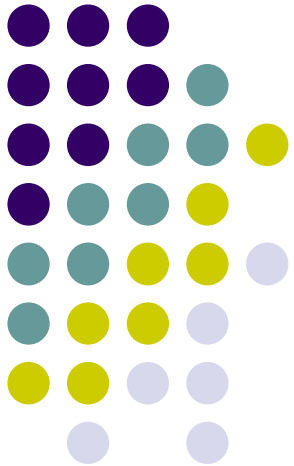
Using Linked Open Data in Recommender Systems

Ladislav Peška and Peter Vojtáš

Department of Software Engineering,
Charles University in Prague,
Czech Republic



Recommending in Czech Second-hand Bookshop

- Mostly **single item in stock**
- Few content-based attributes *(low information value)*
  - **Title**, **author**, **price**, **category**, textual description
  - Hard to define informative attributes
  - Title (and author name) in **Czech**
  - No common book identifier (ISBN mostly inapplicable)
- No explicit feedback
  - Page-view, time on page, buys…
  - Users identified through cookies
- Approx. 9500 active books
  - 50-100 visitors / day
  - 2-4 purchases

Challenge

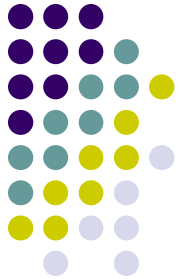
- Recommending for **small e-commerce websites**
  - No explicit feedback
  - Very few visited pages
  - Low user loyalty (they usually never go back)

  ⇒ **Not enough data for collaborative filtering**

- **Moreover for Second-hand bookshop**
  - Single item available in stock => **No best-selling objects**
  - Very few common attributes about books, expensive to fill-in

  ⇒ **Not enough data for content-based filtering**
  - Difficult navigation, high ratio between number of books vs. attributes, **data filtration and attribute search is problematical**

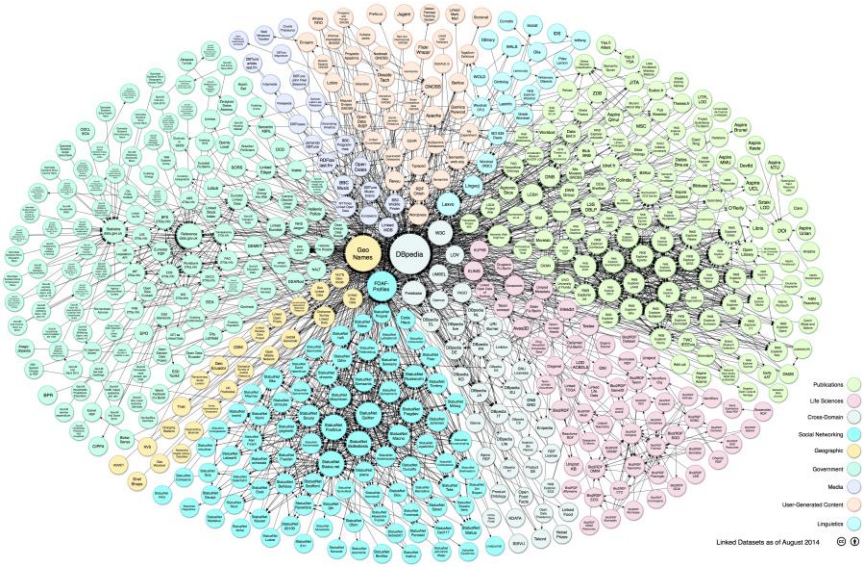
Proposed Solution: use LOD



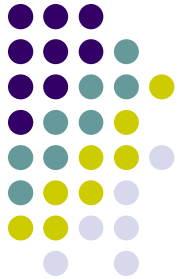
- Use information from LOD
  - Queried by SPARQL
  - In a natural way to the recommender systems
    - attributes of an object
  - Together with Content-based or Hybrid recommender system

Linked Open Data

- Part of Semantic Web movement
  - „lightweighted“ SemWeb
  - Each resource has unique URI
  - You can link resources accross datasets
  - Querying language SPARQL available
- DBPedia.org/sparql



LOD in RecSys

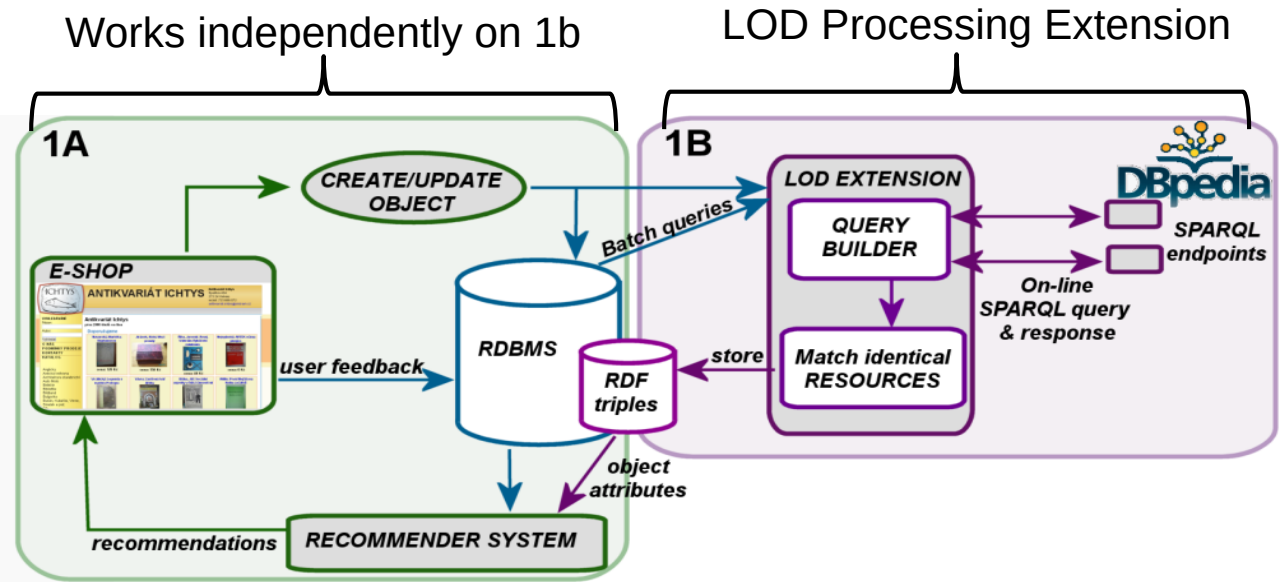


- Similarity of objects (or users) induced by (subset) of LOD
- Various distance metrics, e.g.,
  https://www.insight-centre.org/content/measuring-semantic-distance-linked-open-data-enabled-recommender-systems-0

- Or simply collect relevant attributes and run standard CB algorithm
  - You need to have a fallback if no LOD data was found for some objects
  - Regular updates of resources

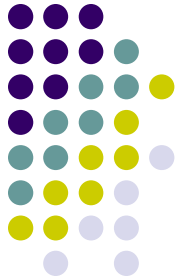
- Another option: build your service on LOD sources
  - LM's bachelor thesis

Using LOD in E-Commerce



- No local LOD repository (high maintenance costs)
  - Store only relevant information for existing E-shop objects
- Query LOD regulary for updated information and upon creates/updates
  - HTTP protocol

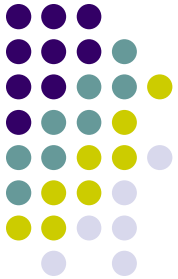
LOD Datasets Selection



- Necessary conditions
  - SPARQL endpoint
  - Good coverage of books domain
    - With relevant non-trivial attributes
  - Czech book names available

- We would really appriciate LOD meta-query language or LOD statistics
  - „Find me datasets with books available in Czech language“

Czech and English DBPedia



- Czech DBPedia
  - \- Less evolved in general
  - \- No sameAs links *(no link to other languages)*
  - \- ***Many books has no infoboxes*** -> no type -> keyword search
  - \+ wikiPageLinks even for unrecognized types
  - \+ Names in Czech
- English DBPedia
  - \- Mapped through owl:sameAs links => Czech Wiki page must exist
  - \- Locally famous authors probably won't be listed
  - \+ Better information for „blockbusters"

Czech and English DBPedia

Pes baskervillský

Infobox missing

Tento článek pojednává o původním románu. Další významy jsou uvedeny v článku Pes baskervillský (rozcestník).

Pes baskervillský (anglicky *The Hound of the Baskervilles*) je román z cyklu románů a povídek o slavném detektivovi Sherlocku Holmesovi, který napsal anglický, potažmo skotský, spisovatel Arthur Conan Doyle. Tento román vyšel na pokračování v magazínu *Strand*, který se proslavil právě vydáváním povídek a románů detektiva z *Baker Street 221B*, v letech 1901 – 1902 a přitom to byl první román po delší pauze, kdy spisovatel oznámil, že Holmesovy příběhy již psát nebude, neboť ho vyčerpávaly a odváděly od pro něj skutečně podstatné tvorby. Přesto je považován za jeden z nejlepších Holmesových příběhů.

Děj [editovat | editovat zdroj]

Varování: Následující část článku vyzrazuje zápletku nebo rozuzlení díla. **Přeskočit**

Příběh se otevírá, když Holmese a jeho přítele Dr. Watsona navštíví v jejich bytě v Londýně Dr. James Mortimer, který přináší příběh záhadné smrti svého pacienta sira Charlese Baskervilla. Mortimer žije v Dartmooru, což je známá oblast blat a vřesovišť v Devonu v jihozápadní Anglii. Sir Charles Baskerville se obává rodinné kletby, která prý od časů anglické občanské války stíhá členy rodu Baskervillů. Neboť prý darebný Hugo Baskerville unesl v té době dceru místního farmáře a uvěznil ji ve svém zámku, odkud mu ale utekla. Společně se svými kumpány ji pronásleduje přes pustinu blat, ale je zabit obrovským černým psem, který od té doby pronásleduje rod Baskervillů.

Sir Charles je nalezen mrtev za podivných okolností a Mortimer, kterému se vše toto nezdá, se vypraví za Holmesem, aby případ objasnil. Navíc je oznámeno, že po siru Charlesovi převezme panství jeho vzdálený synovec z Ameriky, Henry Baskerville. Po jeho příjezdu se Watson společně s ním a Dr. Mortimerem vypraví do Devonu, aby se mohl Henry ujmout svého majetku. Postupně se zde objevují boční linie příběhu, kdy vyjde najevo, že uprchlý vrah Selden je bratrem služebné v zámku a jiná další tajemství. Holmes, o kterém si všichni myslí, že je v Londýně, je však v Devonu a řeší případ tam. Postupně vyjde najevo, že skutečným vrahem sira Charlese byl opravdu pes, žádná kletba, ale pes z masa a kostí, kterého poslal Stapleton, který se chtěl zmocnit majetku rodu Baskervillů a musel sira Charlese odstranit. Nakonec je však jeho plán na zabití sira Henryho překažen a sám Stapleton hyne v Grimpenských bažinách. Holmes případ vyřeší, sir Henry se oženi s manželkou Stapletona, kterou vydával za svoji sestru, a která se jej snažila od jeho plánů odradit a zastavit jeho krvavé tažení za majetkem a příběh končí do jisté míry šťastně.

Konec části článku, která vyzrazuje zápletku nebo rozuzlení díla.

Související články [editovat | editovat zdroj]

The Hound of the Baskervilles

From Wikipedia, the free encyclopedia

For other uses, see The Hound of the Baskervilles (disambiguation).

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(December 2011)*

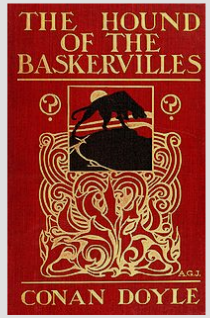
The Hound of the Baskervilles is the third of the crime novels written by Sir Arthur Conan Doyle featuring the detective Sherlock Holmes. Originally serialised in *The Strand Magazine* from August 1901 to April 1902, it is set largely on Dartmoor in Devon in England's West Country and tells the story of an attempted murder inspired by the legend of a fearsome, diabolical hound of supernatural origin. Sherlock Holmes and his companion Dr. Watson investigate the case. This was the first appearance of Holmes since his intended death in "The Final Problem", and the success of *The Hound of the Baskervilles* led to the character's eventual revival.

In 2003, the book was listed as number 128 of 200 on the BBC's The Big Read poll of the UK's "best-loved novel."[2] In 1999, it was listed as the top Holmes novel, with a perfect rating from Sherlockian scholars of 100.[3]

Contents [hide]

Wikipedia Infobox

The Hound of the Baskervilles



Cover of the 1st edition

Author	Arthur Conan Doyle
Illustrator	Sidney Paget
Cover artist	Alfred Garth Jones
Country	United Kingdom
Language	English
Series	Sherlock Holmes
Genre	Detective fiction
Publisher	George Newnes
Publication date	1902[1]
Preceded by	*The Memoirs of Sherlock*

Querying DBPedia

- Simple SPARQL queries for *book* and *author*
  - *With or without check for correct rdf:type*

```
select ?s, ?p, ?o                                    (b)
where {
  ?s a <http://dbpedia.org/ontology/WrittenWork>.
  ?s <http://www.w3.org/2000/01/rdf-schema#label> ?l.
  ?l bif:contains'("Otec" and "Prasátek")'.
  ?s ?p ?o.
}
```

- Processing RDF, finding
  - Related **genres**, **books**, **persons** and **categories**
  - Relevant location and language information
  - Tokenized abstract
  - Publisher, occupation etc.
- Stored as Binary attributes
  - E.g. RelatedPerson_Sherlock_Holmes = 1
  - Ideal for VSM and CBMF recommender systems

Query Statistics

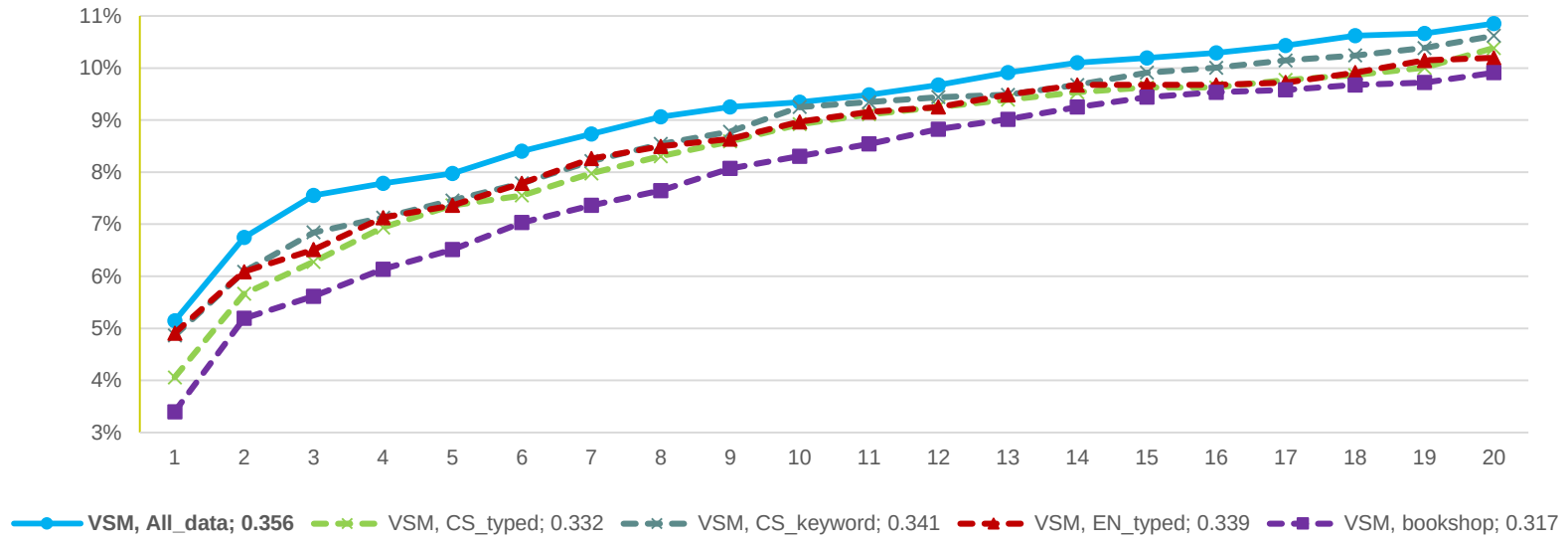
Query type	Found books	Found authors	Total triples
EN_typed	2.5% (0.104s)	31.1% (0.247s)	385K
CS_typed	1.0% (0.018s)	21.4% (0.022s)	651K (15K)
CS_keyword	9.2% (0.017s)	34.4% (0.024s)	1.37M (46K)

Attribute	EN_typed	CS_typed	CS_keyword
Rel. books	5.1% (3K)	4.4% (1.4K)	4.7% (1.4K)
Rel. persons	4.4% (6K)	2.3% (1.2K)	3.2% (1.7K)
Rel. categories	**32.5% (312K)**	**22% (615K)**	**38.5% (1.3M)**
Rel. genres	9.6% (5K)	2.4% (518)	4.8% (1.6K)
Language	1.3% (179)	0.0% (4)	2.1% (356)
Location	**26% (47K)**	**17.9% (15K)**	**18.3% (14K)**
Publisher	0.2% (26)	0.5% (114)	2.7% (1.0K)
Occupation	**17.7% (11K)**	7.6% (2.7K)	7.6% (2.5K)

- Data in total about 20-40% of e-shop objects, however the quality is often questionable
- Categories, genres, (location), related books and persons

Off-line Experiments

- Vector Space Model (VSM) algorithm with TF-IDF
- Train set (2/3 of user data), Test set (1/3)
- Recommender systems tries to predict visited objects
  - **Presence** of visited object **in top-k** recommended, **nDCG**

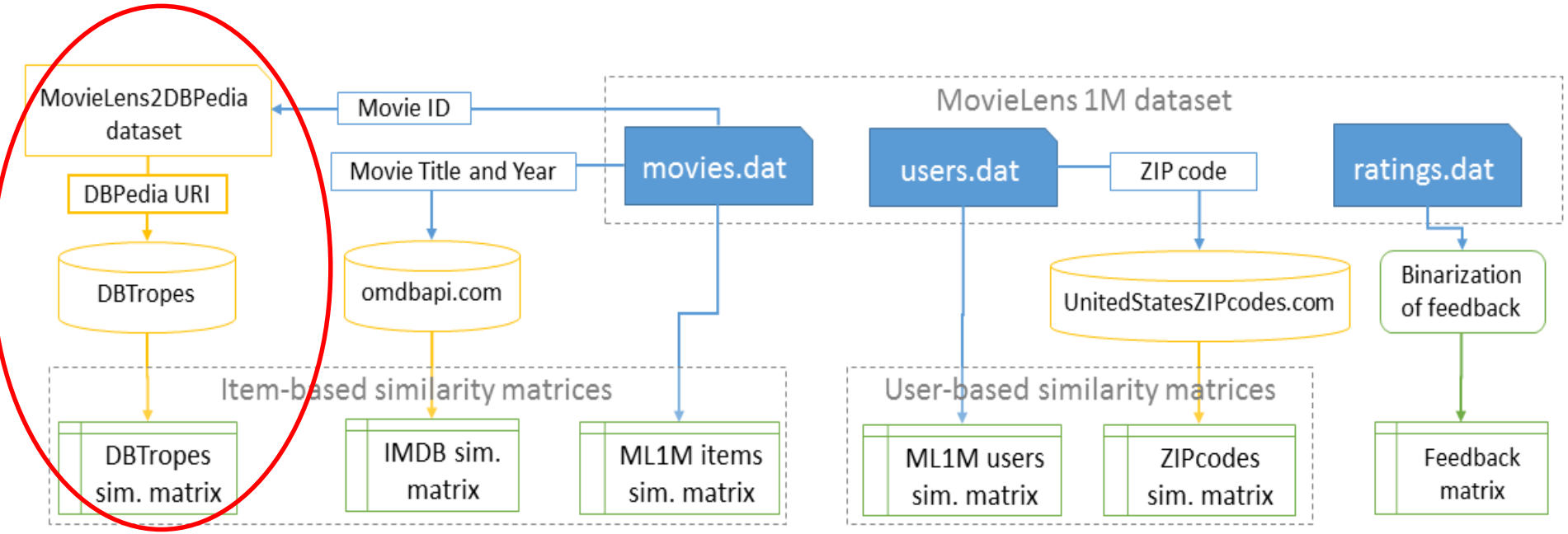


Concluding Remarques

- Applicability of LOD depends on domain coverage
- Automated querying of whole LOD cloud is still problematical
  - Keep in mind for future

- „Bounded“ applications plausible

Schema of MovieLens1M Dataset Extensions



Coverage:

15% 79% 100% 100% 97%