR 2016.

Word2Vec

Results



MovieLens	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,4645	0,4715	0,4921	0,4910	**0,5056**	0,5054	**0,5217**	*0,5022*	0,5141
F1@10	0,5393	0,5469	0,5622	0,5613	**0,5757**	0,5751	**0,5969**	0,5836	0,5928
F1@15	0,5187	0,5254	0,5349	0,5352	0,5672	**0,5674**	**0,5911**	0,5814	0,5876

Word Embedding overcomes **I2I-CF only on F1@5**. Needs to further process content **on less sparse datasets.**

Musto, Cataldo, et al. **"Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems."** *European Conference on Information Retrieval*. ECIR 2016.

Word2Vec

Results

MovieLens	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,4645	0,4715	0,4921	0,4910	**0,5056**	0,5054	**0,5217**	*0,5022*	0,5141
F1@10	0,5393	0,5469	0,5622	0,5613	**0,5757**	0,5751	**0,5969**	0,5836	0,5928
F1@15	0,5187	0,5254	0,5349	0,5352	0,5672	**0,5674**	**0,5911**	0,5814	0,5876

Word Embedding overcomes **I2I-CF only on F1@5**. Needs to further process content **on less sparse datasets.**

DBbook	LSI		RI		W2V		U2U-CF	I2I-CF	BPRMF
	300	500	300	500	300	500			
F1@5	0,5056	0,5076	0,5064	0,5039	0,5183	**0,5186**	0,5193	*0,5111*	**0,5290**
F1@10	0,6256	**0,6260**	0,6239	0,6244	0,6207	0,6209	0,6229	*0,6194*	**0,6263**
F1@15	0,5908	**0,5909**	0,5892	0,5887	0,5829	0,5828	*0,5777*	*0,5776*	*0,5778*

Results comparable to CF and MF on more sparse datasets. LSI is the best-performing approach on F1@15

Musto, Cataldo, et al. **"Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems."** *European Conference on Information Retrieval*. ECIR 2016.

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

Random Indexing

Word2Vec

……

eVSM

- **Enhanced Vector Space Model [*]**
- **Content-based Recommendation Framework**

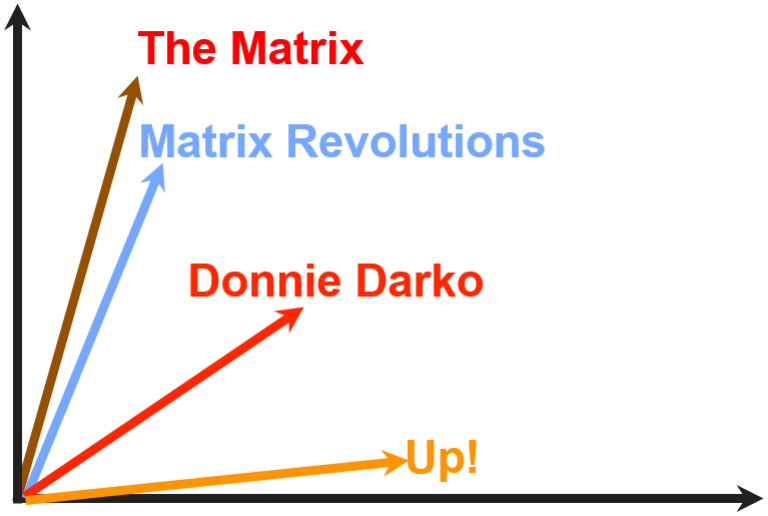
- **Cornerstones**
  - Semantics modeled through **Distributional Models**
  - **Random Indexing** for Dimensionality Reduction
  - Negative Preferences modeled through **Quantum Negation [^]**
  - **User Profiles** as centroid vectors of items representation
  - Recommendations **through Cosine Similarity**

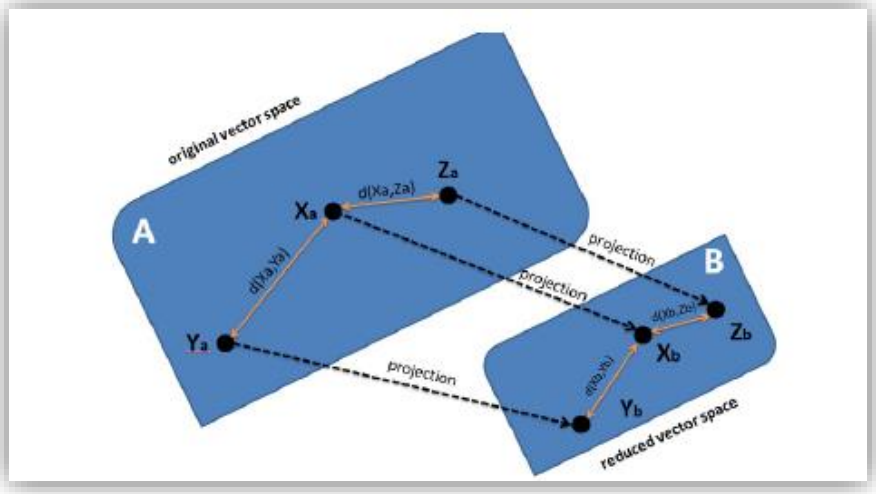
[*] Musto, Cataldo. **"Enhanced vector space models for content-based recommender systems."** *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010.

[^] Widdows, Dominic, and Stanley Peters. **"Word vectors and quantum logic: Experiments with negation and disjunction."** *Mathematics of language* 8.141-154 (2003).

eVSM

Cornerstones





Distributional Models
to build DocSpace of the items

(whole document used as context)

Random Indexing for
Dimensionality Reduction

Musto, Cataldo. **"Enhanced vector space models for content-based recommender systems."** *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010.

eVSM

Cornerstones

- **Given two vectors a e b**
- Through Quantum Negation we can define a Vector **(a ∧¬b)**
- **Formally:**
  - Projection of **vector a** on the subspace **orthogonal to that** generated **by vector b**

- **Intuitively:**
  - Vector «a» models «positive» preferences
  - **Vector «b» models «negative» preferences**
  - **Through quantum negation we get a unique vector modeling both aspects**
    - Close to vectors containing as many as possible features from «a» and as less as possible features from «b»

Widdows, Dominic, and Stanley Peters. **"Word vectors and quantum logic: Experiments with negation and disjunction."** *Mathematics of language* 8.141-154 (2003).

eVSM
Cornerstones

- **User Profiles**
  - Calculated as **centroid vectors of the items** the user liked/disliked

$$I_u = \{d \in D | r(u,d) \geq \beta\}$$

$$\boldsymbol{p}_u = \sum_{i=1}^{|I_u|} \boldsymbol{d}_i * r(u, d_i)$$

Random Indexing-based Profiles **(RI)**

Musto, Cataldo. **"Enhanced vector space models for content-based recommender systems."** *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010.

eVSM

Cornerstones

- **User Profiles**
  - Calculated as **centroid vectors of the items** the user liked/disliked

$$I_u = \{d \in D | r(u,d) \geq \beta\}$$

$$\boldsymbol{p}_u = \sum_{i=1}^{|I_u|} \boldsymbol{d}_i * r(u, d_i)$$

Random Indexing-based Profiles **(W-RI)**

$$I_u = \{d \in D | r(u,d) \geq \beta\}$$

$$I_u^- = \{d \in D | r(u,d_i) < \beta\}$$

$$\boldsymbol{p}_{+u} = \sum_{i=1}^{|I_u^+|} \boldsymbol{d}_i * r(u, d_i)$$

$$\boldsymbol{p}_{-u} = \sum_{i=1}^{|I_u^-|} \boldsymbol{d}_i * (MAX - r(u, d_i))$$

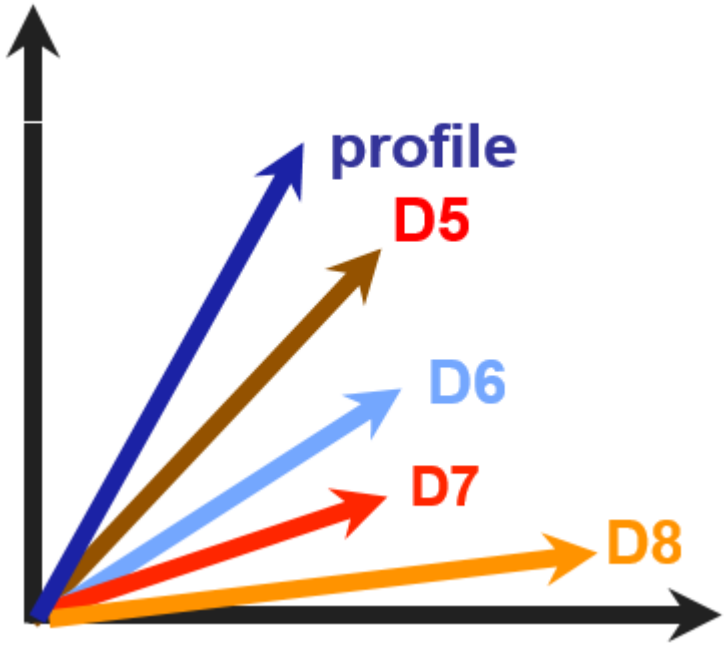
Quantum Negation- based Profiles (W-QN)

Musto, Cataldo. **"Enhanced vector space models for content-based recommender systems."** *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010.

eVSM

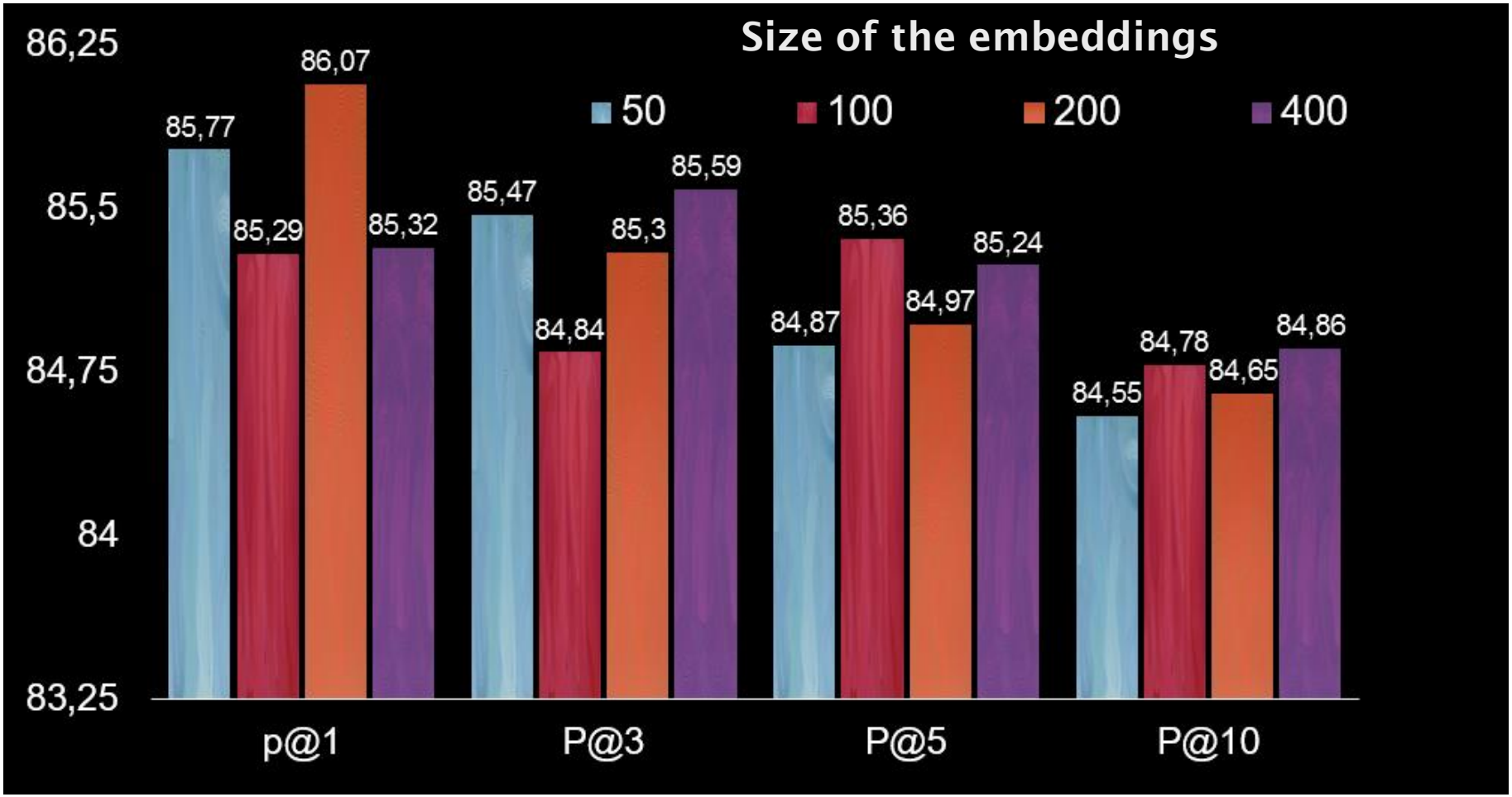
Cornerstones

- **Recommendations**
  - Similarity Calculations on DocSpace



eVSM

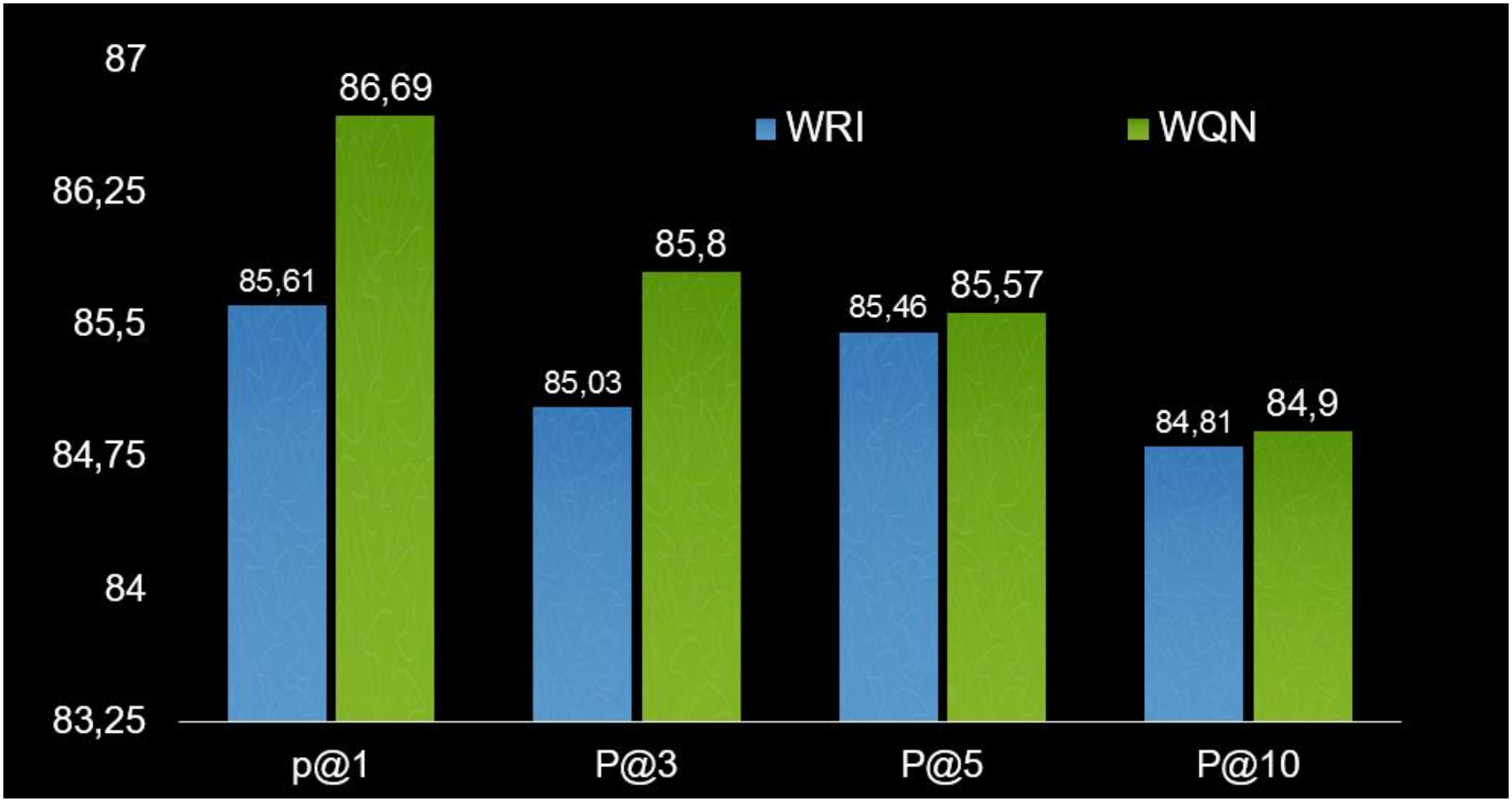
Experiments



The **size** of the embeddings **does not significantly affect** the overall accuracy of eVsm **(MovieLens data)**

eVSM

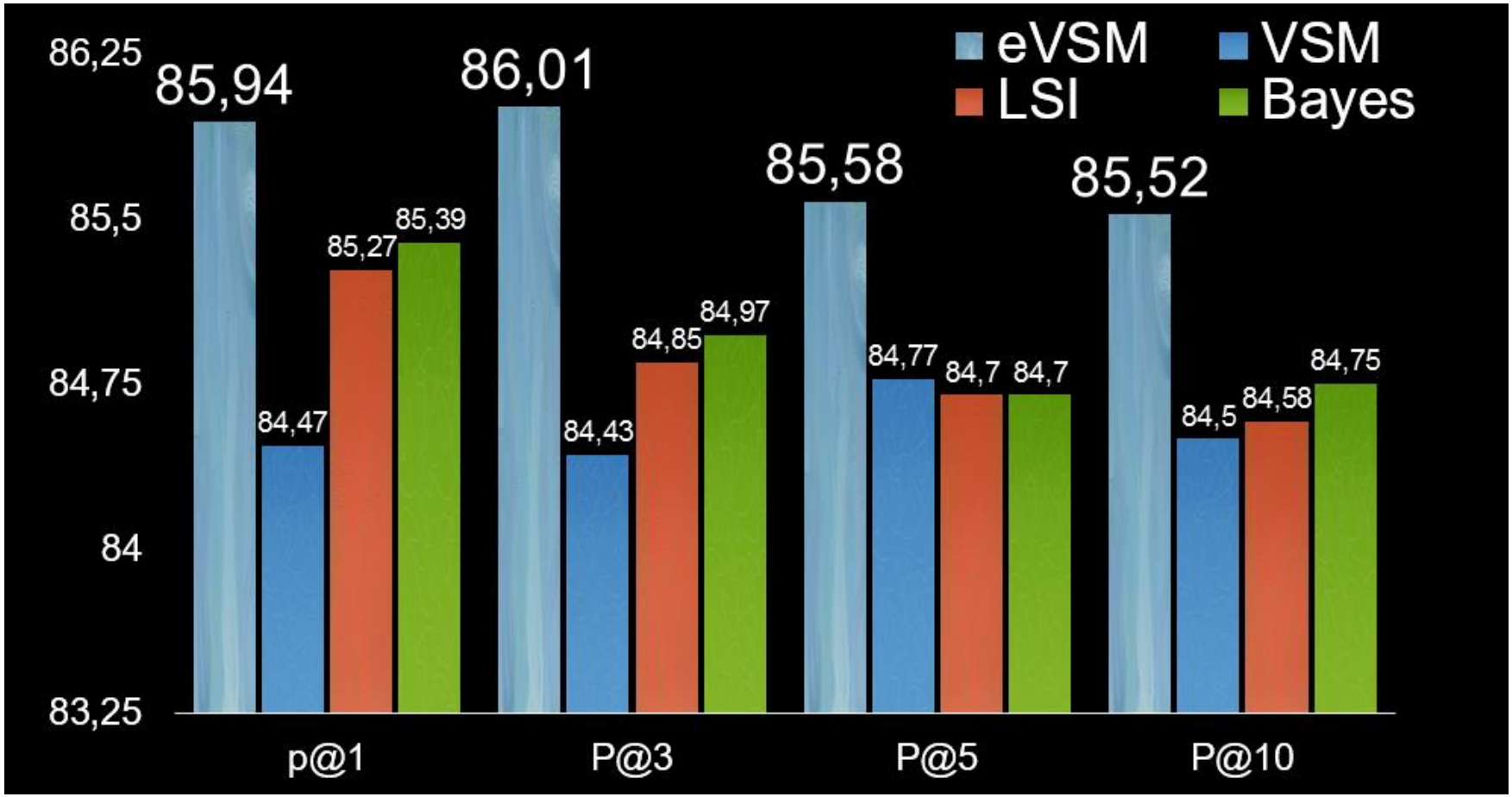
Experiments



Quantum Negation improves the accuracy of the model
(MovieLens data, embedding size=100)

eVSM

Experiments



eVSM significantly overcame all the baselines.

(MovieLens data, embedding size=400)

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Word Sense Disambiguation

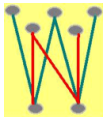
Entity Linking

Explicit Semantic Analysis

Random Indexing

Word2Vec

.....



C-eVSM

- **Contextual Enhanced Vector Space Model [*]**
  - **Extension of eVSM: context-aware Framework**

- **Cornerstones**
  - **Entity Linking** of the content through Tag.me
  - Semantics modeled through **Distributional Models**
  - **Random Indexing** for Dimensionality Reduction
  - **Distributional Models also used to build a representation of the context**
  - **Context-aware User Profiles** as centroid vectors
  - Recommendations **through Cosine Similarity**

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Context-aware User Profiles**

$$\text{C-WRI}(u,c_k,v_j) = \alpha * \text{WRI}(u) + (1-\alpha) * \text{context}(u,c_k,v_j)$$

Let **u** be the **target user**

Let $\mathbf{c_k}$ be a **contextual variable** (e.g. task, mood, etc.)

Let $\mathbf{v_j}$ be **its value** (e.g. task=running, mood=sad, etc.)

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Context-aware User Profiles**

$$\text{C-WRI}(u,c_k,v_j) = \alpha * \text{WRI}(u) + (1-\alpha) * \text{context}(u,c_k,v_j)$$

eVSM profile

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Context-aware User Profiles**

$$C\text{-}WRI(u,c_k,v_j) = \alpha * WRI(u) + (1-\alpha) * context(u,c_k,v_i)$$

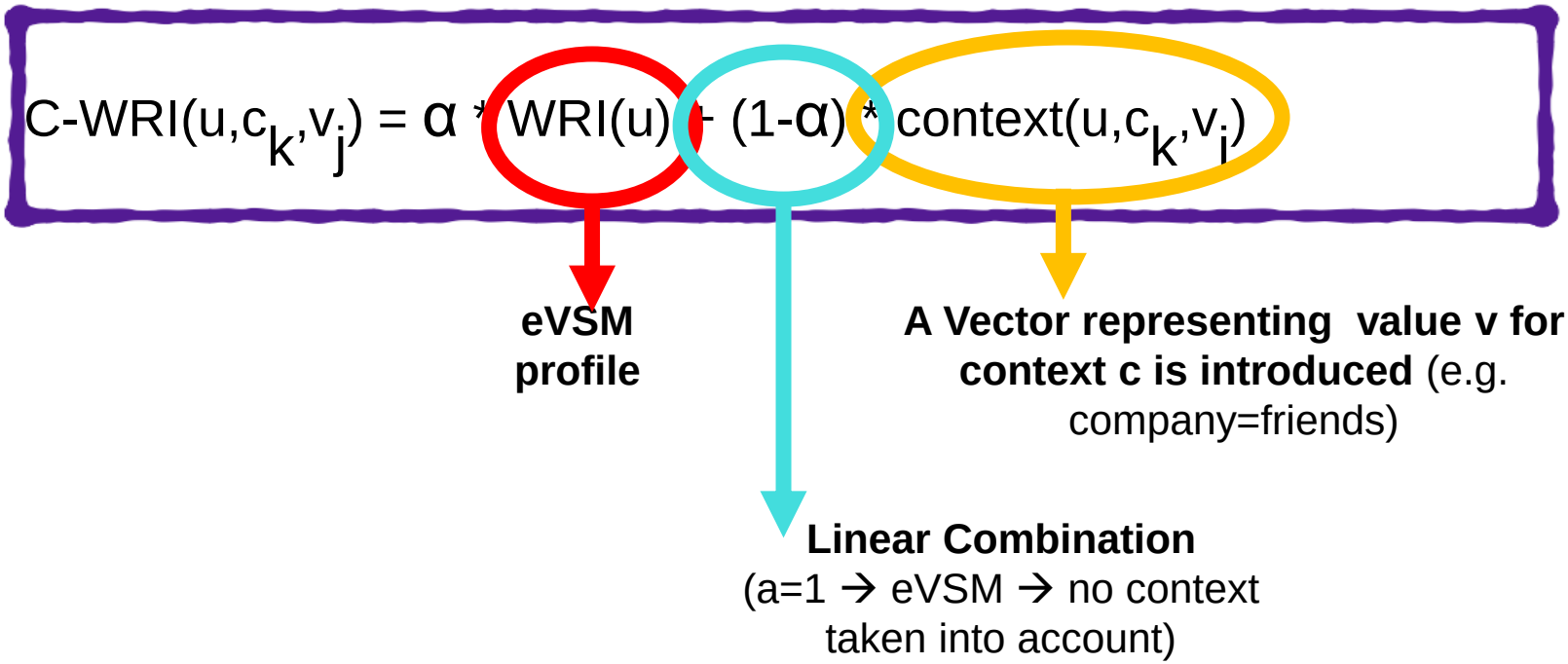
eVSM profile

A Vector representing value v for context c is introduced (e.g. company=friends)

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Context-aware User Profiles**



Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Context-aware User Profiles**

$$C\text{-}WRI(u,c_k,v_j) = \alpha * WRI(u) + (1-\alpha) * context(u,c_k,v_j)$$

eVSM profile

A Vector representing value v for context c is introduced (e.g. company=friends)

Linear Combination
(a=1 → eVSM → no context taken into account)

$$context(u,c_k,v_j) = \sum_{i=1}^{|L(c_k,v_j)|} d_i * \frac{r(u,i,c_k,v_j)}{MAX}$$

items the user liked **in that specific context**

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Context-aware User Profiles**

$$C\text{-}WRI(u,c_k,v_j) = \alpha * WRI(u) + (1-\alpha) * context(u,c_k,v_j)$$

eVSM profile

A Vector representing value v for context c is introduced (e.g. company=friends)

Why this formula?

Ratio: context **is just a factor** which can influence user's perception of an item

if the user **did not express any preference** in that specific contextual setting, $\mathbf{context(u,c_k,v_j) = 0}$

—> **non contextual recommendation**

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Why this formula?**

Insight: it exists a set of terms that is **more descriptive of items relevant** in that specific context




for a romantic dinner, e.g. candlelight, seaview, violin

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

- **Why this formula?**

$$context(u,c_k,v_j) = \sum_{i=1}^{|L(c_k,v_j)|} d_i * \frac{r(u,i,c_k,v_j)}{MAX}$$

Context is represented **on the ground of the items** the user liked in that specific contextual setting

Items are represented on the ground of the **co-occurrences between terms (entities)**

the resulting representation of the **context** is such that a **bigger weight** is given to the entities which **typically occur in the description of the items relevant in that specific context**

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

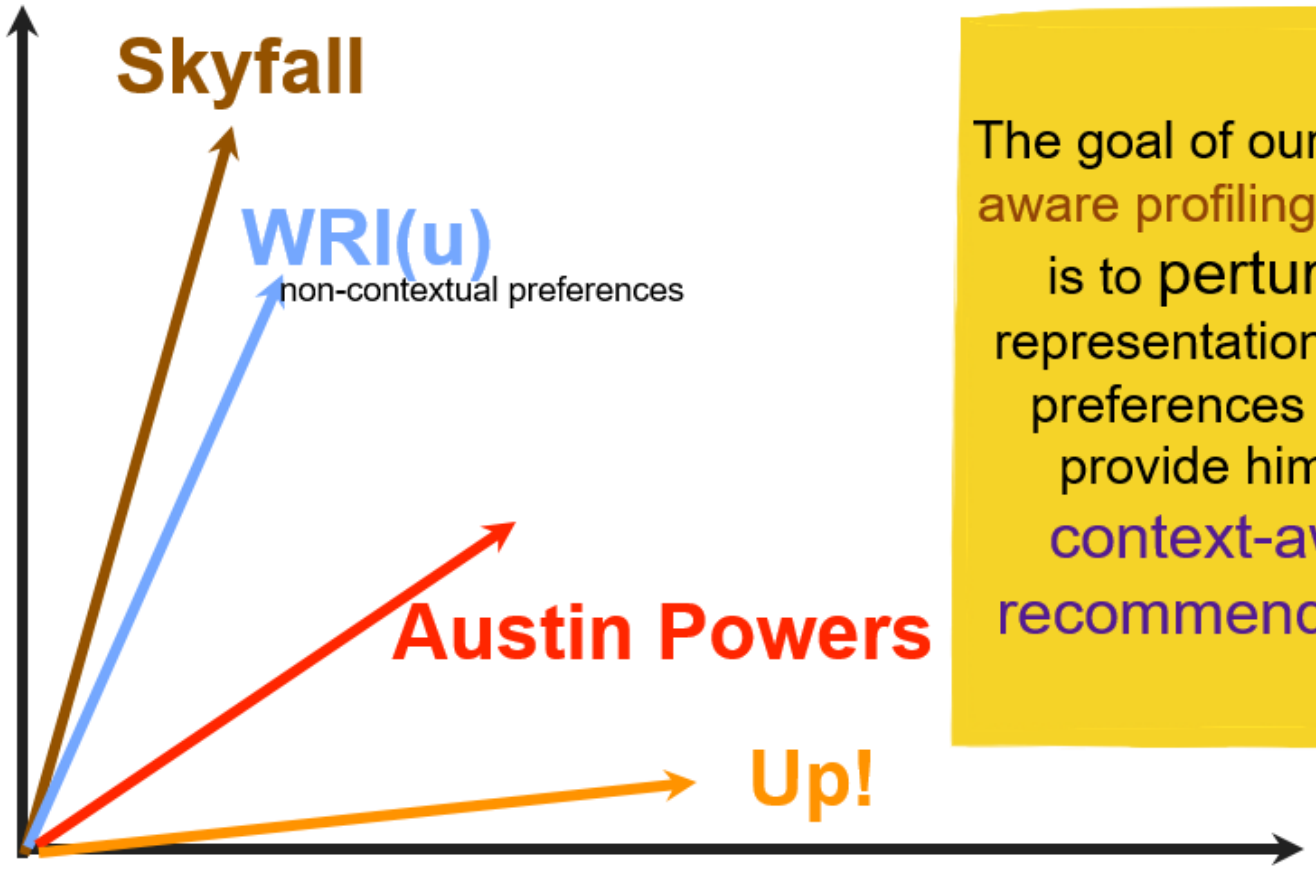
- **Why this formula?**

$$context(u,c_k,v_j) = \sum_{i=1}^{|L(c_k,v_j)|} d_i^* \frac{r(u,i,c_k,v_j)}{MAX}$$

Thanks to **Distributional Semantics Models** it is possible to build a vector-space representation of the context which **emphasize the importance** of those terms, **since they are more used** (→ more important) **in that specific contextual setting.**

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

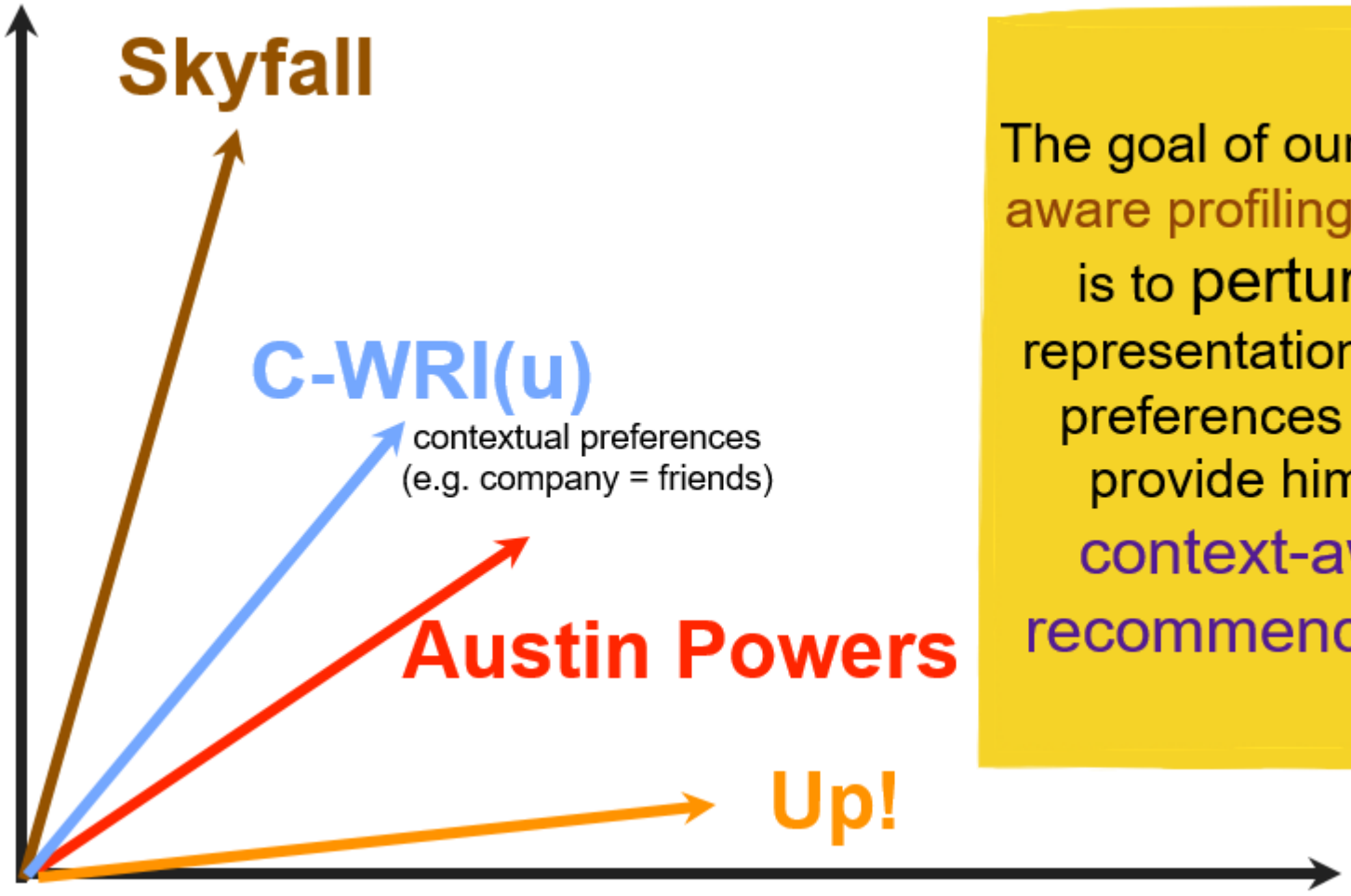
C-eVSM



The goal of our context-aware profiling strategy is to perturb the representation of user preferences and to provide him with context-aware recommendations

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

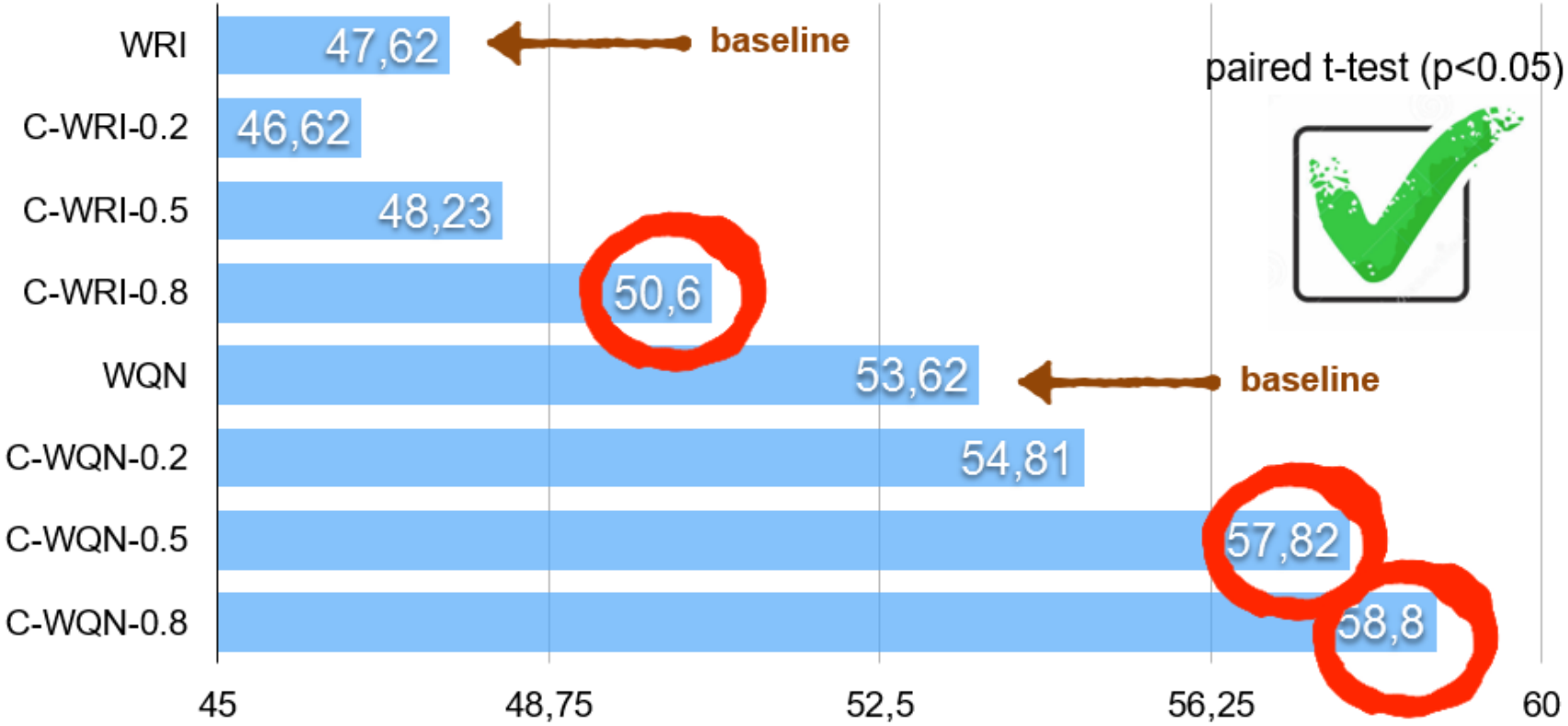


The goal of our context-aware profiling strategy is to perturb the representation of user preferences and to provide him with context-aware recommendations

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

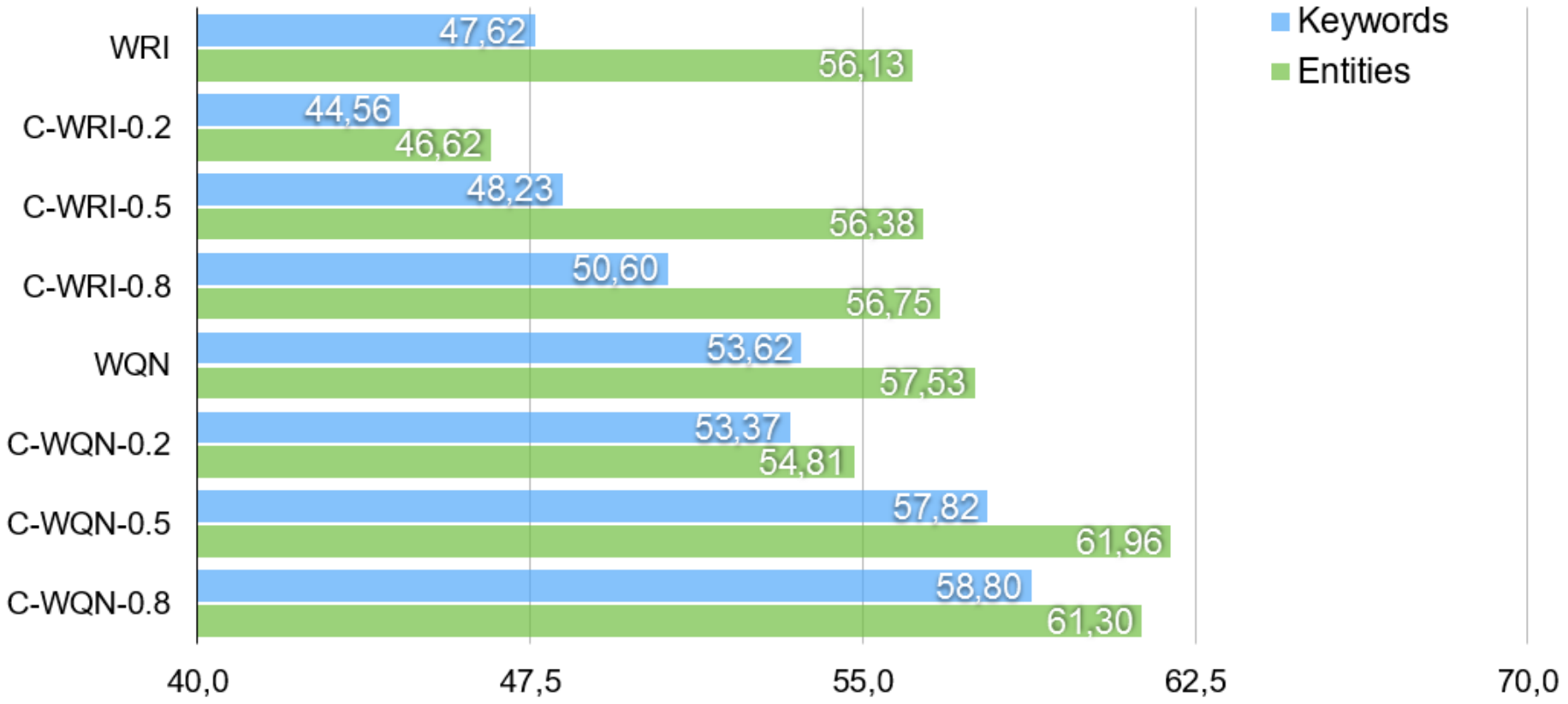
Selection of Results :: HOME segment



Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

Selection of Results



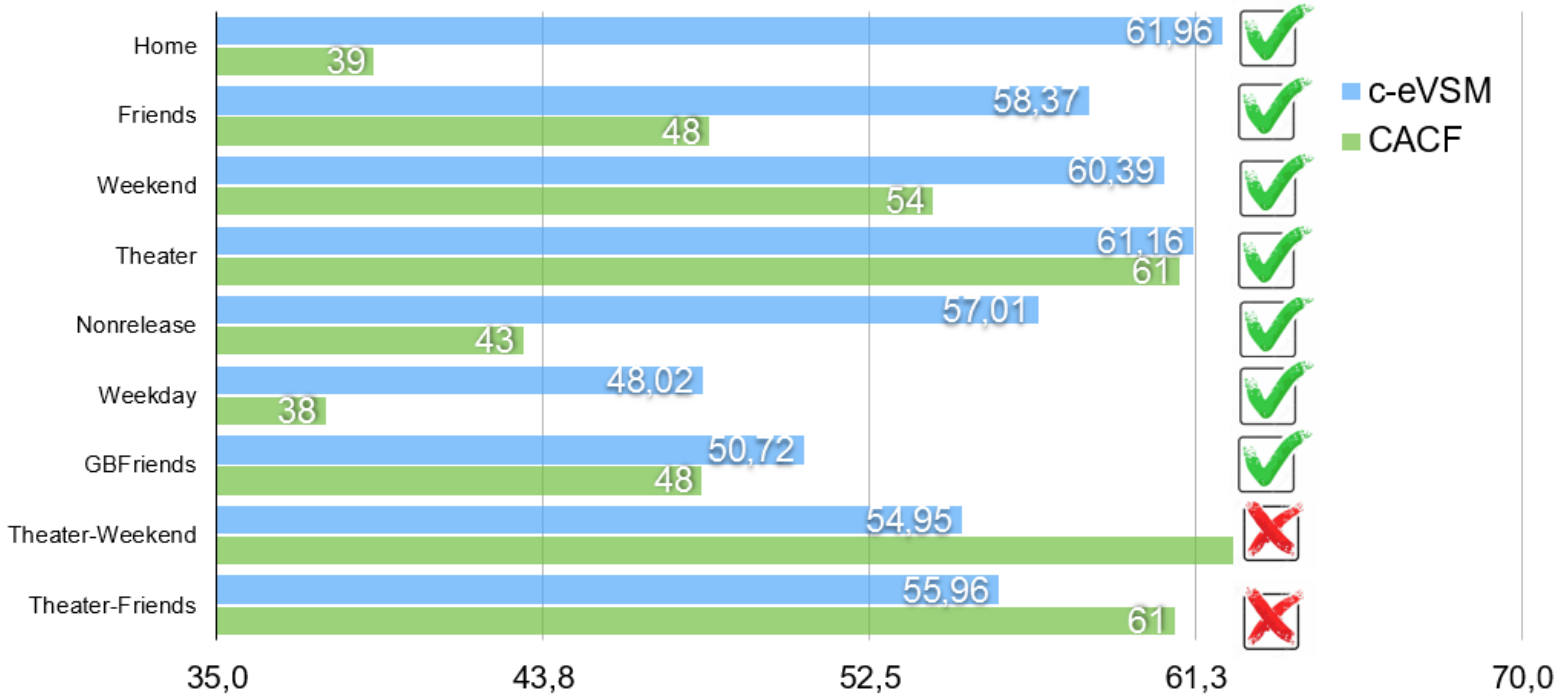
Entities are better than simple keywords!

Musto, Cataldo, et al. "Combining distributional semantics and entity linking for context-aware content-based recommendation." *International Conference on User Modeling, Adaptation, and Personalization*. UMAP 2014

C-eVSM

Selection of Results

Comparison to context-aware CF algorithm



Compared to Context-aware Collaborative Filtering (CACF)
[*] algorithm: better in 7 contextual segments

[*] G.Adomavicius et al. , Incorporating contextual information in recommender systems using a multi-dimensional approach. ACM Trans. Inf. Systems, 2005

Semantic representations

Explicit (Exogenous) Semantics

Implicit (Endogenous) Semantics

Introduce semantics **by mapping the features describing** the item with semantic **concepts**

Introduce semantics **by linking the Item to a knowledge graph**

Distributional semantic models

Explicit Semantic Analysis

Random Indexing

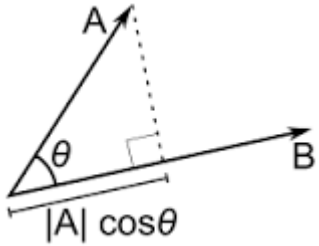
Word2Vec

……

ESA effectively used for

Text Categorization [Gabri09]
experiments on diverse datasets

Semantic relatedness of words and texts [Gabri09]
cosine similarity between vectors of ESA concepts



Information Retrieval [Egozi08, Egozi11]
ESA-based IR algorithm enriching documents and queries

what about ESA for Information Filtering?

[Gabri09] E. Gabrilovich and S. Markovitch. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research* 34:443-498, 2009.

[Egozi08] Ofer Egozi, Evgeniy Gabrilovich, Shaul Markovitch: Concept-Based Feature Generation and Selection for Information Retrieval. *AAAI 2008*, 1132-1137, 2008.

[Egozi11] Ofer Egozi, Shaul Markovitch, Evgeniy Gabrilovich. Concept-Based Information Retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems* 29(2), April 2011.

Information Filtering using ESA

TV-domain

German Electronic Program Guides (EPG)

- better **precision** [Musto12]

Social data

from **Facebook** and **Twitter**

- better **serendipity** [Narducci13]
- i.e. more unexpected and interesting recommendations

Multimedia recommendation

TED lectures

- better results in a **cold-start** setting [Pappas14]

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

[Narducci13] F. Narducci, C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. Leveraging Encyclopedic Knowledge for Transparent and Serendipitous User Profiles. UMAP 2013, pp. 350-352.

[Pappas14] N. Pappas and A. Popescu-Belis. Combining content with user preferences for non-fiction multimedia recommendation: A study on ted lectures. Multimedia Tools and Applications, 2014.

Electronic Program Guides

problem

description of TV shows **too short** or
poorly meaningful to feed a
content-based recommendation algorithm

solution

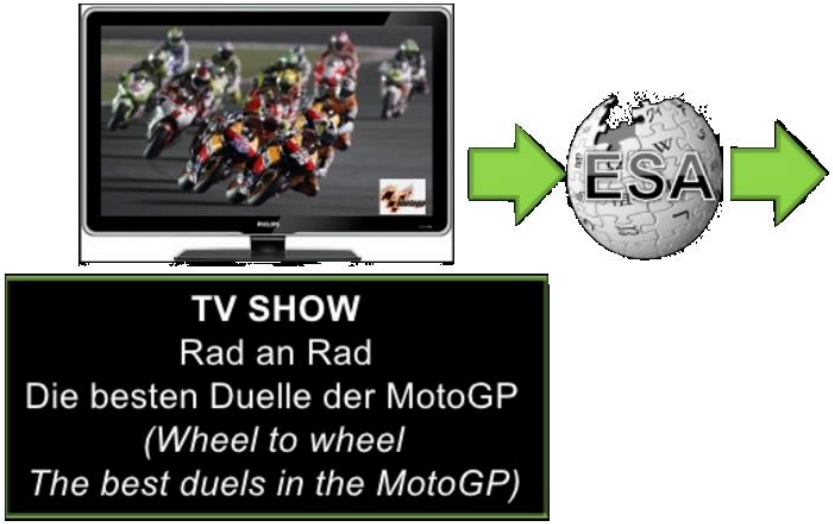
Explicit Semantic Analysis exploited to obtain an
enhanced representation

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. **Enhanced semantic tv-show representation for personalized electronic program guides.** UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

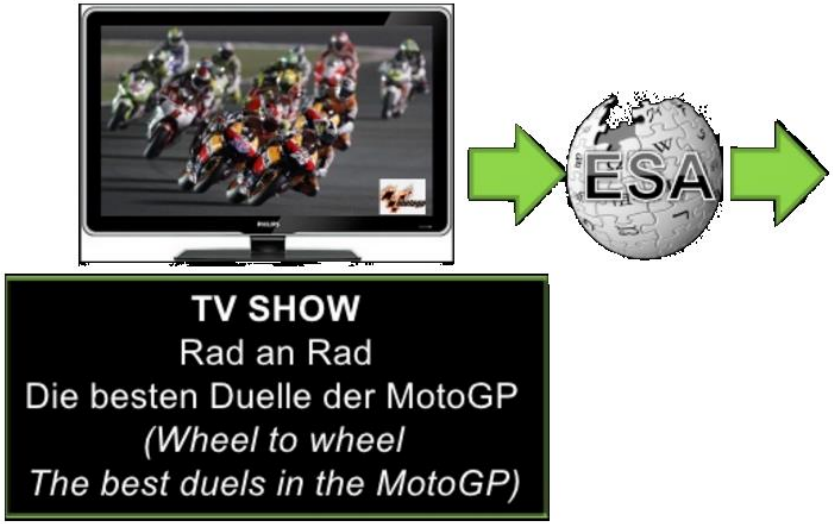
Electronic Program Guides



[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

Wikipedia Articles related to the TV show **are added to the description**



Wikipedia Articles
großer preis von italien (motorrad)
großer preis von malaysia (motorrad)
großer preis von tschechien (motorrad)
scuderia ferrari
valentino rossi
motorrad-wm-saison 2005
motorrad-wm-saison 2006
max biaggi
großer preis der usa (motorrad)
motorrad-wm-saison 2008
rad (heraldik)
loris capirossi
shin'ya nakano
motogp

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
sports
motorbike
...
competition

tv show

2012 Superbike
Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
sports
motorbike
...
competition

No matching!

tv show

2012 Superbike Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
superbike
sports
motorbike
formula 1
…
competition

Through ESA we can add new features to the profile **and we can improve the overlap between textual description**

tv show

2012 Superbike
Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

user profile

motogp
superbike
sports
motorbike
formula 1
…
competition

Matching!

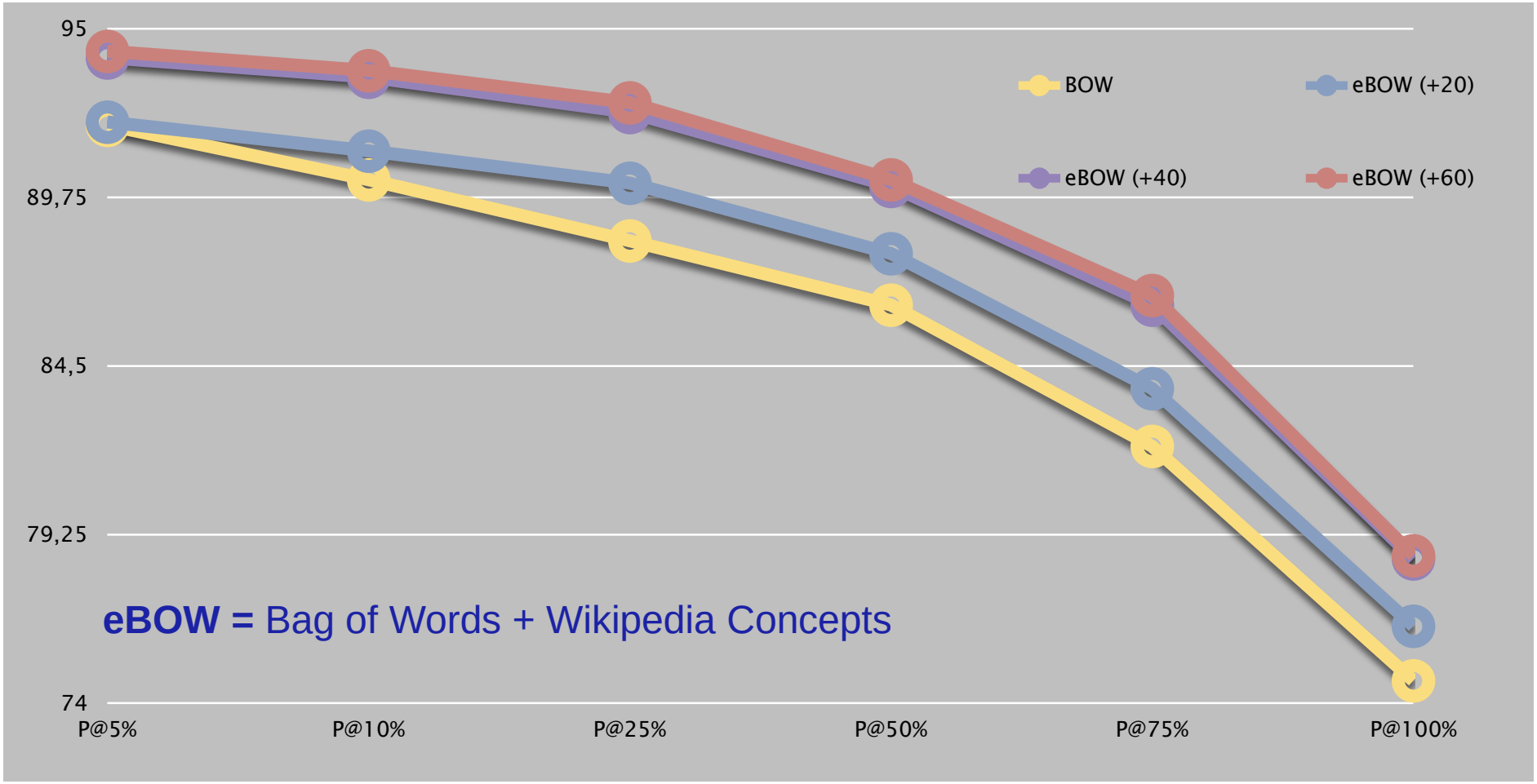
tv show

2012 Superbike
Italian Grand Prix

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

Electronic Program Guides

results on Aprico.tv data



The more Wikipedia Concepts are added to the textual description of the items (eBOW+60), **the best the precision of the algorithm**

Cross-lingual information access

Cross-lingual access: problems

Vocabulary mismatch

use of **different languages**

extreme case of **vocabulary mismatch**

L'Italia prepara la sfida impossibile
Con la Francia un miracolo non basta

SIAULIAI (Lituania), 3 settembre 2011

Gli azzurri devono battere Parker e soci, ancora senza sconfitte, con almeno 13 punti di scarto per continuare a sperare. Il c.t. Pianigiani ammette: "Non valiamo le prime 10 d'Europa"

Gli azzurri festeggiano la vittoria sulla Lettonia. Ansa

PIÙ LETTI | PIÙ COMMENTATI

Juve, la notte dell'orgoglio, Il nuovo stadio emoziona

Moratti: "Vicini a Gasperini, Il caso Forlan è grave"

Milan, la macchina da gol , contro Klose e Cisse

Portieri: sarà l'anno di Mirante?, Mutu, il riscatto

87 | 10 | Mi piace | Tweet | 19 | 1 | DI LA TUA | +1 | Share

Invia articolo
Versione stampabile
Ascolta

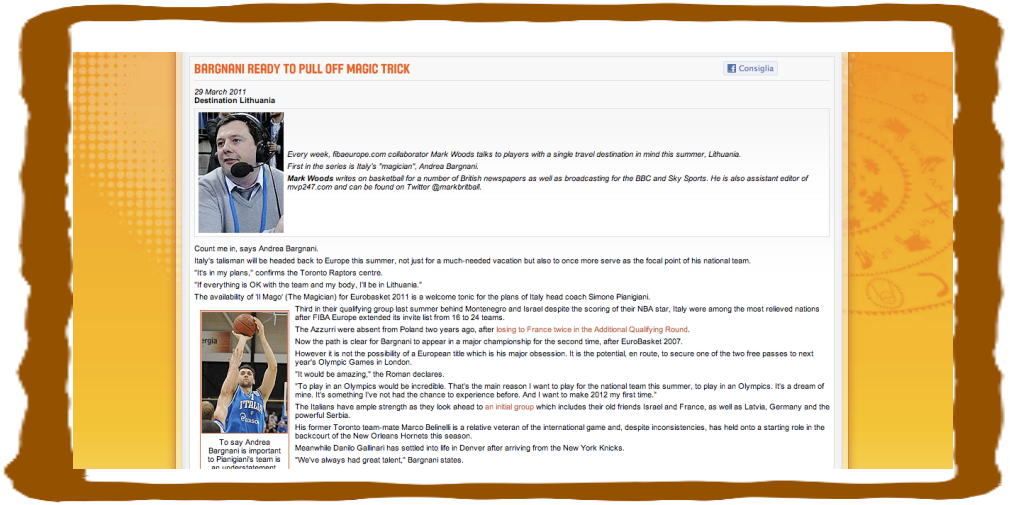
PER SAPERNE DI PIÙ
Risultati e classifiche

MAGIC LIBRO 2011
Non perdere nemmeno un colpo all'asta d'inizio anno! A soli 7,99 € in edicola

EXTRA MAGIC CHAMPIONS
Una squadra più forte di Barça e Red Devils? Tu puoi crearla. Gioca gratis e vinci!

LE GRANDI STORIE DELL'AUTO
Porta le grandi storie dell'auto sempre con te!Scarica subito per iPad e iPhone a SOLI 2.99€.

LINOMANIA



BARGNANI READY TO PULL OFF MAGIC TRICK

Consiglia

29 March 2011
Destination Lithuania

Every week, fibaeurope.com collaborator Mark Woods talks to players with a single travel destination in mind this summer, Lithuania. First in the series is Italy's "magician", Andrea Bargnani.

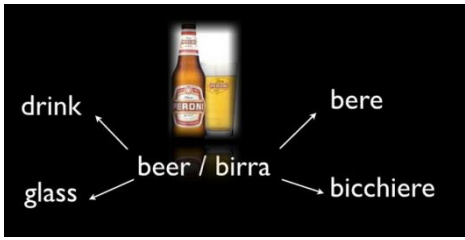
***Mark Woods** writes on basketball for a number of British newspapers as well as broadcasting for the BBC and Sky Sports. He is also assistant editor of mvp247.com and can be found on Twitter @markbritball.*

Count me in, says Andrea Bargnani.
Italy's talisman will be headed back to Europe this summer, not just for a much-needed vacation but also to once more serve as the focal point of his national team.
"It's in my plans," confirms the Toronto Raptors centre.
"If everything is OK with the team and my body, I'll be in Lithuania."
The availability of 'Il Mago' (The Magician) for Eurobasket 2011 is a welcome tonic for the plans of Italy head coach Simone Pianigiani.

Third in their qualifying group last summer behind Montenegro and Israel despite the scoring of their NBA star, Italy were among the most relieved nations after FIBA Europe extended its invite list from 16 to 24 teams.
The Azzurri were absent from Poland two years ago, after losing to France twice in the Additional Qualifying Round.
Now the path is clear for Bargnani to appear in a major championship for the second time, after EuroBasket 2007.
However it is not the possibility of a European title which is his major obsession. It is the potential, en route, to secure one of the two free passes to next year's Olympic Games in London.
"It would be amazing," the Roman declares.
"To play in an Olympics would be incredible. That's the main reason I want to play for the national team this summer, to play in an Olympics. It's a dream of mine. It's something I've not had the chance to experience before. And I want to make 2012 my first time."
The Italians have ample strength as they look ahead to an initial group which includes their old friends Israel and France, as well as Latvia, Germany and the powerful Serbia.
His former Toronto team-mate Marco Belinelli is a relative veteran of the international game and, despite inconsistencies, has held onto a starting role in the backcourt of the New Orleans Hornets this season.
Meanwhile Danilo Gallinari has settled into life in Denver after arriving from the New York Knicks.
"We've always had great talent," Bargnani states.

To say Andrea Bargnani is important to Pianigiani's team is

(Cross-lingual) Concept-based representations



MultiWordNet

Sense-based representations

Word Sense Disambiguation (JIGSAW) based on **Multiwordnet** as **sense repository**

multilingual lexical database that supports English, Italian, Spanish, Portuguese, Hebrew, Romanian, Latin

alignment between synsets in the different languages

semantic relations imported and preserved

MultiWordNet

MultiWordNet ID	Synset	Language	Gloss
n02418562	*chair*		a seat for one person, with a support for the back
	sedia		oggetto su cui ci si siede
	silla		
n78346641	*professorship, chair*		the position of professor
	cattedra, ordinariato, professorato		il posto del professore di ruolo
	cátedra		

Bag of MultiWordNet synsets

Italian news

L'Italia prepara la sfida impossibile
Con la Francia un miracolo non basta

SIAULIAI (Lituania), 3 settembre 2011

Gli azzurri devono battere Parker e soci, ancora senza sconfitte, con almeno 13 punti di scarto per continuare a sperare. Il c.t. Pianigiani ammette: "Non valiamo le prime 10 d'Europa"

Gli azzurri festeggiano la vittoria sulla Lettonia. Ansa

87 10 Mi piace Tweet 19 1 DI LA TUA +1 Share

Invia articolo
Versione stampabile
Ascolta

PER SAPERNE DI PIÙ
Risultati e classifiche

PIÙ LETTI | PIÙ COMMENTATI

Juve, la notte dell'orgoglio, Il nuovo stadio emoziona
Moratti: "Vicini a Gasperini, Il caso Forlan è grave"
Milan, la macchina da gol , contro Klose e Cisse
Portieri: sarà l'anno di Mirante?, Mutu, il riscatto

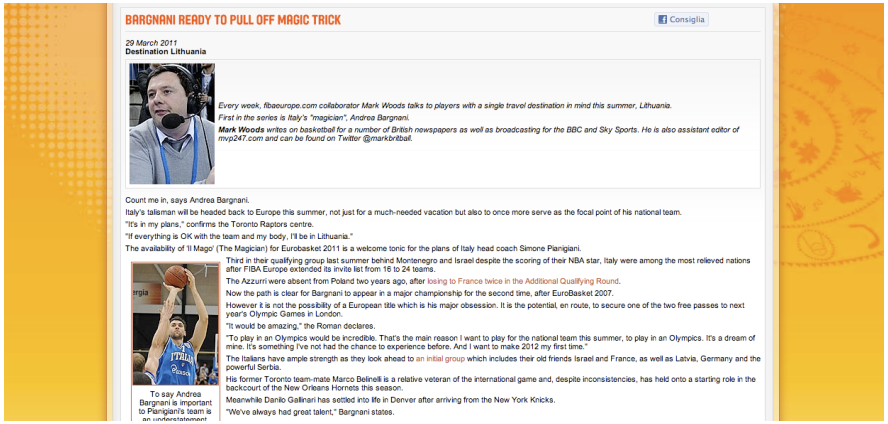
MAGIC LIBRO 2011
Non perdere nemmeno un colpo all'asta d'inizio anno! A soli 7,99 € in edicola

EXTRA MAGIC CHAMPIONS
Una squadra più forte di Barça e Red Devils? Tu puoi crearla. Gioca gratis e vinci!

LE GRANDI STORIE DELL'AUTO
Porta le grandi storie dell'auto sempre con te!Scarica subito per iPad e iPhone a SOLI 2.99€.

LINOMANIA

French News



BARGNANI READY TO PULL OFF MAGIC TRICK

Consiglia

29 March 2011
Destination Lithuania

Every week, fibaeurope.com collaborator Mark Woods talks to players with a single travel destination in mind this summer, Lithuania.
First in the series is Italy's "magician", Andrea Bargnani.
***Mark Woods** writes on basketball for a number of British newspapers as well as broadcasting for the BBC and Sky Sports. He is also assistant editor of mvp247.com and can be found on Twitter @markbritball.*

Count me in, says Andrea Bargnani.
Italy's talisman will be headed back to Europe this summer, not just for a much-needed vacation but also to once more serve as the focal point of his national team.
"It's in my plans," confirms the Toronto Raptors centre.
"If everything is OK with the team and my body, I'll be in Lithuania."
The availability of 'Il Mago' (The Magician) for Eurobasket 2011 is a welcome tonic for the plans of Italy head coach Simone Pianigiani.

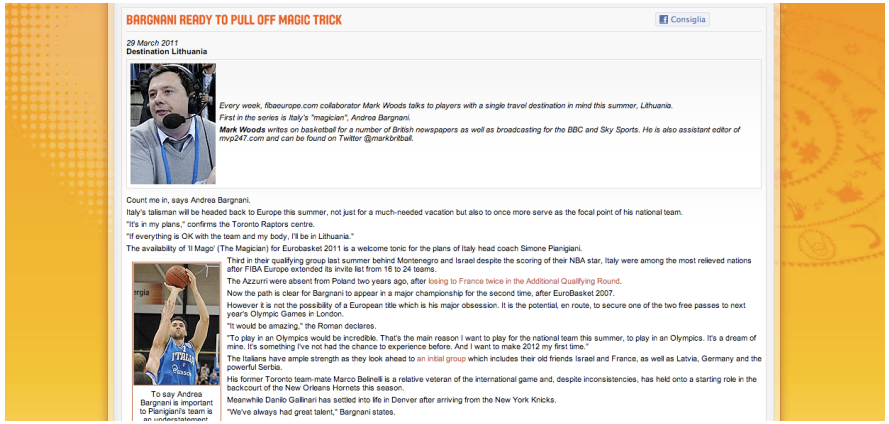
To say Andrea Bargnani is important to Pianigiani's team is

Third in their qualifying group last summer behind Montenegro and Israel despite the scoring of their NBA star, Italy were among the most relieved nations after FIBA Europe extended its invite list from 16 to 24 teams.
The Azzurri were absent from Poland two years ago, after losing to France twice in the Additional Qualifying Round.
Now the path is clear for Bargnani to appear in a major championship for the second time, after EuroBasket 2007.
However it is not the possibility of a European title which is his major obsession. It is the potential, en route, to secure one of the two free passes to next year's Olympic Games in London.
"It would be amazing," the Roman declares.
"To play in an Olympics would be incredible. That's the main reason I want to play for the national team this summer, to play in an Olympics. It's a dream of mine. It's something I've not had the chance to experience before. And I want to make 2012 my first time."
The Italians have ample strength as they look ahead to an initial group which includes their old friends Israel and France, as well as Latvia, Germany and the powerful Serbia.
His former Toronto team-mate Marco Belinelli is a relative veteran of the international game and, despite inconsistencies, has held onto a starting role in the backcourt of the New Orleans Hornets this season.
Meanwhile Danilo Gallinari has settled into life in Denver after arriving from the New York Knicks.
"We've always had great talent," Bargnani states.

Bag of MultiWordNet synsets

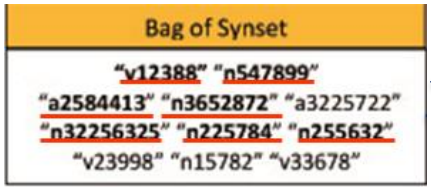
Italian news

French News



bag of MultiWordNet synsets

Match between senses



bag of MultiWordNet synsets

Some results

cross-language movie recommendation scenario

profiles learned from **ENG/ITA** descriptions
recommendation provided on **ITA/ENG** descriptions
MovieLens dataset, **F1** measure, **Wikipedia** source for descriptions

Profiles \ Recommendations		
	64.91	63.98
	63.70	63.71

P. Lops, C. Musto, F. Narducci, M. de Gemmis, P. Basile, G. Semeraro: Cross-Language Personalization through a Semantic Content-Based Recommender System. D. Dicheva, D. Dochev (Eds.): Artificial Intelligence: Methodology, Systems, and Applications, 14th In. Conference, AIMSA 2010. LNCS 6304 Springer 2010, pp.52-60

Cross-lingual representation: Tagme

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American-Australian science fiction action film written and directed by The Wachowski Brothers, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix

Tagged text | **Topics**

The Matrix | Science fiction film | Action film | Screenwriter | Film director | The Wachowskis | Keanu Reeves | Laurence Fishburne | Carrie-Anne Moss | Joe Pantoliano | Hugo Weaving | Dystopia | Perception | Human | Simulated reality | Cyberspace

Cross-lingual representation: Tagme

The Matrix

From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see The Matrix (franchise). For other uses, see Matrix (disambiguation).

The Matrix is a 1999 American-Australian science fiction action film written and directed by The Wachowski Brothers, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix

Tagged text | **Topics**

The Matrix | Science fiction film | Action film | Screenwriter | Film director | The Wachowskis | Keanu Reeves | Laurence Fishburne | Carrie-Anne Moss | Joe Pantoliano | Hugo Weaving | Dystopia | Perception | Human | Simulated reality | Cyberspace

The Wachowskis

From Wikipedia, the free encyclopedia

Lana Wachowski (formerly **Laurence "Larry" Wachowski**, born June 21, 1965)[4] and **Lilly Wachowski** (formerly **Andrew Paul "Andy" Wachowski**, born December 29, 1967)[5] are sibling American film directors, screenwriters, and producers.[6] They are both openly transgender women.[7][8][9][10] Known together professionally as **The Wachowskis**[11] and formerly as **The Wachowski Brothers**, the pair made their directing debut in 1996 with *Bound*, and reached fame with their second film *The Matrix* (1999), a major box office success for which they won the Saturn Award for Best Director. They wrote and directed its two sequels: *The Matrix Reloaded* and *The Matrix Revolutions* (both in 2003), and were deeply involved in the writing and production of other works in the franchise.

Following the commercial success of *The Matrix* series, they wrote and produced the 2006 film *V for Vendetta* (an adaptation of the comic of the same name by Alan Moore), and in 2008 released the film *Speed Racer*, which was a live-action adaptation of the Japanese anime series of the same name. Their next film, *Cloud Atlas*, based on the novel of the same

Lana Wachowski

Lilly (then known as Andy) and Lana Wachowski at the Fantastic Fest screening of *Cloud Atlas* in September 2012

Born	Laurence Wachowski June 21, 1965 (age 51) Chicago, Illinois, United States
Other names	Larry Wachowski (before 2010)
Occupation	Film and TV director,

Keanu Reeves

From Wikipedia, the free encyclopedia
(Redirected from Keanu reeves)

This article is about the Canadian actor. For the Philippine actress, see Keanna Reeves.
"Keanu" redirects here. For other uses, see Keanu (disambiguation).

Keanu Charles Reeves (/keɪˈɑːnuː/ *kay-AH-noo*[citation needed]; born September 2, 1964)[1][2] is a Canadian actor, producer, director and musician.

Reeves is best known for his acting career, beginning in 1985 and spanning more than three decades. He gained fame for his starring role performances in several blockbuster films including comedies from the *Bill and Ted franchise* (1989–1991), action thrillers *Point Break* (1991) and *Speed* (1994), and the science fiction-action trilogy *The Matrix* (1999–2003). He has also appeared in dramatic films such as *Dangerous Liaisons* (1988), *My Own Private Idaho* (1991), and *Little Buddha* (1993), as well as the romantic horror *Bram Stoker's Dracula* (1992).

Since becoming active in the film industry, Reeves' abilities have earned critical acclaim. One *New York Times* critic praised Reeves' versatility, saying that he "displays considerable discipline and range. He moves easily between the buttoned-down demeanor that suits a police procedural story and the loose-jointed manner of his comic roles."[3]

Keanu Reeves

Reeves at the 2014 Fantastic Fest

Born	Keanu Charles Reeves

Laurence Fishburne

From Wikipedia, the free encyclopedia

Laurence John Fishburne III[1] (born July 30, 1961) is an American actor, playwright, director and producer, best known for playing Morpheus in *The Matrix trilogy*, Jason "Furious" Styles in the 1991 drama film *Boyz n the Hood* and Tyrone "Mr. Clean" Miller in the 1979 war film *Apocalypse Now*.

Fishburne became the first African-American to portray Othello in a motion picture by a major studio when he appeared in Oliver Parker's 1995 film adaptation of the Shakespeare play. For his portrayal of Ike Turner in *What's Love Got to Do With It*, he was nominated for an Academy Award for Best Actor. He won a Tony Award for Best Featured Actor in a Play for his performance in *Two Trains Running* (1992), and an Emmy Award for Drama Series Guest Actor for his performance in *TriBeCa* (1993).

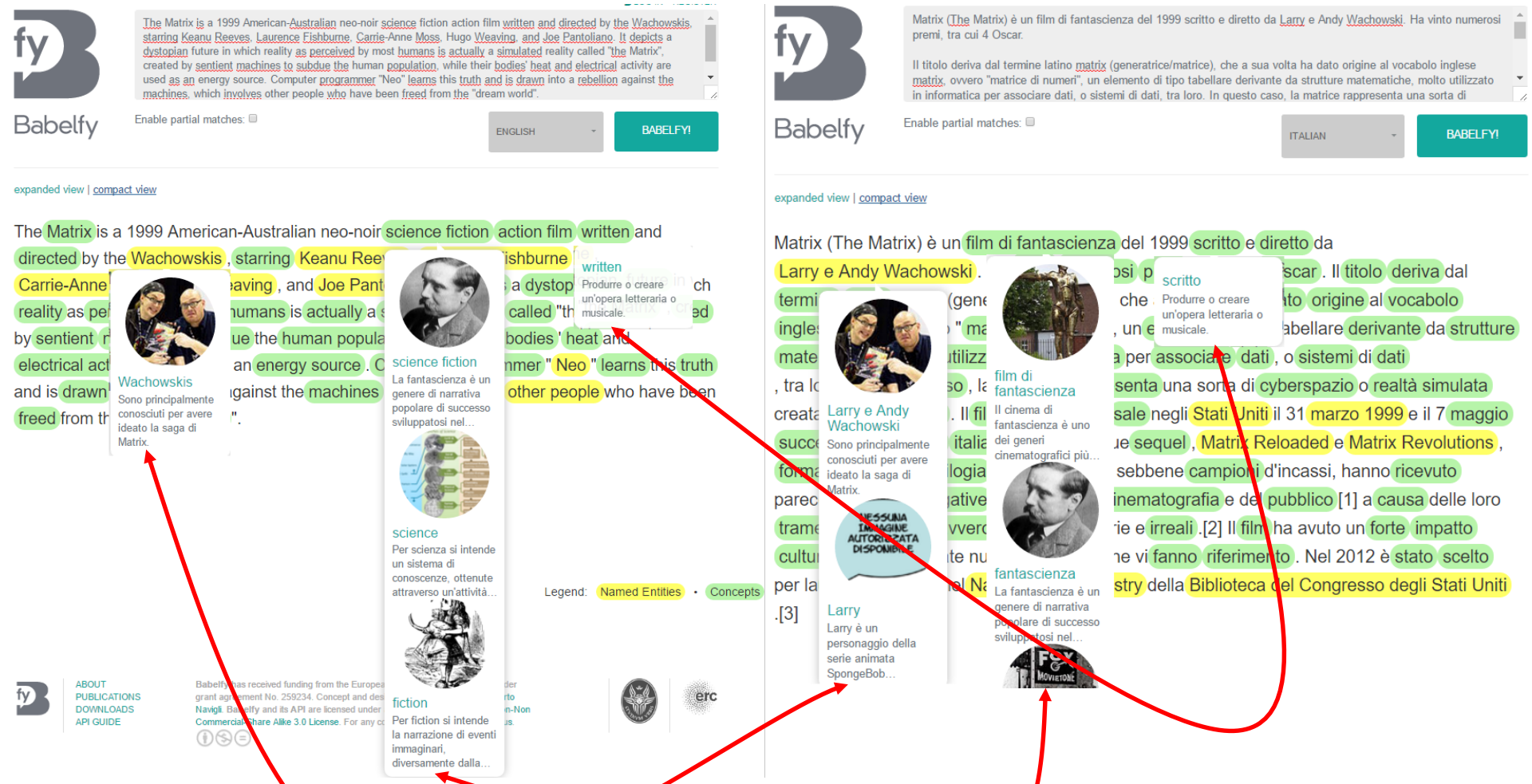
Fishburne starred in several cult classics, including *Deep Cover* and *King of New York*. From 2008 to 2011, he starred as Dr. Raymond Langston on the CBS crime drama *CSI: Crime Scene Investigation* and from 2013 to 2015 starred as Special Agent Jack Crawford on the NBC thriller

Laurence Fishburne

Fishburne in September 2009

Born	Laurence John Fishburne III July 30, 1961 (age 54) Augusta, Georgia, U.S.

Cross-lingual representation: Babelfy



Cross-language links

https://en.wikipedia.org/wiki/The_Wachowskis

Not logged in Talk Contributions Create account Log in

https://en.wikipedia.org/wiki/The_Wachowskis

Languages

العربية
Aragonés
Azərbaycanca
বাংলা
Български
Bosanski
Català
Čeština
Cymraeg
Dansk
Deutsch
Eesti
Ελληνικά
Español
Esperanto
Euskara
فارسی
Français
Galego
한국어
Հայերեն
Hrvatski
Ido
Italiano
עברית
ქართული
Latina

https://it.wikipedia.org/wiki/The_Wachowskis

Accesso non effettuato discussioni contributi Registrati Entra

WIKIPEDIA
L'enciclopedia libera

Pagina principale
Ultime modifiche
Una voce a caso
Vetrina
Aiuto

Comunità
Portale Comunità
Bar
Il Wikipediano
Fai una donazione
Contatti

Strumenti
Puntano qui
Modifiche correlate
Carica su Commons
Pagine speciali
Link permanente
Informazioni sulla pagina
Elemento Wikidata
Cita questa voce

Voce Discussione Leggi Modifica Modifica wikitesto Cronologia Ricerca

Incontra i wikipediani dal 24 al 26 Giugno a Wikimania 2016 a Esino Lario!
Ogni dettaglio sulla pagina dell'evento
[Dai una mano con le traduzioni!]

Lana e Lilly Wachowski

Da Wikipedia, l'enciclopedia libera.

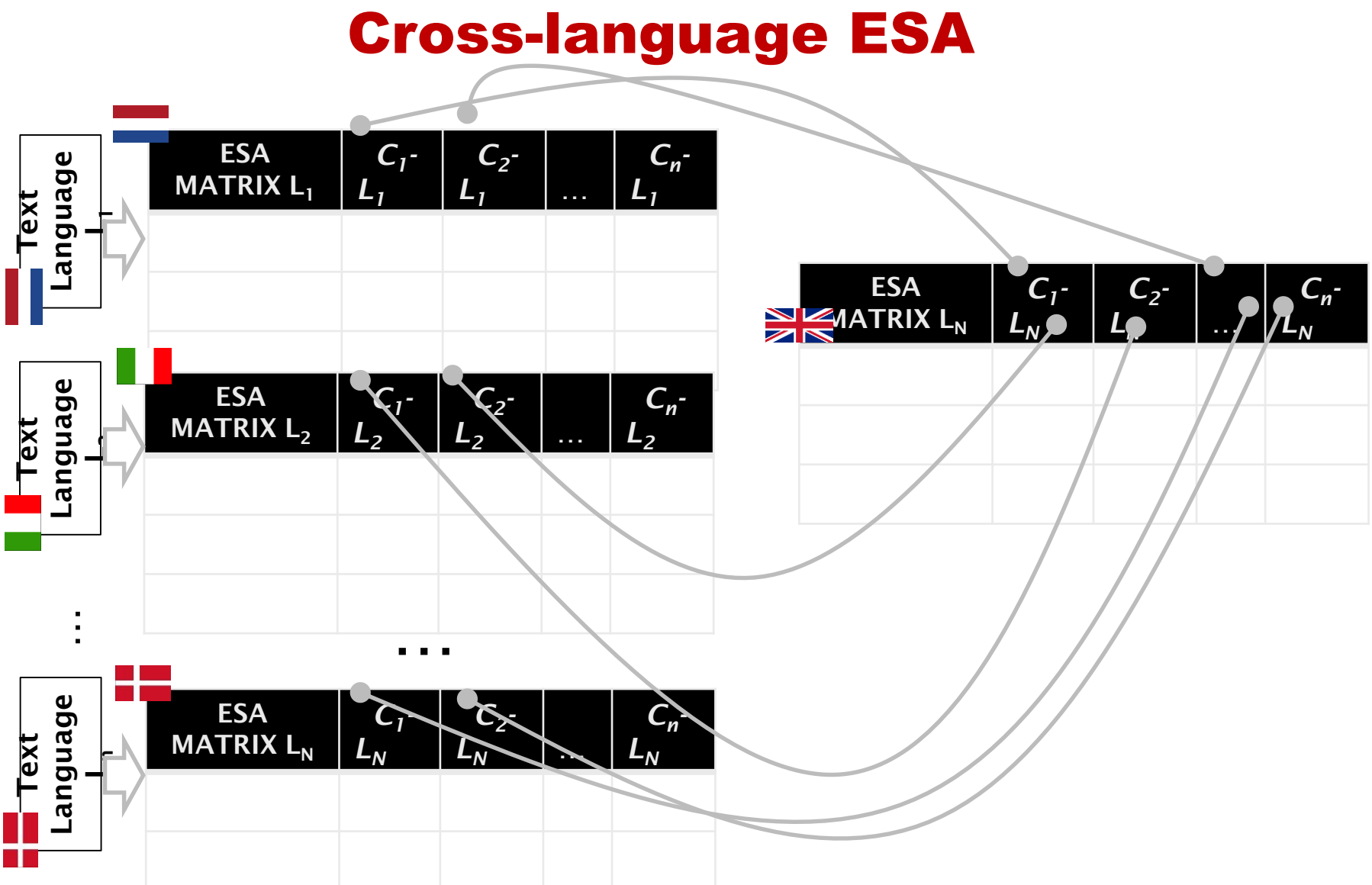
Questa voce o sezione deve essere rivista e aggiornata appena possibile.
Sembra infatti che questa voce contenga informazioni superate e/o obsolete. Se puoi, contribuisci ad aggiornarla.

Lana Wachowski, nata **Laurence** e nota un tempo come **Larry** (Chicago, 21 giugno 1965), e **Lilly Wachowski**, nata **Andrew Paul** e nota un tempo come **Andy** (Chicago, 29 dicembre 1967), sono due sorelle transgender cineaste statunitensi, un tempo note come **Fratelli Wachowski**, principalmente conosciute per avere ideato la saga di *Matrix*.

Indice [nascondi]
1 Biografia
2 Stile
3 Fumetti
4 Controversie
5 Filmografia

Lilly (a sinistra) e Lana Wachowski al Fantastic Fest nel 2012, per la presentazione di *Cloud Atlas*.

Cross-lingual representation: ESA



Preliminary results

effectiveness of **knowledge-based strategies** to provide **cross-lingual recommendations**

MovieLens and **DBbook** datasets

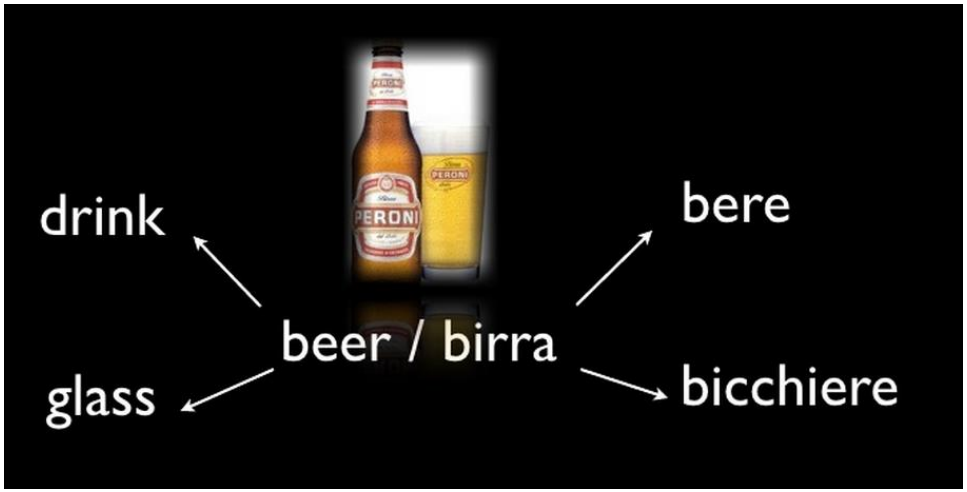
F1 measure, **Wikipedia** source for descriptions

English and **Italian** languages

Representation	MovieLens	DBbook
Keywords	51.70	47.50
TAGME	48.60	49.50
	50.90	50.70
Babelfy	54.10	53.10

Cross-lingual representation: Distributional models

distribution of the terms
(almost) the same in **different languages**

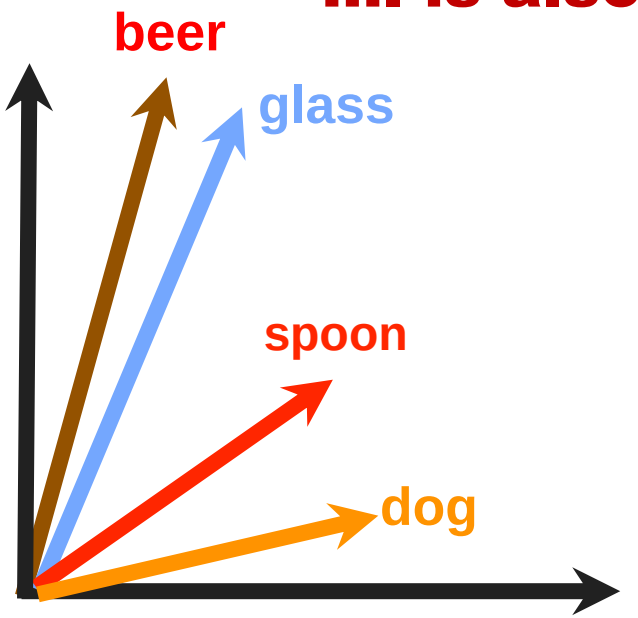


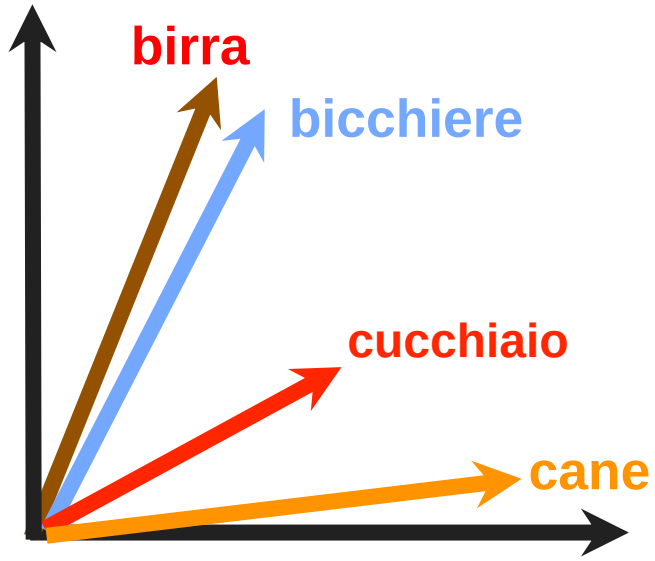
- **cross-lingual representation comes with no costs** thanks to the distributional hypothesis

C. Musto, F. Narducci, P. Basile, P. Lops, M. de Gemmis, G. Semeraro: Cross-Language Information Filtering: Word Sense Disambiguation vs. Distributional Models. AI*IA 2011: 250-261

Distributional Semantics

.... is also multilingual! (Recap)



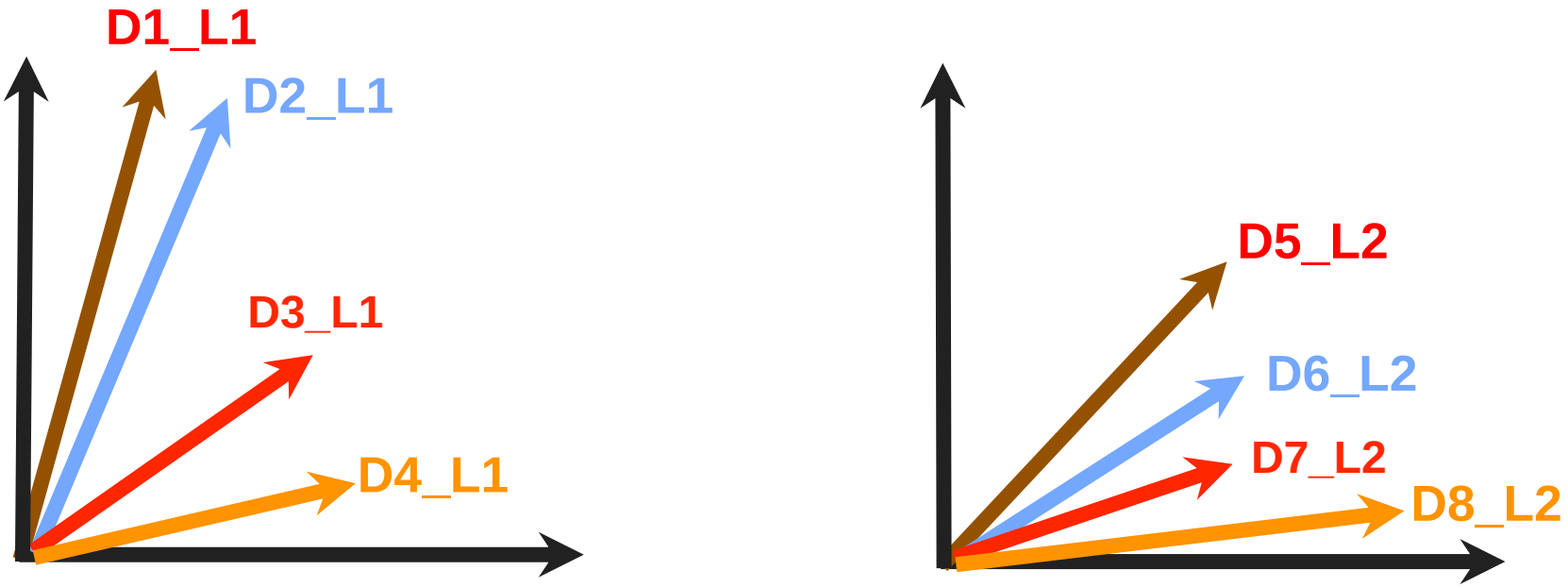


English WordSpace **Italian WordSpace**

The position in the space can be slightly different, but the relations similarity between terms **still hold**

Distributional Semantics

Multilingual DocSpace



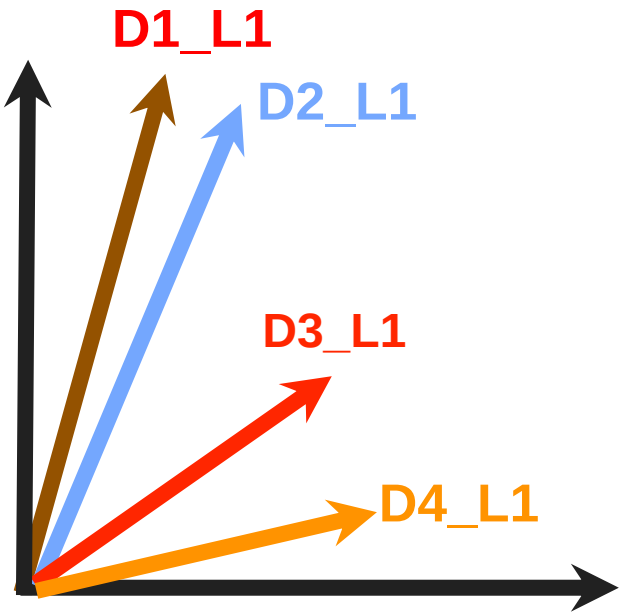
English WordSpace **Italian WordSpace**

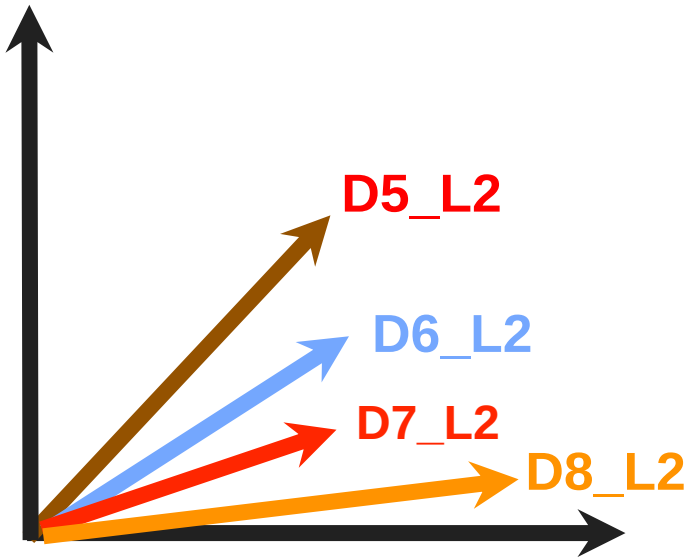
By following the same procedure we can obtain a multilingual DocSpace

Different documents in different languages are represented in a uniform space

Distributional Semantics

.... is also multilingual!





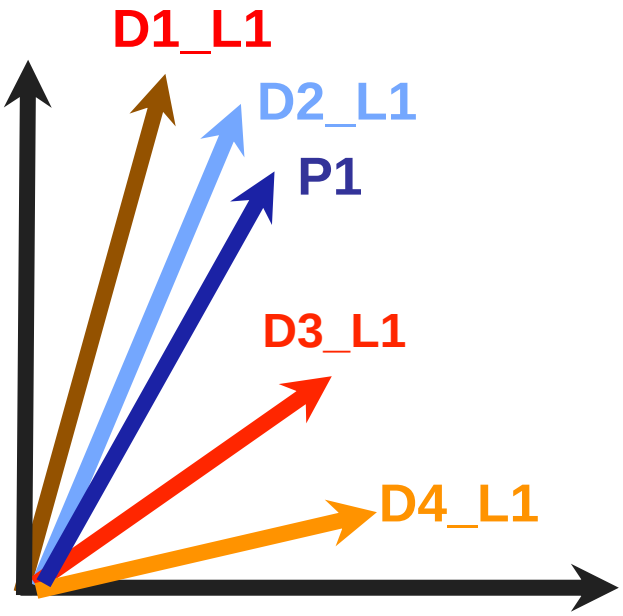
English WordSpace **Italian WordSpace**

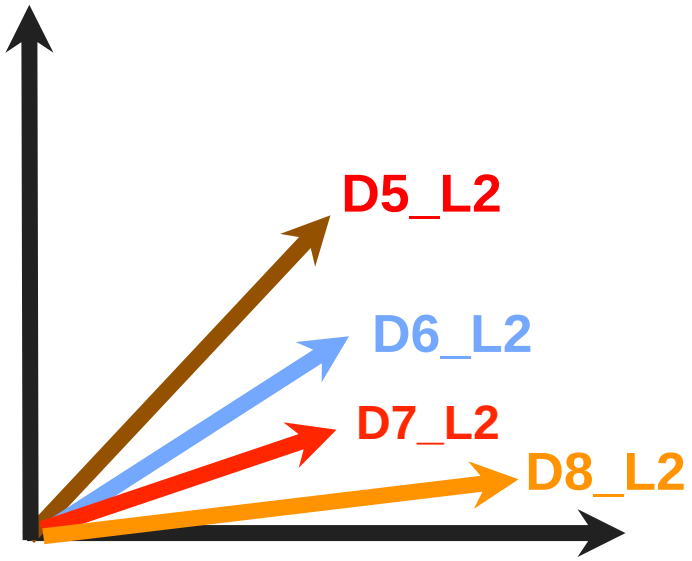
By following the same procedure we can obtain a multilingual DocSpace

How to build a cross-lingual recommender?

Distributional Semantics

.... is also multilingual!





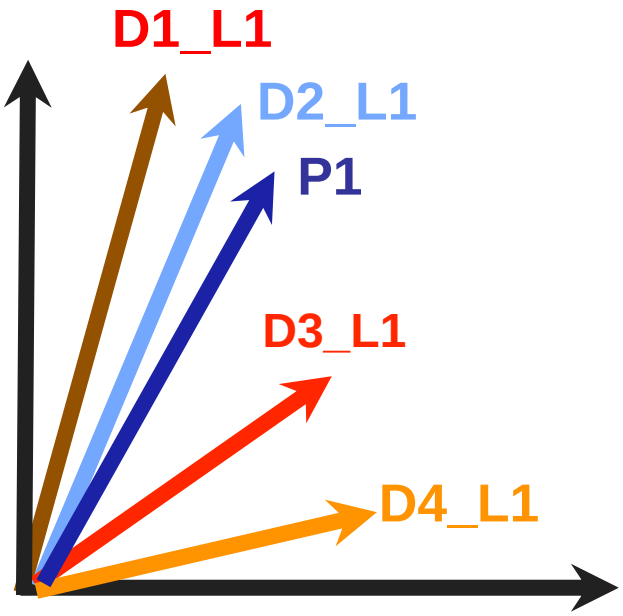
English WordSpace **Italian WordSpace**

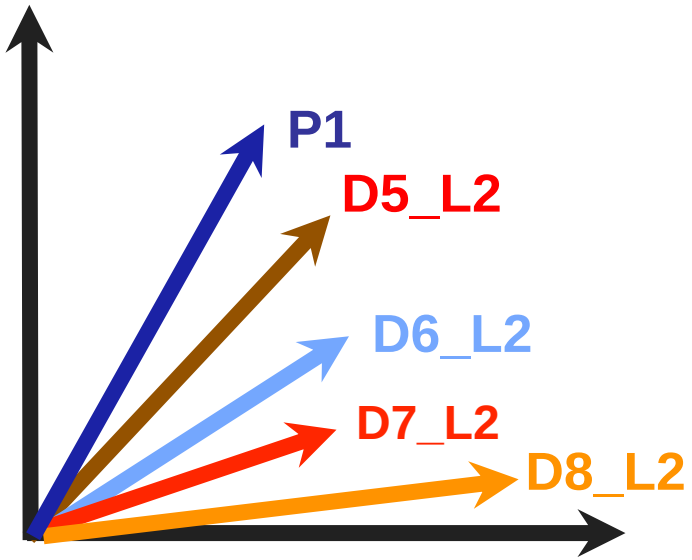
How to build a cross-lingual recommender?

We build a user profile in L1 (Italian DocSpace)

Distributional Semantics

.... is also multilingual!





English WordSpace **Italian WordSpace**

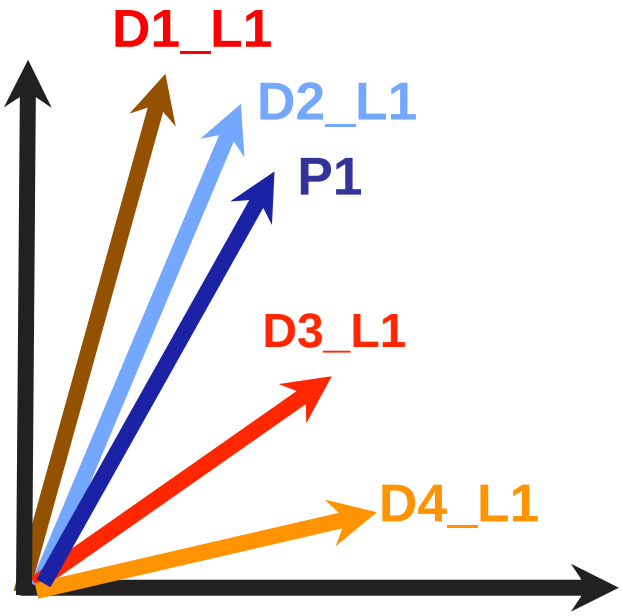
How to build a cross-lingual recommender?

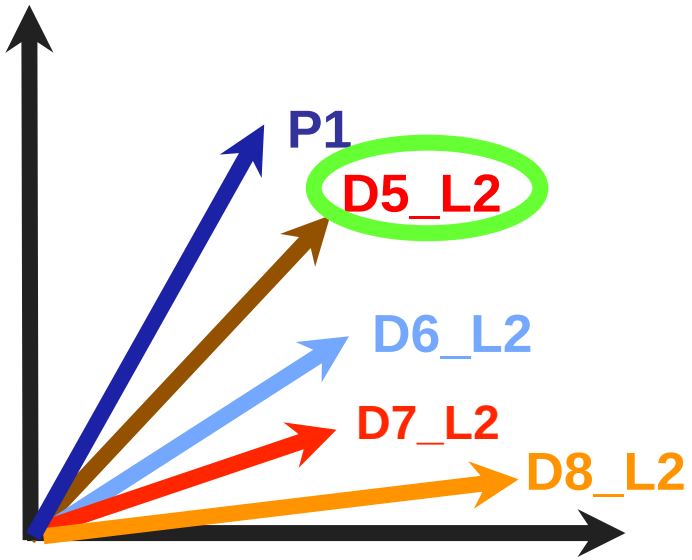
We build a user profile in L1 (English DocSpace)

We can «move» the profile in L2 (Italian DocSpace)

Distributional Semantics

.... is also multilingual!





English WordSpace **Italian WordSpace**

How to build a cross-lingual recommender?

We build a user profile in L1 (English DocSpace)

We can «move» the profile in L2 (Italian DocSpace)

We can use similarity measures to suggest items in different language

Some results

effectiveness of **knowledge-based strategies** to provide **cross-lingual recommendations**

MovieLens dataset

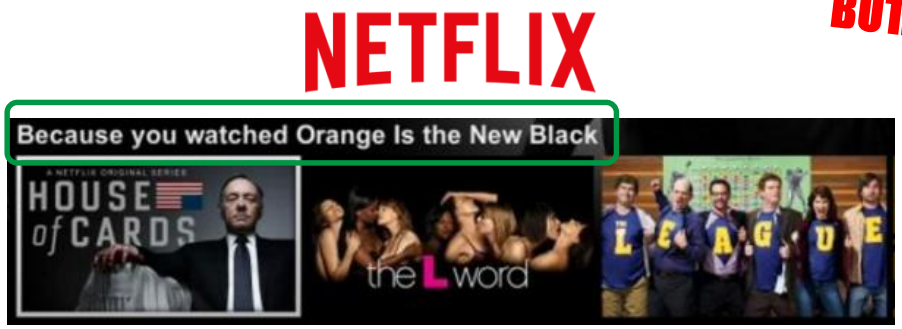
F1 measure

	Distributional Models	Bayes classifier + Synsets
 → 	84.65	85.61
 → 	84.63	85.20

comparable results (gap not statistically significant)

C. Musto, F. Narducci, P. Basile, P. Lops, M. de Gemmis, G. Semeraro:**"Cross-language information filtering: Word sense disambiguation vs. distributional models."** *AI*IA 2011: 250-261*

A different application: Explanation in Recommender Systems



Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco de Gemmis, Giovanni Semeraro: **ExpLOD: A Framework for Explaining Recommendations based on the Linked Open Data Cloud.** RecSys 2016: 151-154

Problem

I Suggest you…

Recommendation

Problem

?

Recommendation

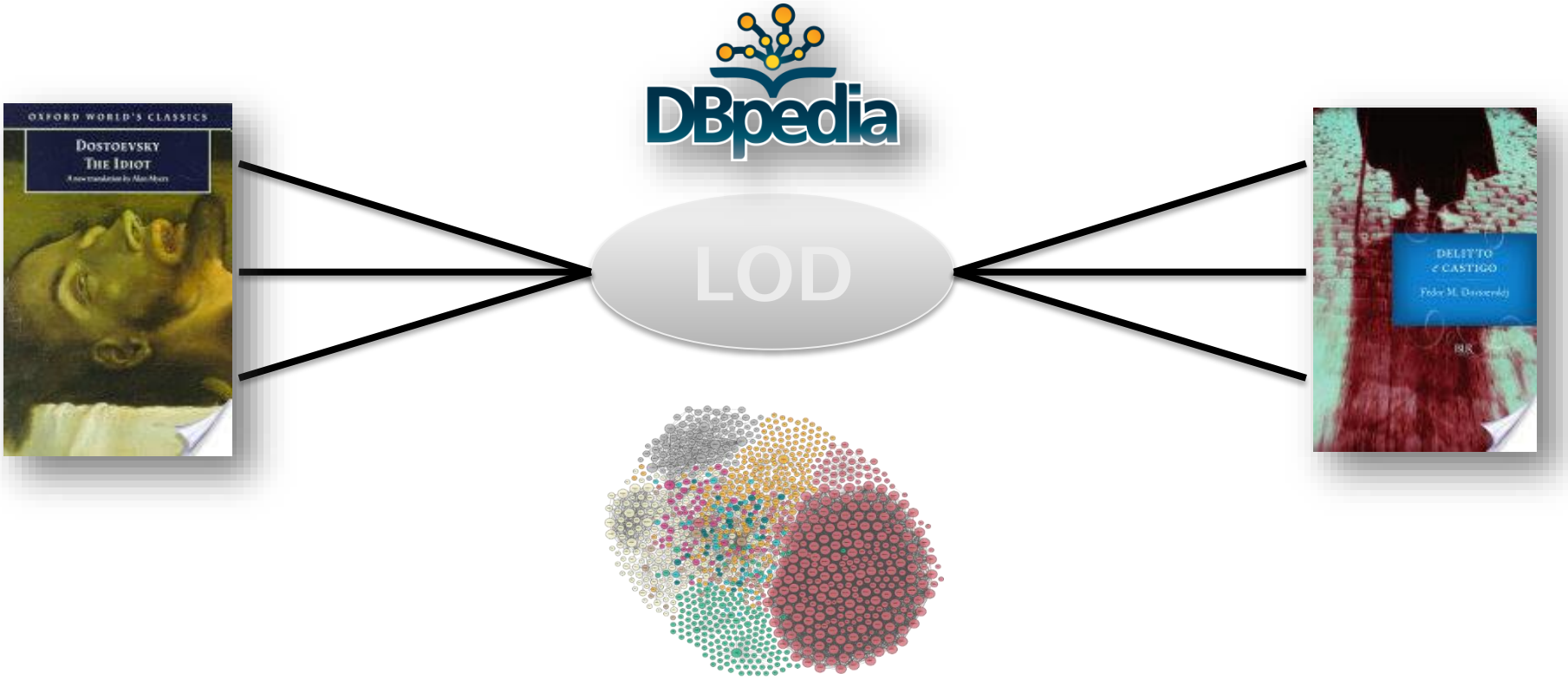
A possible solution

I suggest Crime and Punishment
because
you like books written by Fyodor
Dostoevskij
as The Brothers Karamazov.
Furthermore, you often like
Psychological Russian Novels as Anna
Karenina and
War Stories.

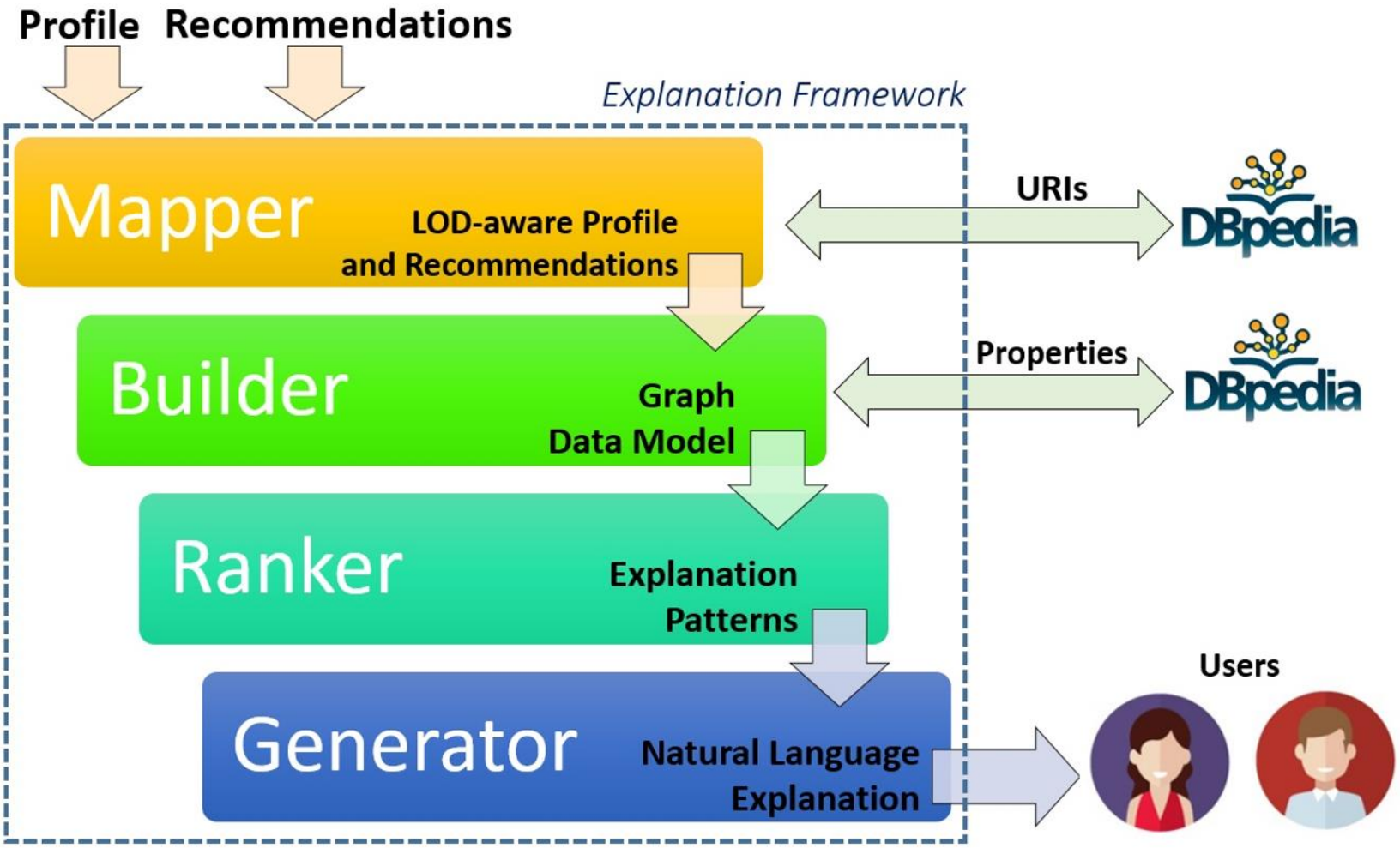
Recommendation

Insight

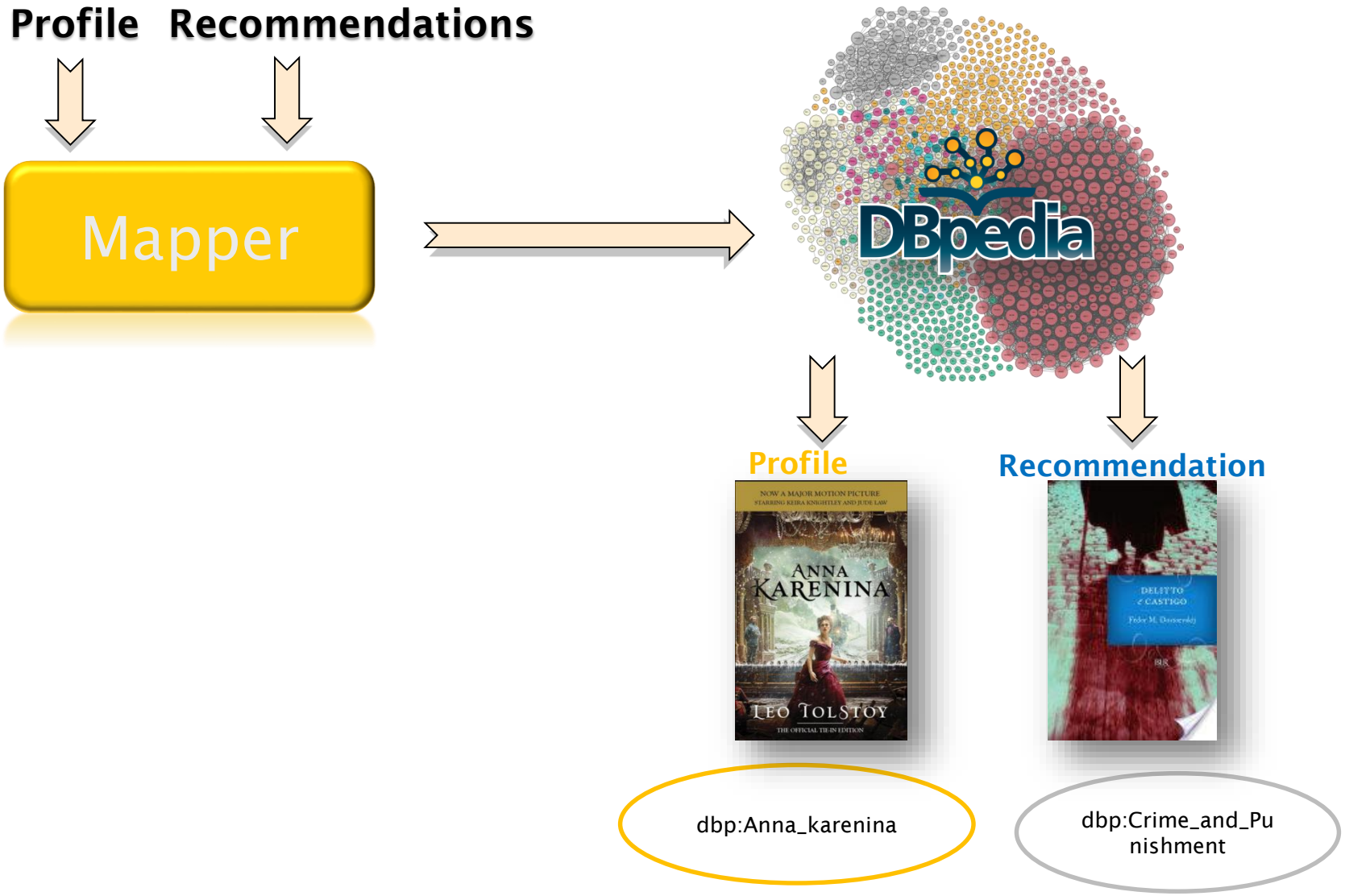
To connect the properties **describing the items the user** liked to those **describing the recommendations** via the properties **available in the LOD cloud**



EXPLOD: Framework

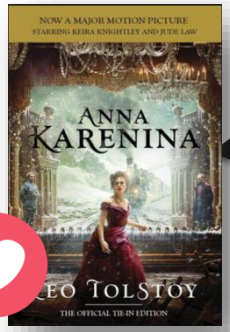


EXPLOD: Mapper

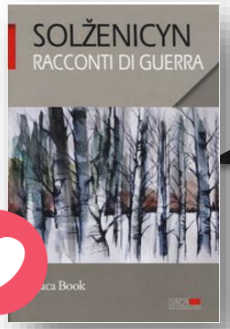


EXPLOD: Builder

dbp:The_Brothers_Karamazov



dbp:Anna_karenina



dbp:War_stories

author

author

genre

genre

genre

genre

subject

subject

Fyodor Dostoevskij

dbp:Fyodor_Dostoyevsky

Philosophical Fiction

dbp:Category:Philosophical_fiction

Psychological Russian Novel

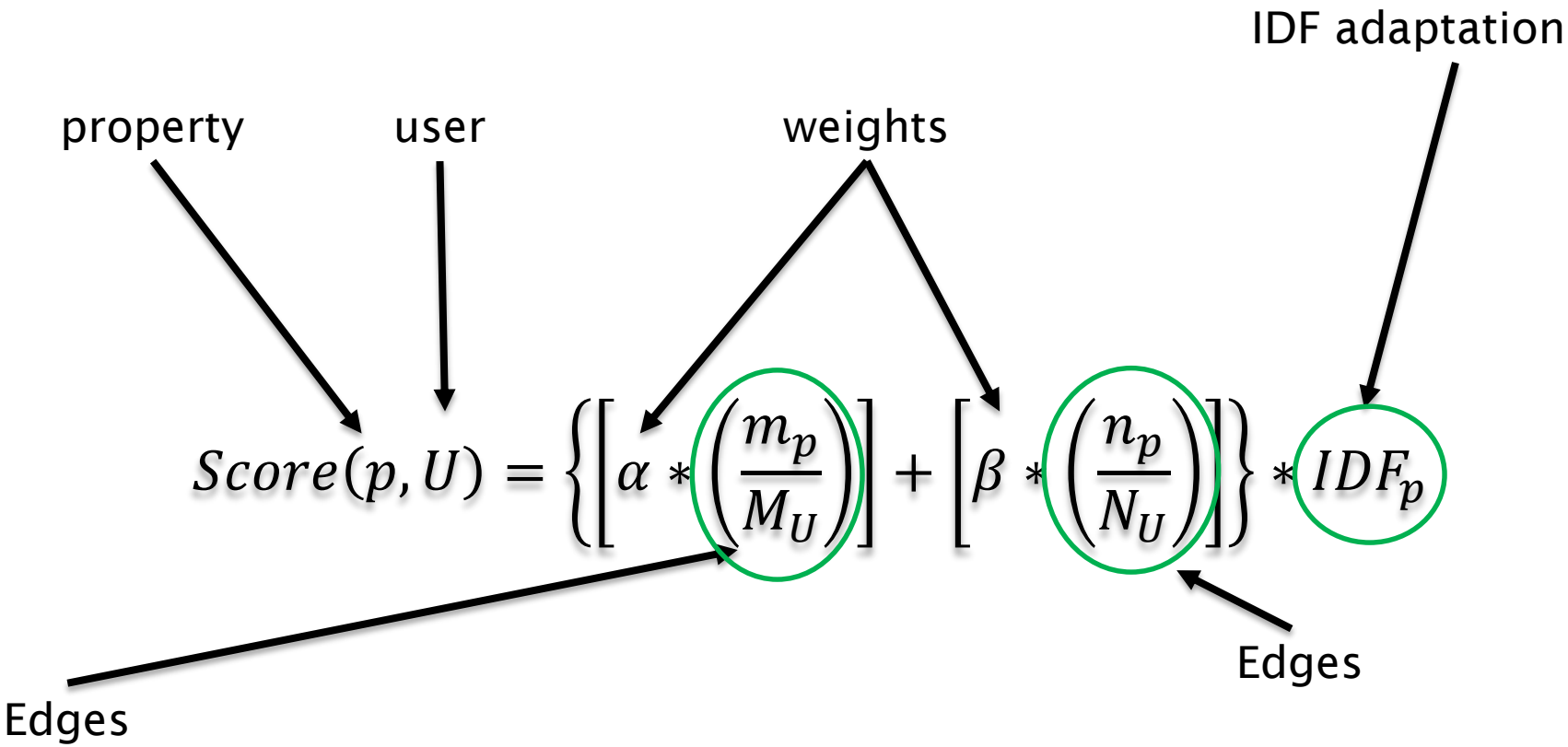
dbp:Category:Psycological_novel

Narcisism In Fiction

dbp:Category:Narcissism_in_fiction

dbp:Crime_and_Punishment

EXPLOD: Ranker



- Properties in the graph **are ranked**

EXPLOD: Ranker

dbp:The_Brothers_Karamazov

author

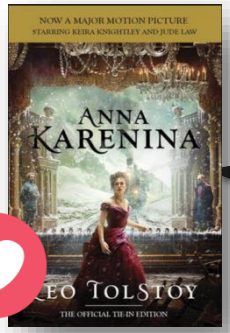
Fyodor Dostoevskij: Score: 4,875

dbp:Fyodor_Dostoyevsky

author

Recommendation

dbp:Crime_and_Punishment



dbp:Anna_karenina

genre

Philosophical Fiction Score: 3,534

dbp:Category:Philosophical_fiction

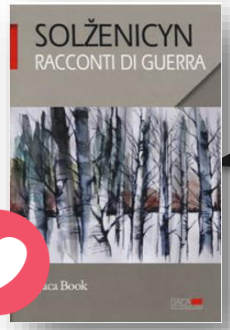
genre

genre

Psychological Russian Novel Score: 3,231

dbp:Category:Psycological_novel

genre



dbp:War_stories

genre

subject

Narcisism In Fiction Score: 3,132

dbp:Category:Narcissism_in_fiction

subject

EXPLOD: Ranker

Returns the top-3 properties

dbp:The_Brothers_Karamazov

author

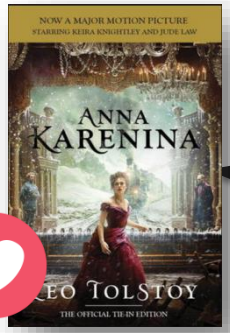
Fyodor Dostoevskij: Score: 4,875

dbp:Fyodor_Dostoyevsky

author

Recommendation

dbp:Crime_and_Punishment



dbp:Anna_karenina

genre

Philosophical Fiction Score: 3,534

dbp:Category:Philosophical_fiction

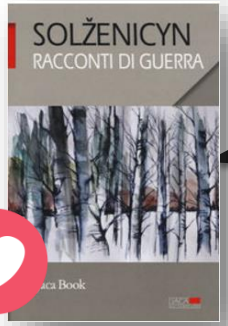
genre

genre

Psychological Russian Novel Score: 3,231

dbp:Category:Psycological_novel

genre



dbp:War_stories

genre

subject

Narcisism In Fiction Score: 3,152

dbp:Category:Narcissism_in_fiction

subject

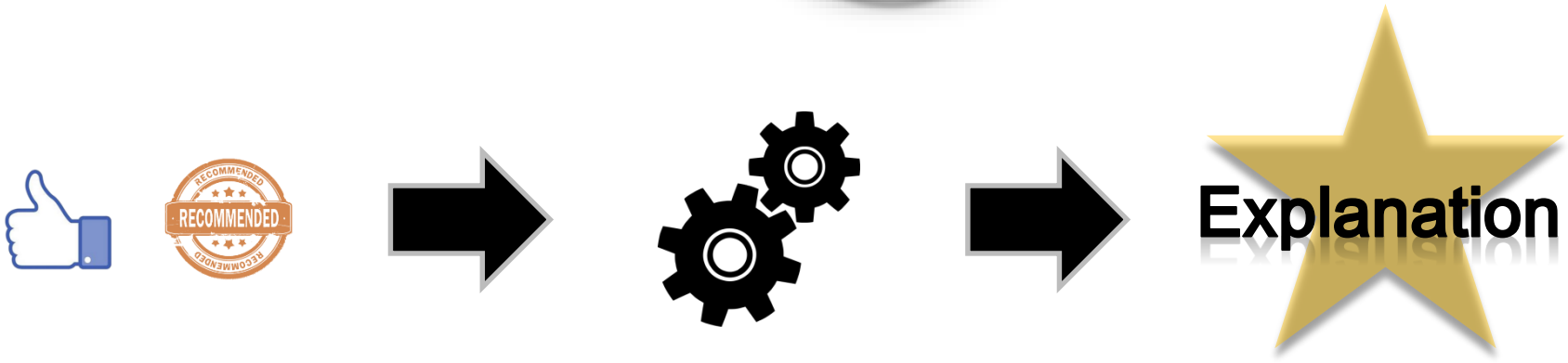
EXPLOD: Generator

Input:

- User Profile
- Recommended Items
- Top-k properties

Output:

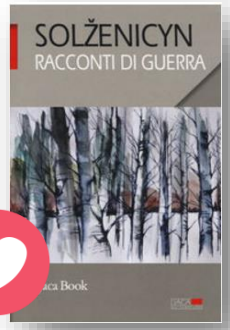
- Natural Language Explanation



EXPLOD: Generator

dbp:The_Brothers_Karamazov

dbp:Anna_karenina



dbp:War_stories

I suggest Crime and Punishment...

Recommendation

dbp:Crime_and_Punishment

EXPLOD: Generator

Item profilo

dbp:The_Brothers_Karamazov

author

Fyodor Dostoevskij: Score: 4,875

dbp:Fyodor_Dostoyevsky

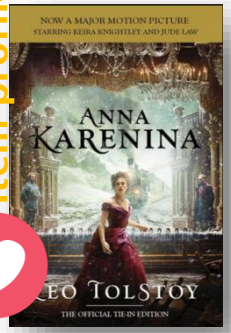
Top-1 property

author

Recommendation

dbp:Crime_and_Punishment

Item profilo



dbp:Anna_karenina

Item profilo

dbp:War_stories

I suggest Crime and Punishment because you like books written by Fyodor Dostoevskij as The Brothers Karamazov.

EXPLOD: Generator

dbp:The_Brothers_Karamazov

Top-2 properties

author

Fyodor Dostoevskij: Score: 4,875

dbp:Fyodor_Dostoyevsky

author

Recommendation

dbp:Crime_and_Punishment

dbp:Anna_karenina

genre

Philosophical Fiction Score: 3,534

dbp:Category:Philosophical_fiction

genre

dbp:War_stories

I suggest **Crime and Punishment** because you like books **written by Fyodor Dostoevskij** as **The Brothers Karamazov**. Furthermore, you like **Philosophical Fiction**, as **Anna Karenina**.

EXPLOD: Generator

dbp:The_Brothers_Karamazov

author

Fyodor Dostoevskij: Score: 4,875

dbp:Fyodor_Dostoyevsky

Top-3 Properties

author

Recommendation

dbp:Crime_and_Punishment

dbp:Anna_karenina

genre

Philosophical Fiction Score: 3,534

dbp:Category:Philosophical_fiction

genre

genre

Psycological Russian Novel Score: 3,231

dbp:Category:Psycological_novel

genre

genre

Item profilo

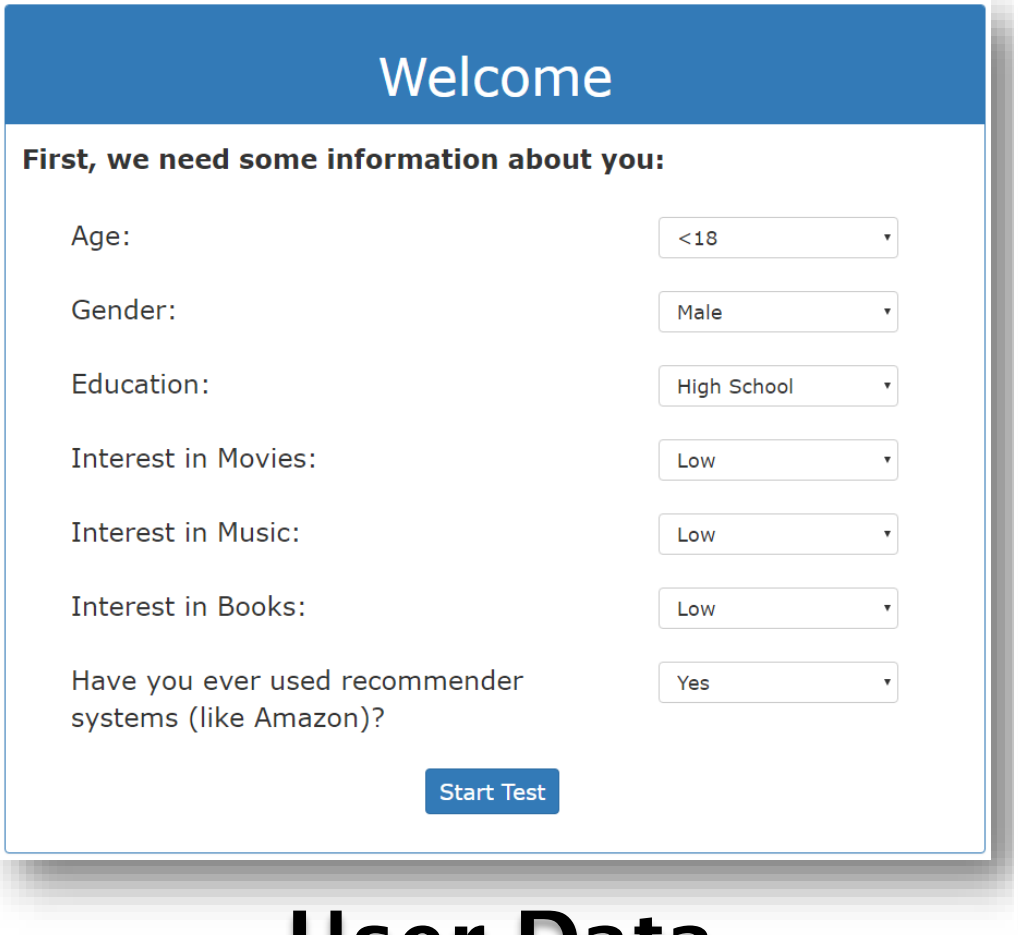
dbp:War_stories

I suggest **Crime and Punishment** because you like books **written by Fyodor Dostoevskij** as **The Brothers Karamazov**. Furthermore, you like **Philosophical Fiction**, as **Anna Karenina**. Finally, you **often** like **Psychological Russian Novel**, as **Anna Karenina** and **War Stories**.

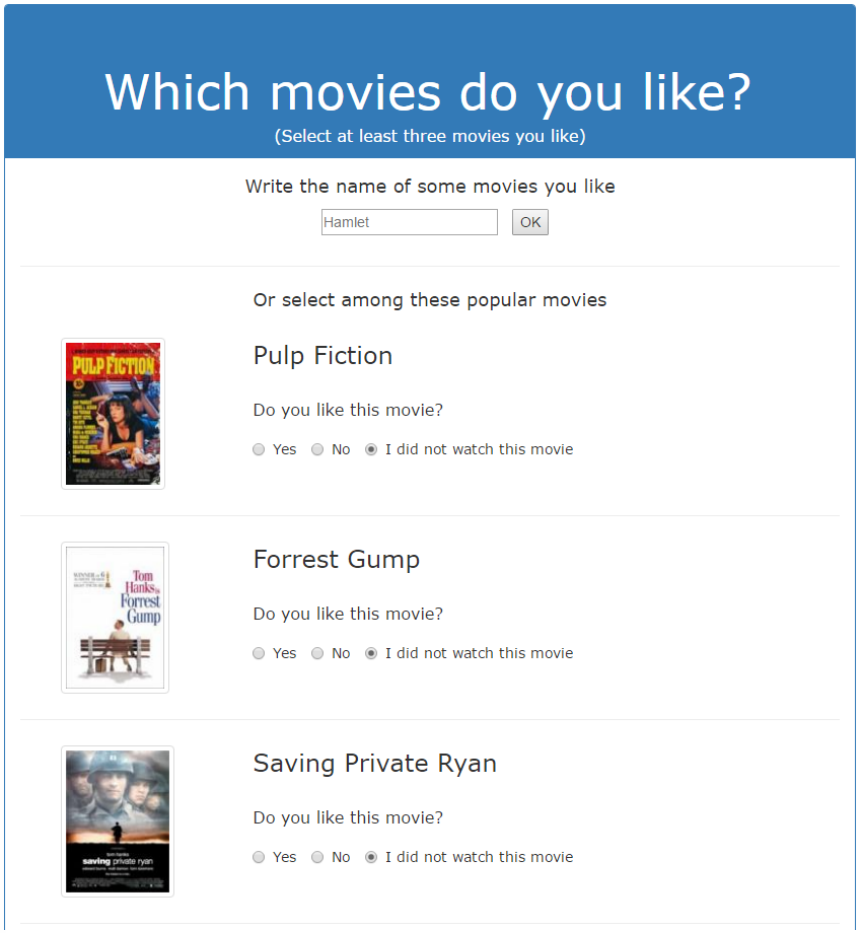
Experimental Evaluation

- **User Study**
  - ✓ Movie Domain, 306 users involved
  - ✓ **Protocol:**
    - Web Application → Building User Profiles → Recommendations + Explanations → Questionnaire + Ex-post Evaluation
  - ✓ **Explanation Aims**
    - Transparency, Engagement, Persuasion, Trust, Effectiveness
- **Three configurations compared**
  - ✓ Popularity-based Explanation (baseline)
  - ✓ Non-Personalized Explanation based on LOD
  - ✓ EXPLOD

Web Application



User Data



User Profile

Web Application

Recommendation for you

Iron Man 2

That's my explanation:

I suggest you **Iron Man 2** because you sometimes like *movies produced by Cinema of Southern California*, as **Pulp Fiction**, **The Shining** and **Iron Man**.

Besides, you sometimes like *Films shot in the United States*, as **The Shining**.

Finally, you sometimes like *Science fiction action films*, as **Iron Man**.

Rate this recommendation (read the explanation first!)

Questionnaire:

I understood why this movie was recommended to me

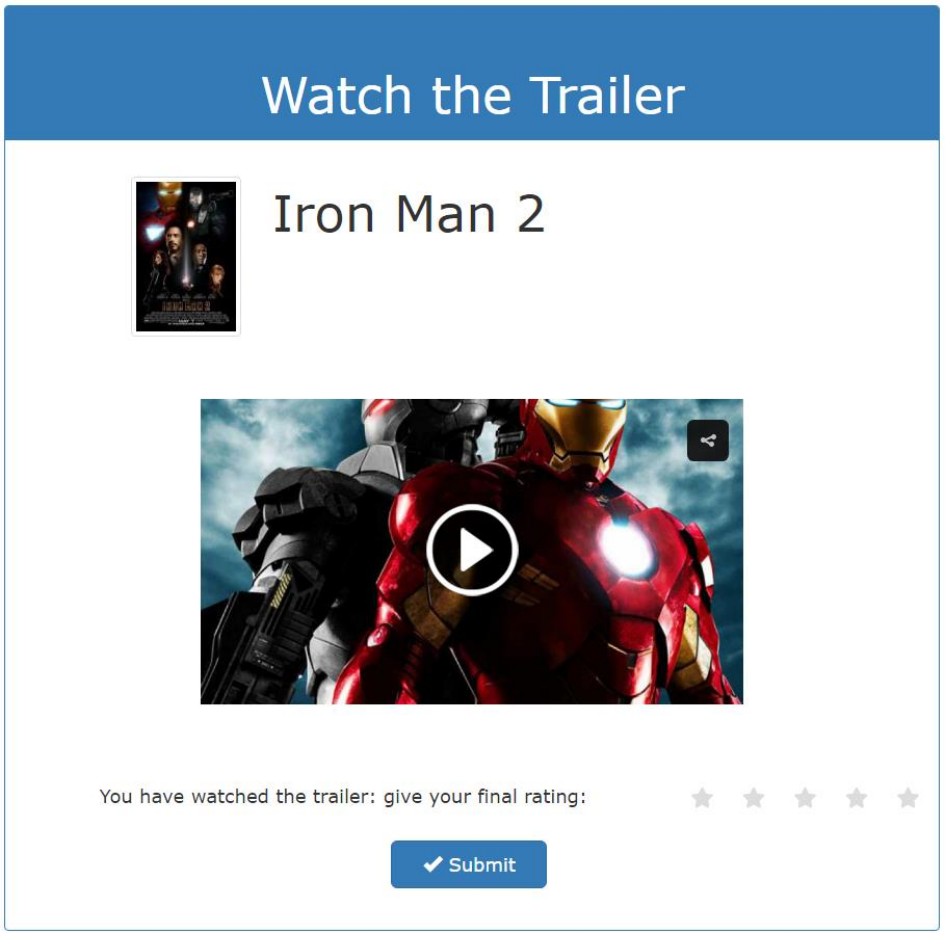
The explanation made the recommendation more convincing

The explanation helped me discover new information about this movie

The explanation increased my trust in the recommender system

Explanation and Questionnaire

Web Application



Ex-post Evaluation

Explanations - Results

	ExpLOD	Non-Personalized	Baseline
Transparency	**4.18**	3.04	3.01
Persuasion	**3.41**	2.84	2.59
Engagement	**3.48**	3.28	2.31
Trust	**3.39**	2.81	2.67
Effectiveness	0.72	**0.66**	0.93

Explanations - Results

	ExpLOD	Non-Personalized	Baseline
Transparency	**4.18**	3.04	3.01
Persuasion	**3.41**	2.84	2.59
Engagement	**3.48**	3.28	2.31
Trust	**3.39**	2.81	2.67
Effectiveness	0.72	**0.66**	0.93

Significant improvement in four out of five metrics

Non-significant gaps in terms of **effectiveness**.

Explanations - Results

aim	question		
transparency	I understood why this movie was recommended to me	○ topic ○ director	○ distributor ○ music composer
persuasion	The explanation made the recommendation more convincing	○ awards ○ director	○ location ○ producer
engagement	The explanation helped me discover new information about this movie	○ writer ○ director	○ producer ○ distributor
trust	The explanation increased my trust in the	○ awards ○ compos	○ producer ○ topic

Readings

Semantics-aware Recommender Systems

- V. W. Anelli, V. Bellini, T. Di Noia, W. La Bruna, P. Tomeo, E. Di Sciascio: ***An analysis on Time- and Session-aware diversification in recommender systems.*** UMAP - 25th Conference on User Modeling, Adaptation and Personalization - 2017
- A. Ragone, P. Tomeo, C. Magarelli, T. Di Noia, M. Palmonari, A. Maurino, E. Di Sciascio: ***Schema-summarization in Linked-Data-based feature selection for recommender systems.*** 32nd ACM SIGAPP Symposium On Applied Computing - 2017
- T. Di Noia, J. Rosati, P. Tomeo, E. Di Sciascio: ***Adaptive multi-attribute diversity for recommender systems.*** Inf. Sci. 382-383: 234-253(2017)
- S. Oramas, V. C. Ostuni, T. Di Noia, X. Serra, E. Di Sciascio: ***Sound and Music Recommendation with Knowledge Graphs.*** ACM TIST 8(2): 21:1-21:21 (2017)
- C. Musto, G. Semeraro, M. de Gemmis, P. Lops: ***Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems***. ECIR 2016: 729-734
- T. Di Noia, V. C. Ostuni, P. Tomeo, E. Di Sciascio: ***SPrank: Semantic Path-Based Ranking for Top-N Recommendations Using Linked Open Data.*** ACM TIST 8(1): 9:1-9:34 (2016)
- P. Tomeo, I. Fernández-Tobías, T. Di Noia, I. Cantador: ***Exploiting Linked Open Data in Cold-start Recommendations with Positive-only Feedback.*** CERI 2016: 11
- J. Rosati, P. Ristoski, T. Di Noia, R. De Leone, H. Paulheim: ***RDF Graph Embeddings for Content-based Recommender Systems.*** CBRecSys@RecSys 2016: 23-30
- I. Fernández-Tobías, P. Tomeo, I. Cantador, T. Di Noia, E. Di Sciascio***: Accuracy and Diversity in Cross-domain Recommendations for Cold-start Users with Positive-only Feedback.*** RecSys 2016: 119-122
- C. Musto, G. Semeraro, M. de Gemmis, P. Lops: ***Word Embedding Techniques for Content-based Recommender Systems: An Empirical Evaluation***. RecSys Posters 2015
- C. Musto, P. Basile, M. de Gemmis, P. Lops, G. Semeraro, S. Rutigliano: ***Automatic Selection of Linked Open Data Features in Graph-based Recommender Systems***. CBRecSys@RecSys 2015: 10-13

Readings

Semantics-aware Recommender Systems

- M. de Gemmis, P. Lops, C. Musto, F.Narducci, G. Semeraro: ***Semantics-Aware Content-Based Recommender Systems***. Recommender Systems Handbook 2015: 119-159
- P. Tomeo, T. Di Noia, M. de Gemmis, P. Lops, G. Semeraro, E. Di Sciascio: ***Exploiting Regression Trees as User Models for Intent-Aware Multi-attribute Diversity.*** CBRecSys@RecSys 2015: 2-9
- T. Di Noia, V. C. Ostuni: ***Recommender Systems and Linked Open Data.*** Reasoning Web 2015: 88-113
- P. T. Nguyen, P. Tomeo, T. Di Noia, E. Di Sciascio: ***Content-Based Recommendations via DBpedia and Freebase: A Case Study in the Music Domain.*** International Semantic Web Conference (1) 2015: 605-621
- P. Nguyen, P. Tomeo, T. Di Noia, E. Di Sciascio***: An evaluation of Sim: Rank and Personalized PageRank to build a recommender system for the Web of Data.*** WWW (Companion Volume) 2015: 1477-1482
- P. Basile, C. Musto, M. de Gemmis, P. Lops, F. Narducci, G. Semeraro: ***Content-Based Recommender Systems + DBpedia Knowledge = Semantics-Aware Recommender Systems***. SemWebEval@ESWC 2014: 163-169
- C. Musto, P. Basile, P. Lops, M. de Gemmis, G. Semeraro: ***Linked Open Data-enabled Strategies for Top-N Recommendations***. CBRecSys@RecSys 2014: 49-56
- C. Musto, G. Semeraro, P. Lops, M. de Gemmis: ***Combining Distributional Semantics and Entity Linking for Context-Aware Content-Based Recommendation***. UMAP 2014: 381-392
- V. C. Ostuni, T. Di Noia, R. Mirizzi, E. Di Sciascio: ***A Linked Data Recommender System Using a Neighborhood-Based Graph Kernel.*** EC-Web 2014: 89-100
- V. C. Ostuni, G. Gentile, T. Di Noia, R. Mirizzi, D. Romito, E. Di Sciascio: ***Mobile Movie Recommendations with Linked Data.*** CD-ARES 2013: 400-415
- V. C. Ostuni, T. Di Noia, E. Di Sciascio, R. Mirizzi: ***Top-N recommendations from implicit feedback leveraging linked open data.*** RecSys 2013: 85-92

Readings

Semantics-aware Recommender Systems

- C. Musto, G. Semeraro, P. Lops, M. de Gemmis: ***Contextual eVSM: A Content-Based Context-Aware Recommendation Framework Based on Distributional Semantics***. EC-Web 2013: 125-136
- T. Di Noia, R. Mirizzi, V. C. Ostuni, D. Romito, M. Zanker: ***Linked open data to support content-based recommender systems.*** I-SEMANTICS 2012: 1-8
- C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, R. Clout: ***Enhanced Semantic TV-Show Representation for Personalized Electronic Program Guides***. UMAP 2012: 188-199
- M. Degemmis, P. Lops, G. Semeraro: ***A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation***. User Model. User-Adapt. Interact. 17(3): 217-255 (2007)
- G. Semeraro, M. Degemmis, P. Lops, P. Basile: ***Combining Learning and Word Sense Disambiguation for Intelligent User Profiling***. IJCAI 2007: 2856-2861

Cross-language Recommender Systems

- C. Musto, F. Narducci, P. Basile, P. Lops, M. de Gemmis, G. Semeraro: ***Cross-Language Information Filtering: Word Sense Disambiguation vs. Distributional Models***. AI*IA 2011: 250-261
- P. Lops, C. Musto, F. Narducci, M. de Gemmis, P. Basile, G. Semeraro: ***Cross-Language Personalization through a Semantic Content-Based Recommender System***. AIMSA 2010: 52-60

Explanations

- C. Musto, F. Narducci, P. Lops, M. de Gemmis, G. Semeraro: ***ExpLOD: a framework for Explaining Recommendations based on the Linked Open Data cloud***. RecSys 2016, to appear