RecSys.Scifi: RECOMMENDER SYSTEMS DATASETS IN SCIENTIFIC FIELDS

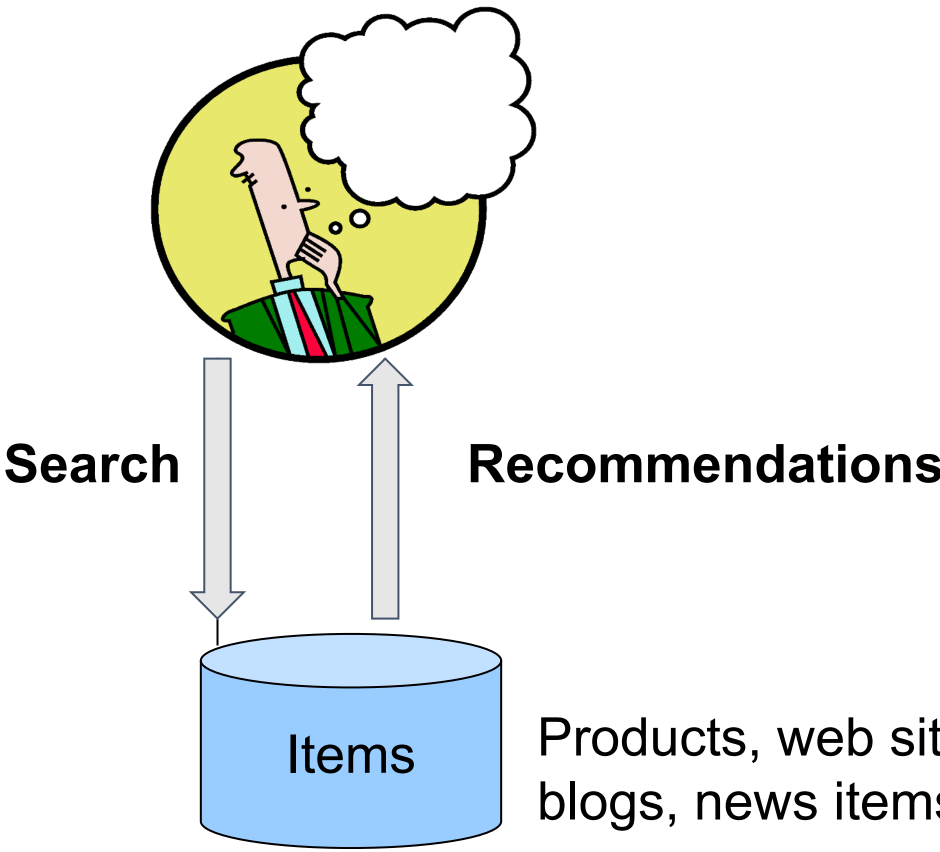
KDD2021 Lecture-style Tutorial

PART 1

Introduction to Recommender Systems

Francisco M. Couto

Recommendations



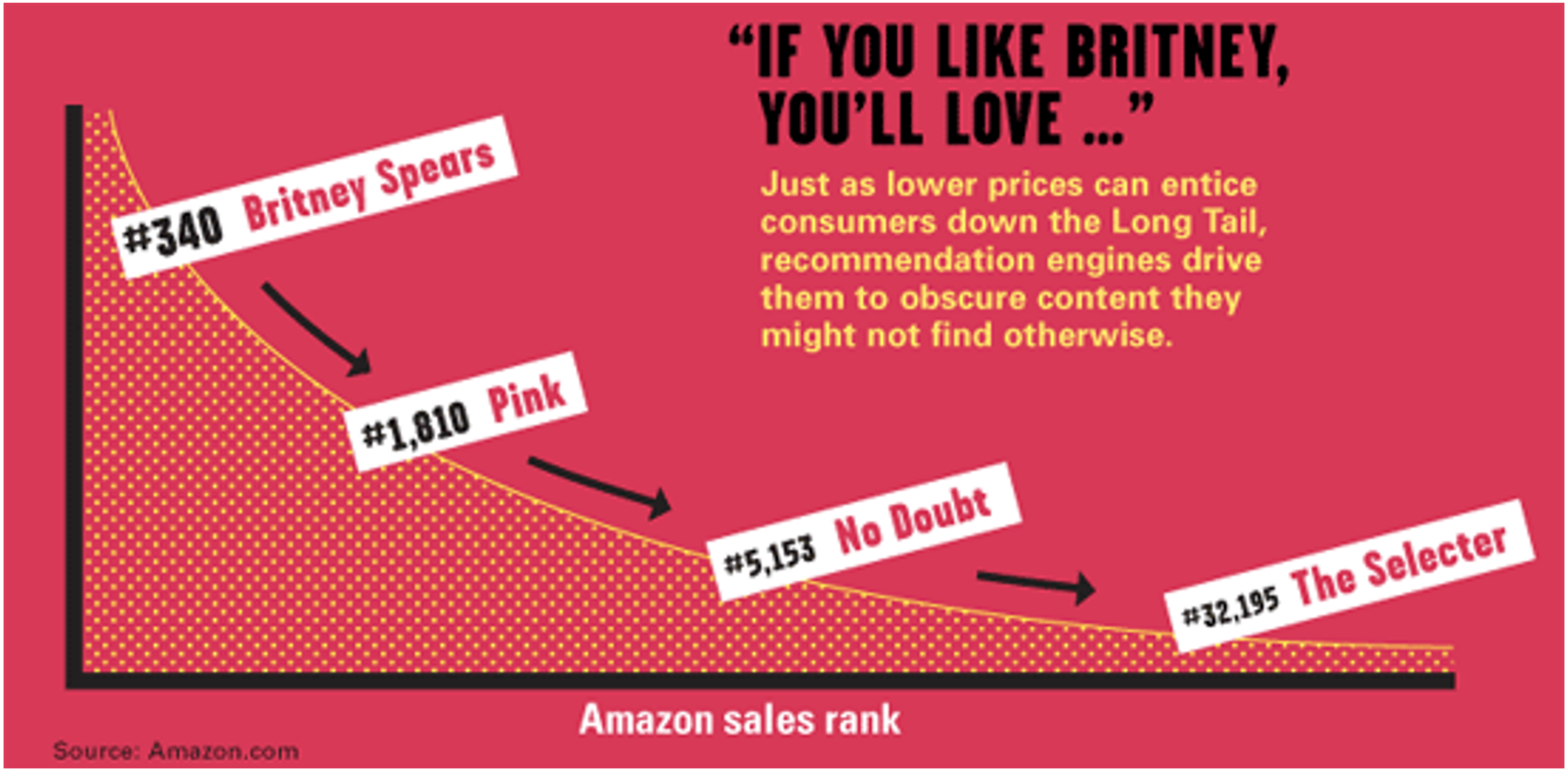
Products, web sites, blogs, news items, …

Examples:

amazon.com
PANDORA
StumbleUpon
del.icio.us
NETFLIX
movielens
helping you find the *right* movies
last.fm
the social music revolution
Google News
YouTube
XBOX LIVE

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Long Tail



Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Types of Recommendations

Editorial and hand curated

- List of favorites
- Lists of “essential” items

Simple aggregates

- Top 10, Most Popular, Recent Uploads

Tailored to individual users

- Amazon, Netflix, …

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Formal Model

X = set of **Customers**

S = set of **Items**

Utility function ***u***: ***X*** × ***S*** → ***R***

- ***R*** = set of ratings
- ***R*** is a totally ordered set
- e.g., **0-5** stars, real number in **[0,1]**

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Utility Matrix

	Avatar	LOTR	Matrix	Pirates
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Key Problems

(1) Gathering “known” ratings for matrix

- How to collect the data in the utility matrix

(2) Extrapolate unknown ratings from the known ones

- Mainly interested in high unknown ratings
  - We are not interested in knowing what you don’t like but what you like

(3) Evaluating extrapolation methods

- How to measure success/performance of recommendation methods

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Gathering Ratings

Explicit

- Ask people to rate items
- Doesn't work well in practice – people can't be bothered

Implicit

- Learn ratings from user actions
  - E.g., purchase implies high rating
- What about low ratings?

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Extrapolating Utilities

Key problem: Utility matrix ***U*** is **sparse**

- Most people have not rated most items
- **Cold start:**
  - New items have no ratings
  - New users have no history

Three approaches to recommender systems:

- **1)** Content-based
- **2)** Collaborative
- **3)** Hybrid

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Content-based

Main idea: Recommend items to customer ***x*** similar to previous items rated highly by ***x***

Example:

Movie recommendations

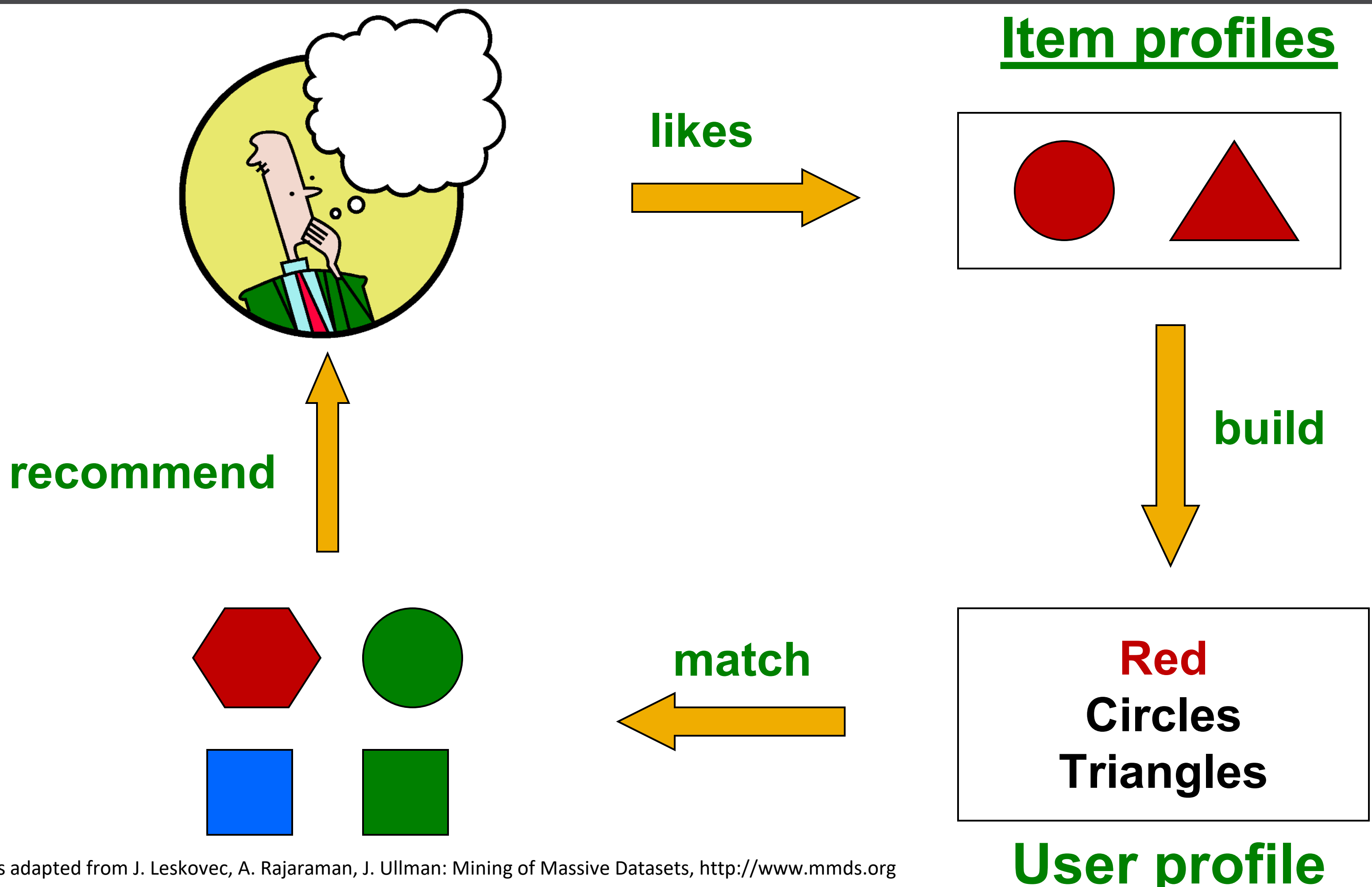
- Recommend movies with same actor(s), director, genre, …

Websites, blogs, news

- Recommend other sites with “similar” content

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Plan of Action



Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Item Profiles

For each item, create an **item profile**

Profile is a set (vector) of features

- **Movies:** author, title, actor, director,…
- **Text:** Set of "important" words in document

How to pick important features?

- Usual heuristic from text mining is **TF-IDF** (Term frequency * Inverse Doc Frequency)
  - **Term … Feature**
  - **Document … Item**

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Users Profiles and Prediction

- **User profile possibilities:**
  - Weighted average of rated item profiles
  - **Variation:** weight by difference from average rating for item
  - ...
- **Prediction heuristic:**
  - Given user profile ***x*** and item profile ***i***, estimate

$$u(\boldsymbol{x}, \boldsymbol{i}) = \cos(\boldsymbol{x}, \boldsymbol{i}) = \frac{\boldsymbol{x} \cdot \boldsymbol{i}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{i}||}$$

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Pros: Content-based

+: No need for data on other users

- No cold-start or sparsity problems

+: Able to recommend to users with unique tastes

+: Able to recommend new & unpopular items

- No first-rater problem

+: Able to provide explanations

- Can provide explanations of recommended items by listing content-features that caused an item to be recommended

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Cons: Content-based

–: Finding the appropriate features is hard

- E.g., images, movies, music

–: Recommendations for new users

- **How to build a user profile?**

–: Overspecialization

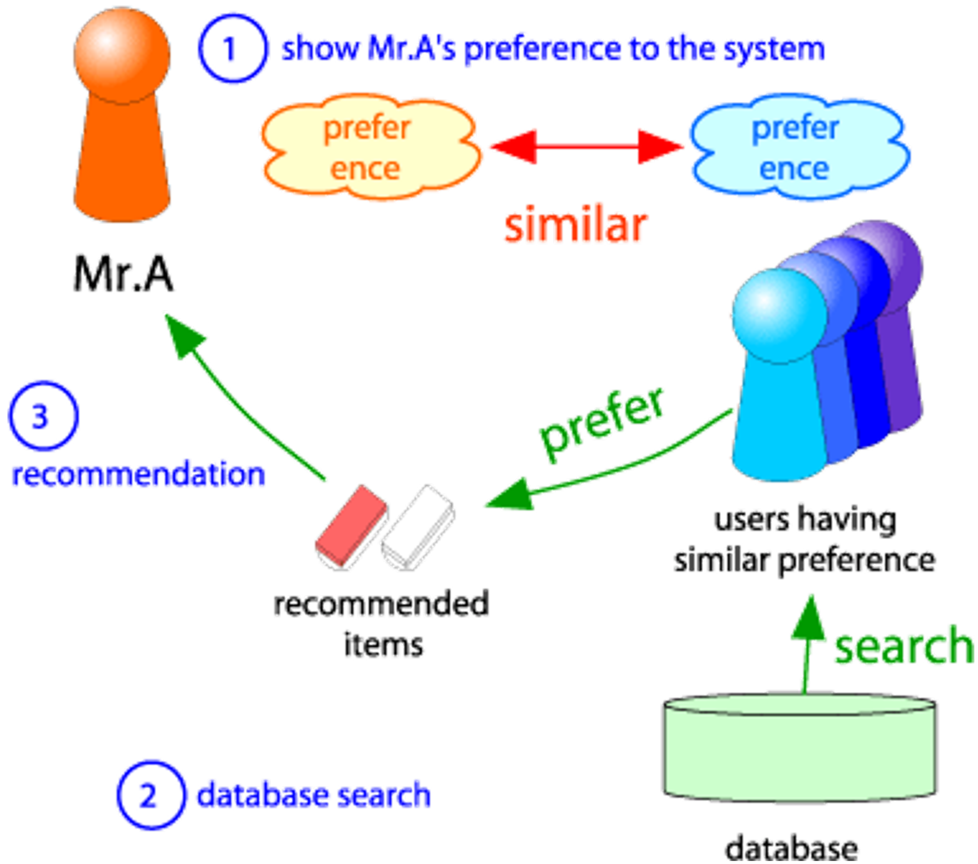
- Never recommends items outside user's content profile
- People might have multiple interests
- **Unable to exploit quality judgments of other users**

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Consider user ***x***

Find set ***N*** of other users whose ratings are "**similar**" to ***x***'s ratings

Estimate ***x***'s ratings based on ratings of users in ***N***



Pros/Cons Collab. Filtering

+ Works for any kind of item

- No feature selection needed

- Cold Start:

- Need enough users in the system to find a match

- Sparsity:

- The user/ratings matrix is sparse
- Hard to find users that have rated the same items

- First rater:

- Cannot recommend an item that has not been previously rated
- New items, Esoteric items

- Popularity bias:

- Cannot recommend items to someone with unique taste
- Tends to recommend popular items

Offline Evaluation

movies

users

1	3	4			
	3	5			5
		4	5		5
		3			
		3			
2			?		?
				?	
	2	1			?
	3			?	
1					

Test Data Set

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Evaluating Predictions

Predict the rating a user would give to an item

- Root-mean-square error (RMSE)
  - differences between the real and predicted ratings for all items
- Rank Correlation:
  - Spearman’s correlation between system and user complete rankings

Recommend a ranked list of (top@k) items

- Precision@k , Recall@k and F_measure@k
- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (nDGC)

Slides adapted from J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Online Evaluation

A/B testing

- Test different algorithms on-the-fly
- Measure #Recommendations followed
- Pros: measure real impact on users
- Cons: only available to data platform owners

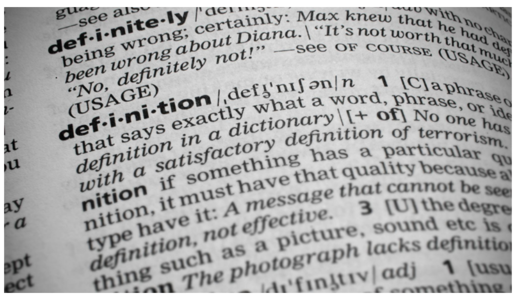
Scientific recommender systems

Matilde Pato

Scientific fields

Concepts and definitions

What is mean **Scientific Fields**?



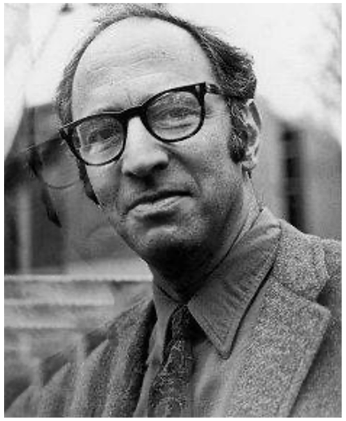
Oxford Dictionaries. 2021.

“*particular branches of study or spheres of activity or interest*”

Scientific fields

Concepts and definitions

What is mean **Scientific Fields**?



Thomas S. Kuhn (1922-1996)

"*Acquisition of a paradigm and of the more esoteric type of research it permits a sign of maturity in the development of any given scientific field*" (Structure of Scientific Revolutions, 1970)

Years after the publication of "The Structure of Scientific Revolutions", Kuhn dropped the concept of a paradigm and began to focus on the semantic aspects of scientific theories … (The Road since Structure, 2000)

Scientific fields

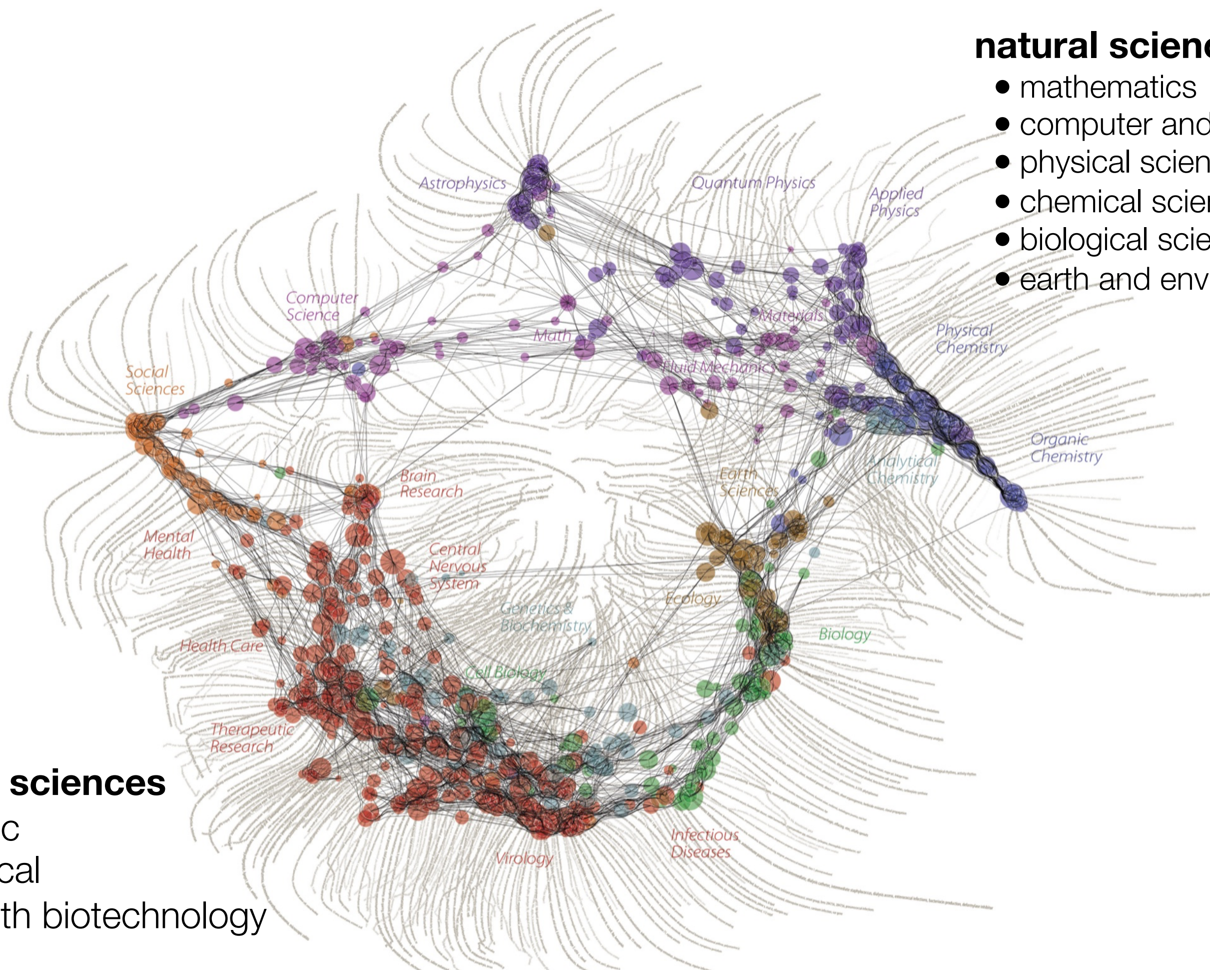
Branches of science

LASIGE driven by excellence

natural sciences

- mathematics
- computer and information system
- physical sciences
- chemical sciences
- biological sciences
- earth and environmental sciences

social sciences



health sciences

- basic
- clinical
- health biotechnology

Research & Node Layout: Kevin Boyack and Dick Klavans (mapofscience.com); Data: Thompson ISI; Graphics & Typography: W. Bradford Paley (didi.com/brad); Commissioned Katy Börner (scimaps.org)

Ciências ULisboa FCT Fundação para a Ciência e a Tecnologia

Scientific items

Concepts and definitions

What is mean **Scientific Items**?

IEEE Access

Received October 21, 2019, accepted November 19, 2019, date of publication December 5, 2019, date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958002

Using Research Literature to Generate Datasets of Implicit Feedback for Recommending Scientific Items

MÁRCIA BARROS[1,2], ANDRÉ MOITINHO[2], AND FRANCISCO M. COUTO[1]

[1]LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal

[2]CENTRA, Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal

Corresponding author: Márcia Barros (mcbarros@fc.ul.pt)

This work was supported by the Fundação para a Ciência e Tecnologia (FCT), under LASIGE Strategic Project - UID/CEC/00408/2019, CENTRA Strategic Project UID/FIS/00099/2019, FCT funded project PTDC/CCI-BIO/28685/2017 and PhD Scholarship SFRH/BD/128840/2017.

ABSTRACT In an age of information overload, we are faced with seemingly endless options from which a small number of choices must be made. For applications such as search engines and online stores, Recommender Systems have long become the key tool for assisting users in their choices. Interestingly, the use of Recommender Systems for recommending scientific items remains a rarity. One difficulty is that the development of such systems depends on the availability of adequate datasets of users' feedback. While there are several datasets available with the ratings of the users for books, music, or films, there is a lack of similar datasets for scientific fields, such as Astronomy and Life and Health Sciences. To address this issue, we propose a methodology that explores scientific literature for generating utility matrices of implicit feedback. The proposed methodology consists in identifying a list of items, finding research articles related to them, extracting the authors from each article, and finally creating a dataset where users are unique authors from the collected articles, and the rating values are the number of articles a unique author wrote about an item. Considering that literature is available for every scientific field, the methodology is in principle applicable to Recommender Systems in any scientific field. The methodology, which we call LIBRETTI (LIterature Based RecommEndaTion of scienTific Items), was assessed in two distinct study cases, Astronomy and Chemistry. Several evaluation metrics for the datasets generated with LIBRETTI were compared to those derived from other available datasets using the same set of recommender algorithms. The results were found to be similar, which provides a solid indication that LIBRETTI is a promising approach for generating datasets of implicit feedback for recommending scientific items.

INDEX TERMS Recommender systems, collaborative filtering, scientific literature, dataset, astronomy, chemical compounds.

I. INTRODUCTION

In the last years, scientific literature has increased in size and complexity [1]. Scientific literature has several applications and purposes, but the main goal is to disseminate the work and the discoveries of researchers. Recommender Systems (RSs) have been a useful help to that end, by improving the discoverability of research articles.

The goal of out article is to provide a methodology for generating datasets of implicit feedback, suitable for evaluating recommender algorithms in scientific areas, by going beyond the recommendation of topics and articles, and support the recommendation of scientific items. For the purposes of this work, we define **scientific item** as an entity belonging to the universe, that may be modeled, characterized by multiple features using a computational representation, and an object of research. Some examples of scientific items are genes, phenotypes, chemical entities, plants, diseases, stars, and groups of stars, such as Open Clusters and Galaxies.

RSs are software tools that provide suggestions for items that are presumably of interest to a particular user [2], which have been used in the recommendation of a wide range of products, for example, movies, books, research articles, or e-commerce [3]–[5]. Some well-known platforms integrating

The associate editor coordinating the review of this manuscript and approving it for publication was Pasquale De Meo.

"(...) an entity belonging to universe, that may be modeled, characterized by multiple features using computational representation, and an object of research."

Barros et al in IEEE Access 2019, 7, pp. 176668-176680

"(...) genes, phenotypes, ***chemical entities***, plants, **disease**, stars"

Literature review

Search process

4 databases

1. ACM Digital Library
2. IEEE Computer Science
3. Elsevier
4. Springer Link

2 search engines

1. Google Scholar
2. Semantic Scholar

search algorithm

{ {recommender OR recommendation} AND {system OR engine} AND ...

1. {drug OR medication} }
2. {chemical compounds OR drug} }
3. {disease} }
4. include "AND {dataset}"

Literature review

Search process

include

1. conference proceeding and journal published after 2009 to present
2. studies focusing on *scientific item* recommendation systems

exclude

1. manuscripts written in other language than english
2. technical reports, master and PhD dissertation
3. surveys

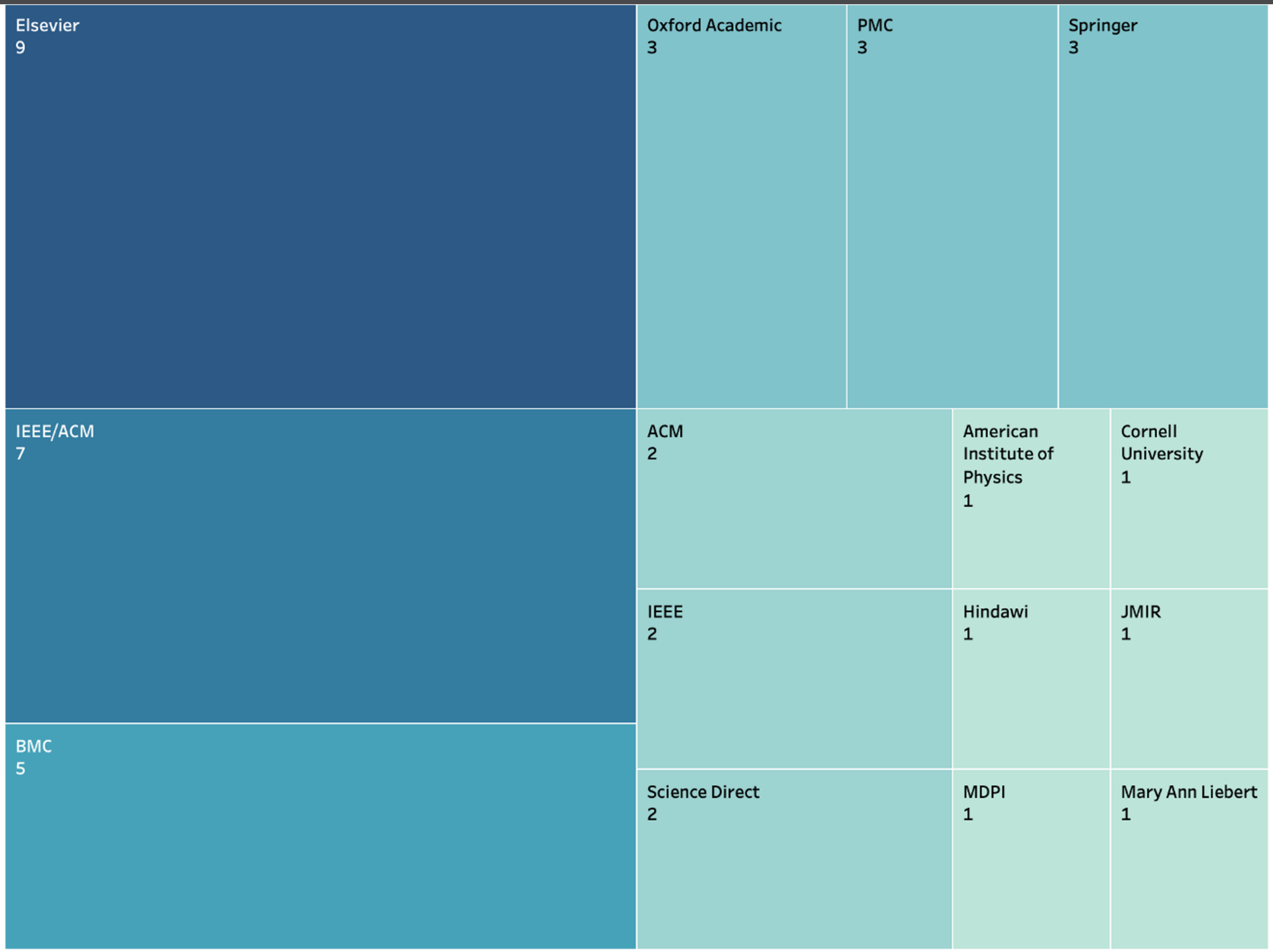
criteria of selection

1. clearly stated objectives, results and findings on the domain of knowledge
2. well-presented and justified arguments
3. well-referenced with a minimum of 10 sources

Trends of a Glance: publisher

LASIGE driven by excellence

#articles

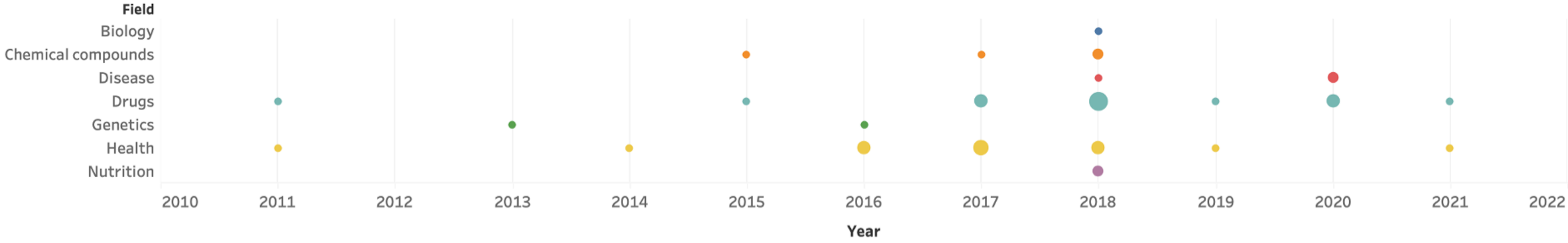


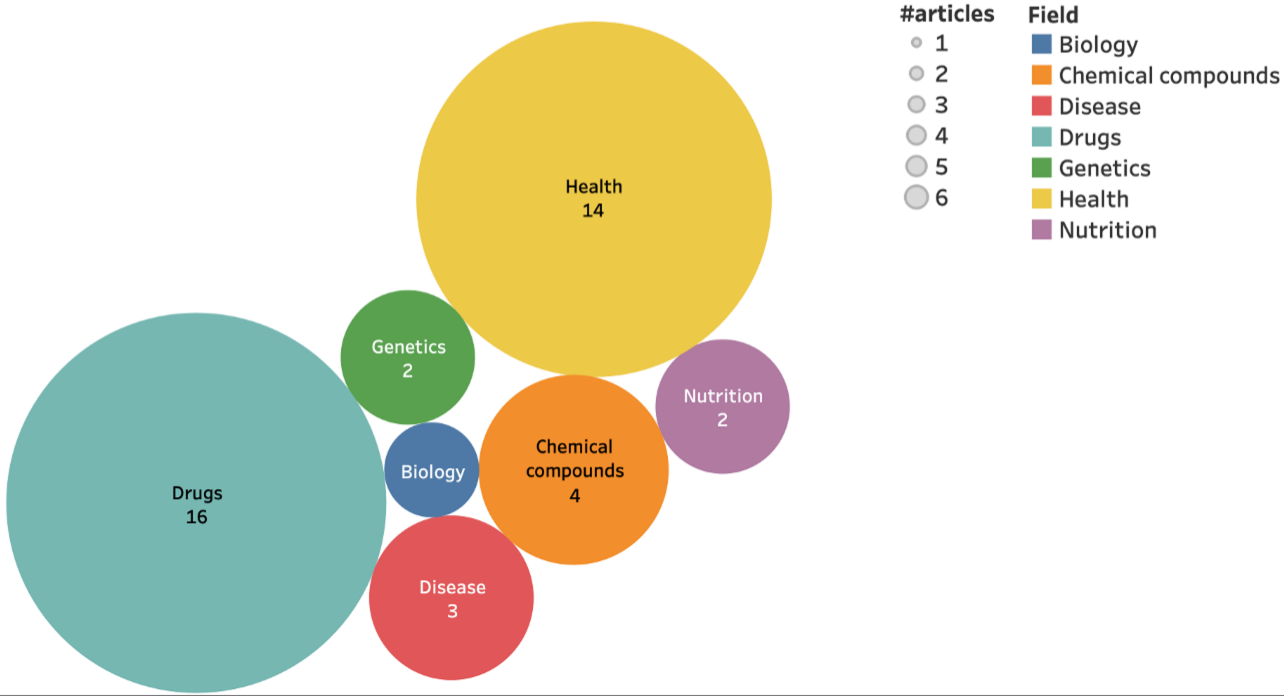
Ciências ULisboa FCT Fundação para a Ciência e a Tecnologia

Journal and conference

Scientific fields

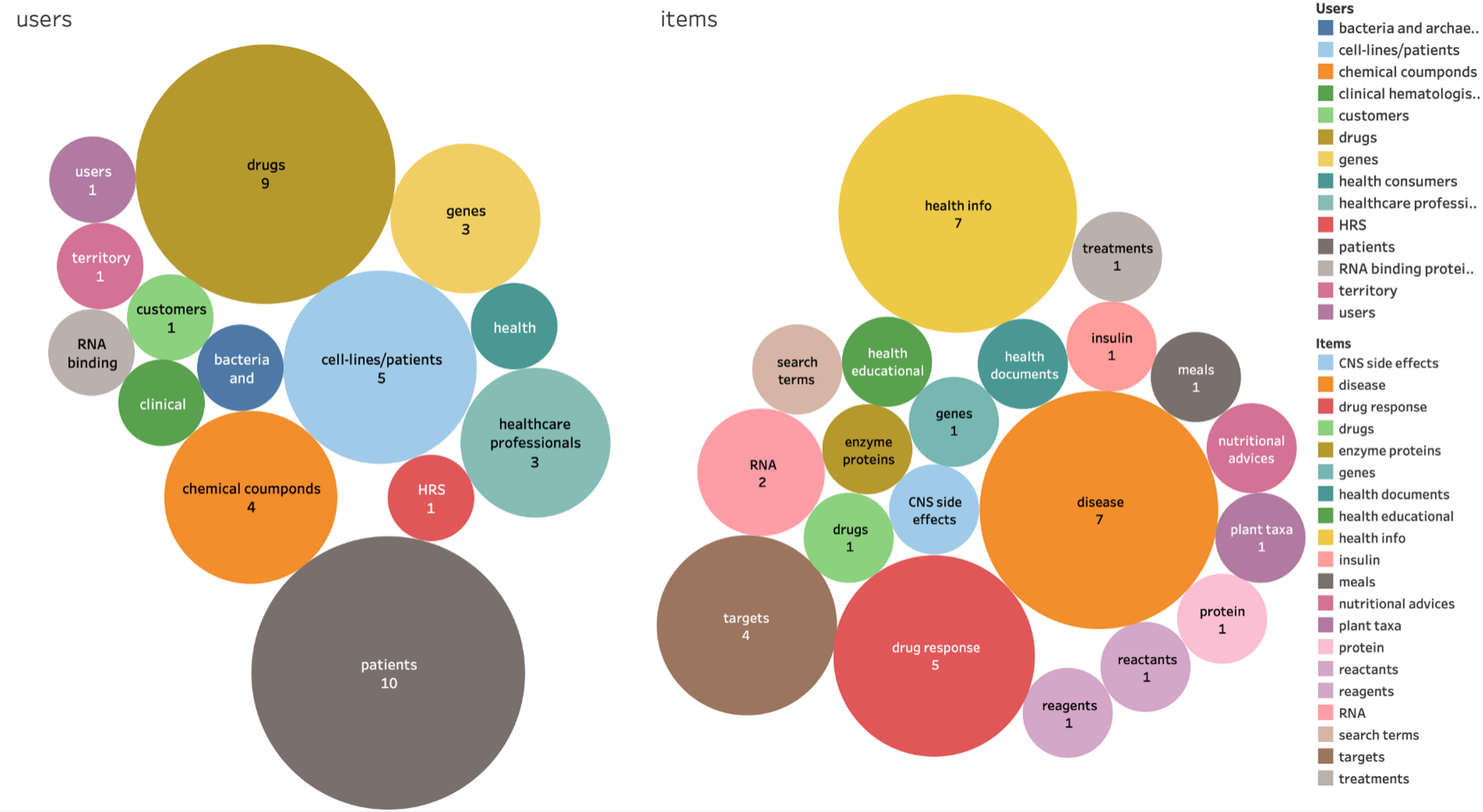
#articles





Matrix Factorization

LASIGE driven by excellence

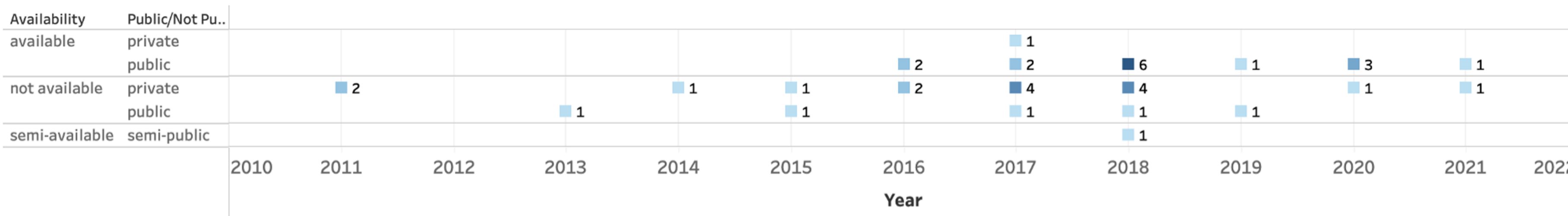


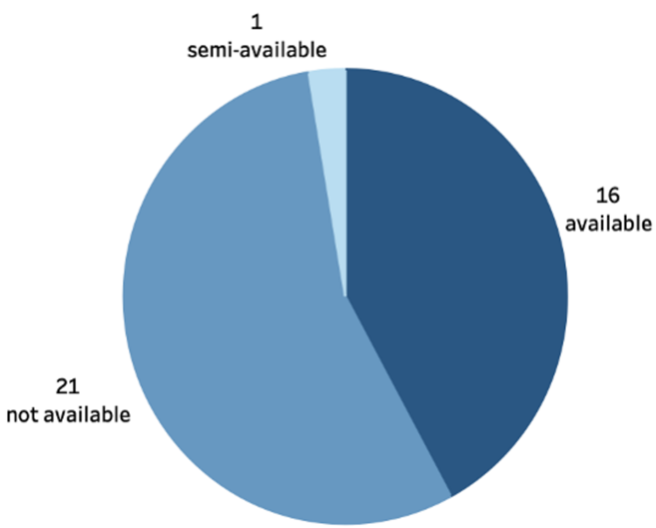
Ciências ULisboa

FCT Fundação para a Ciência e a Tecnologia

tuple: < user, item >

Availability of the dataset





Source of the dataset

LASIGE driven by excellence

Ciências ULisboa

FCT Fundação para a Ciência e a Tecnologia

Introduction to Named Entity Recognition (NER) and Named Entity Linking (NEL)

Pedro Ruas

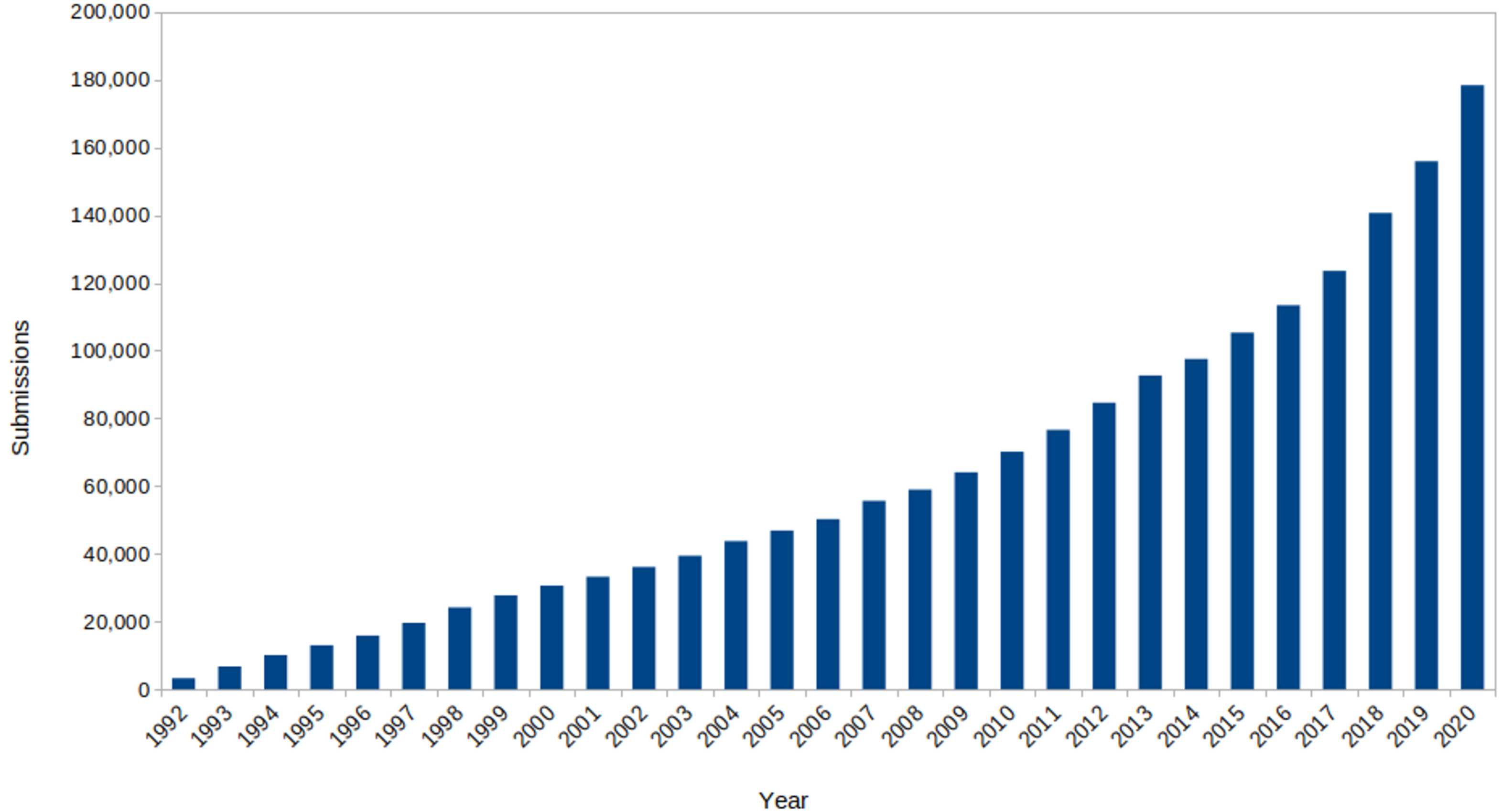
Scientific/biomedical knowledge

Described in text in:

- published papers
- electronic health records
- clinical trials
- patents
- database entries
- ...

Scientific/biomedical knowledge

New submissions to arXiv by year[1]

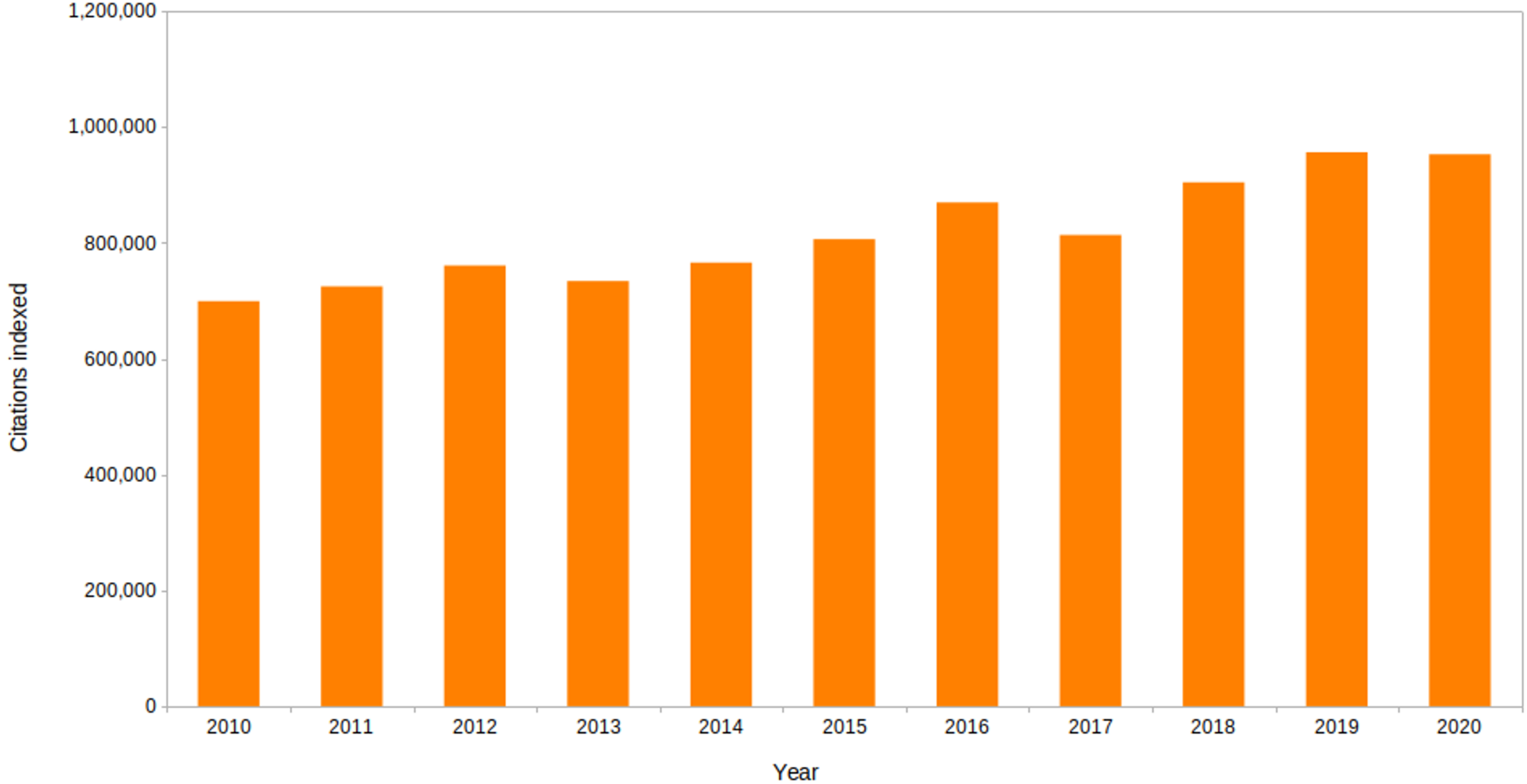


[1]Based on: https://arxiv.org/stats/monthly_submissions

Scientific/biomedical knowledge

LASIGE driven by excellence

Citations indexed to MEDILINE Pubmed by year[1]

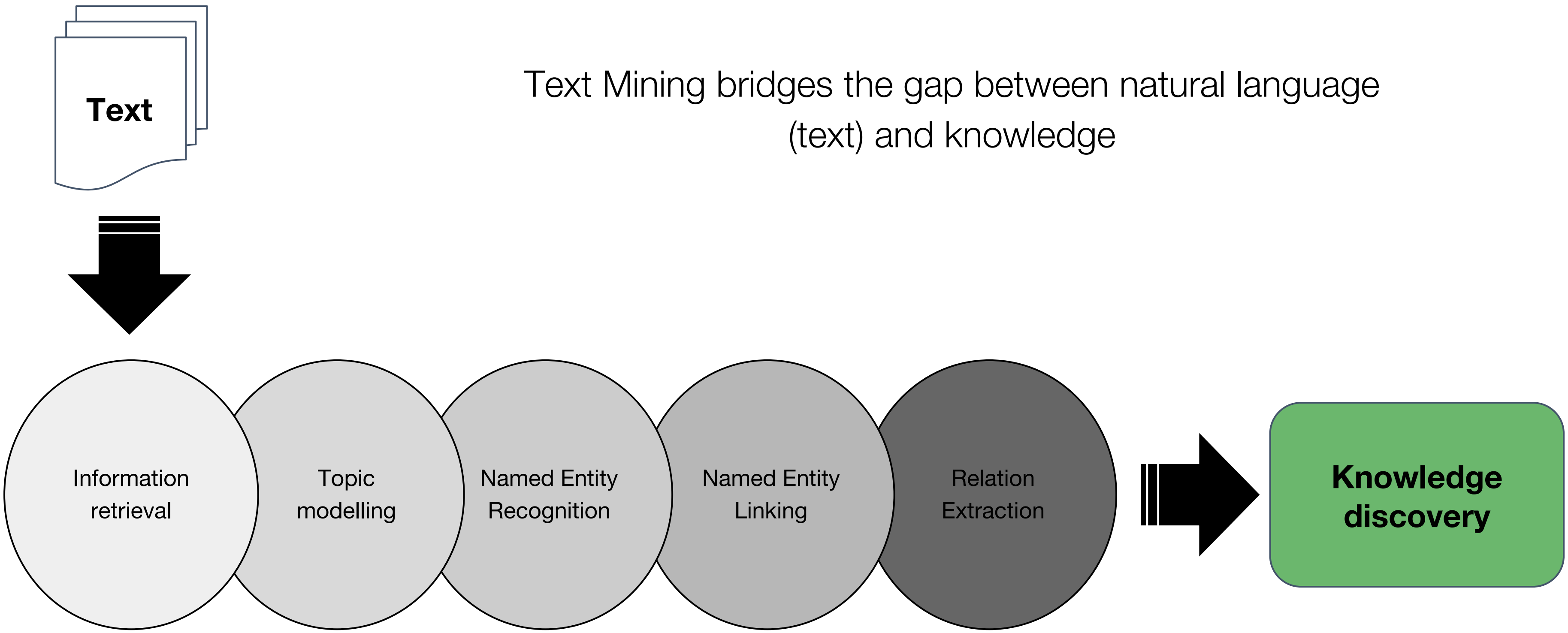


[1]Based on: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

Ciências ULisboa FCT Fundação para a Ciência e a Tecnologia

Text Mining

Text Mining bridges the gap between natural language (text) and knowledge



Text Mining tasks

Named Entity Recognition

Definition

Task introduced in the MUC-6 evaluation (1995):

«Named Entity (NE) -- Insert SGML tags into the text to mark each string that represents a person, organization, or location name, or a date or time stamp, or a currency or percentage figure.» [1]

Another definition (2018):

*«Named Entity Recognition and Classification, an important sub-task of Information Extraction, points to **identify** and **classify** members of rigid designators from data suited to different types of named entities such as organizations, persons, locations, etc.»* [2]

[1]B. M. Sundheim, "Overview of results of the MUC-6 evaluation," in *MUC6 '95: Proceedings of the 6th conference on Message understanding*, 1995, pp. 13–31, doi: https://doi.org/10.3115/1072399.1072402.
[2]A. Goyal, V. Gupta, and M. Kumar, "Recent Named Entity Recognition and Classification techniques: A systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–43, 2018, doi: 10.1016/j.cosrev.2018.06.001.

Named Entity Recognition

Definition

Example

time virus location

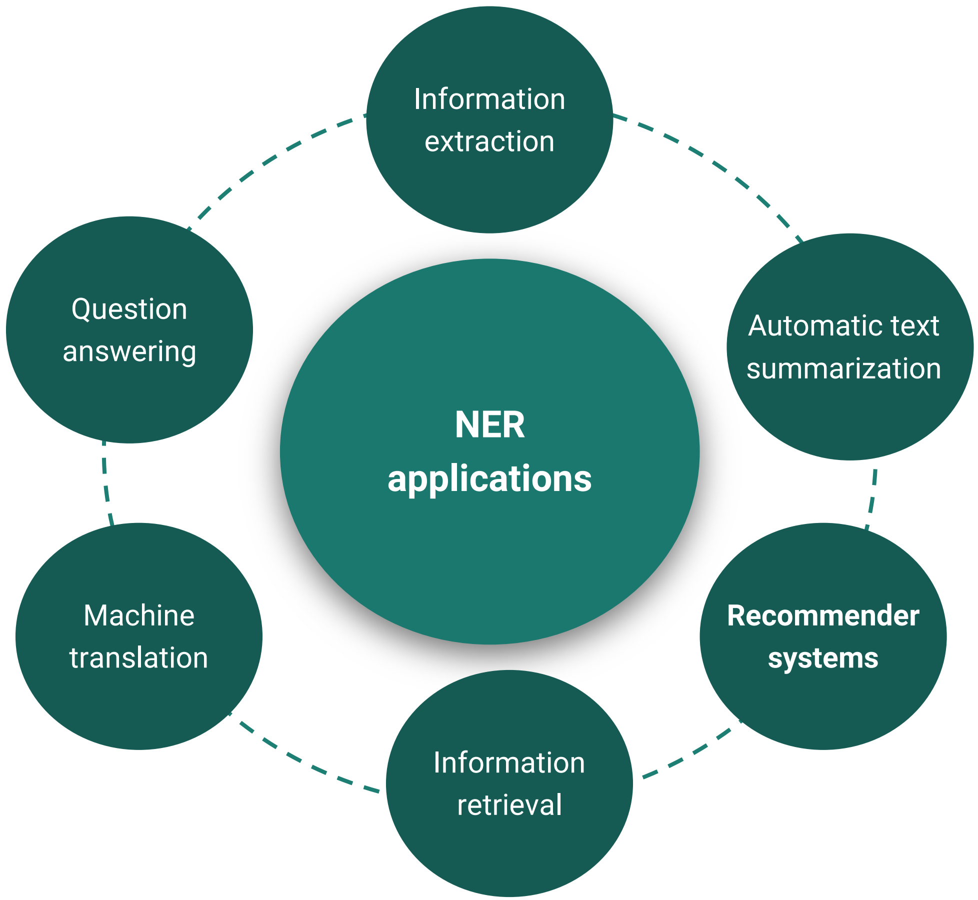
«At the end of December 2019, a novel coronavirus named SARS-CoV-2 in Wuhan, a central city in China, was announced by the World Health Organization»

location organization

Entity	Begin	End	Category
“December 2019”	13	26	time
“SARS-CoV-2”	54	64	virus
“Wuhan”	68	73	location
“China”	93	98	location
“World Health Organization”	121	146	organization

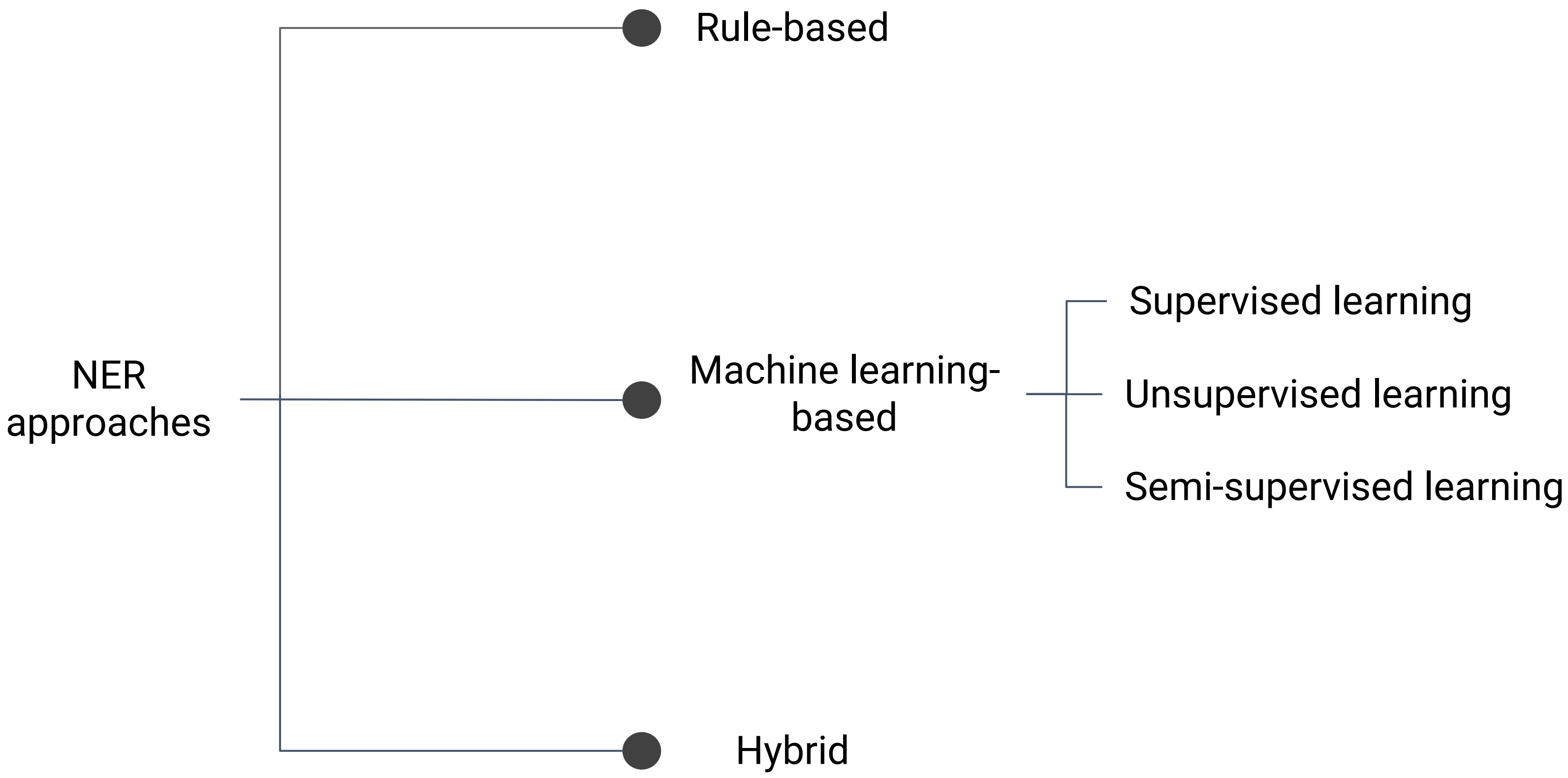
Named Entity Recognition

Applications



Named Entity Recognition

Types of systems

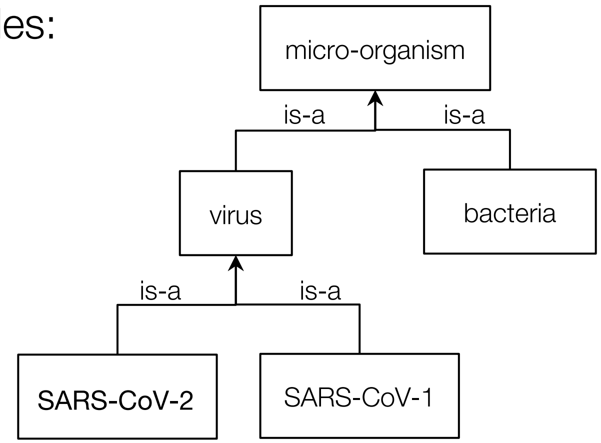


Based on A. Goyal, V. Gupta, and M. Kumar, “Recent Named Entity Recognition and Classification techniques: A systematic review,” *Comput. Sci. Rev.*, vol. 29, pp. 21–43, 2018, doi: 10.1016/j.cosrev.2018.06.001.

Knowledge Bases

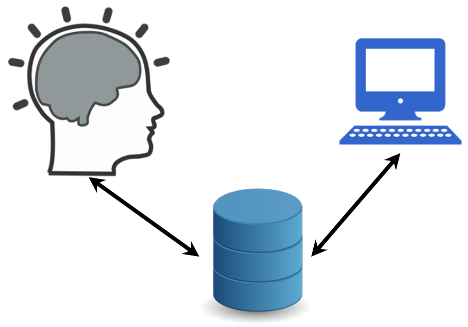
Formal representation of the reality or a part of it, which includes:

- ➔ concepts
- ➔ definitions
- ➔ attributes
- ➔ relations between concepts



Advantages

- shared understanding of reality/knowledge
- integration of knowledge
- accessible by both computer reasoning and humans



Knowledge Bases

General domain

Wikipedia

DBpedia

In the biomedical domain

GENEONTOLOGY Unifying Biology

DISEASE ONTOLOGY

ChEBI

MeSH

Named Entity Linking (NEL)

Definition

*«**Entity Linking**, also referred to as record linkage or entity resolution, involves aligning a textual mention of a **named-entity** to an appropriate **entry in a knowledge base**, which may or may not contain the entity.»* [1]

Text Mining Pipeline

Named Entity Recognition ⇒ **Named Entity Linking** ⇒ Relation extraction ⇒ …

[1]D. Rao, P. McNamee, and M. Dredze, “Entity Linking: Finding Extracted Entities in a Knowledge Base,” in *Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing.*, P. J. Poibeau T., Saggion H., Ed. Springer, Berlin, Heidelberg, 2013, pp. 93–115.

Named Entity Linking (NEL)

Definition

Example

time | virus | location

«At the end of December 2019, a novel coronavirus named SARS-CoV-2 in Wuhan, a central city in China, was announced by the World Health Organization»

location | organization

WIKIPEDIA

Entity				Category	Begin	End	URL
"December 2019"			time		13	26	
	https://en.wikipedia.org/wiki/2019#December						
"SARS-CoV-2"			virus		54	64	
	https://en.wikipedia.org/wiki/Severe_acute_respiratory_syndrome_coronavirus_2						
"Wuhan"			location		68	73	
	ttps://en.wikipedia.org/wiki/Wuhan						
"China"				location		93	98
	https://en.wikipedia.org/wiki/China						
"World Health Organization"		organization	121	146	https://en.wikipedia.org/wiki/World_Health_Organization		

Named Entity Linking (NEL)

Challenges

- **Ambiguity**

iris → circular structure of the eye

iris → a genus containing species of plants

- **Entity name variations** (abbreviations, synonyms, acronyms)

AF4

PBM1

AFF1

MLLT2

same human gene

myocardial infarction

heart attack

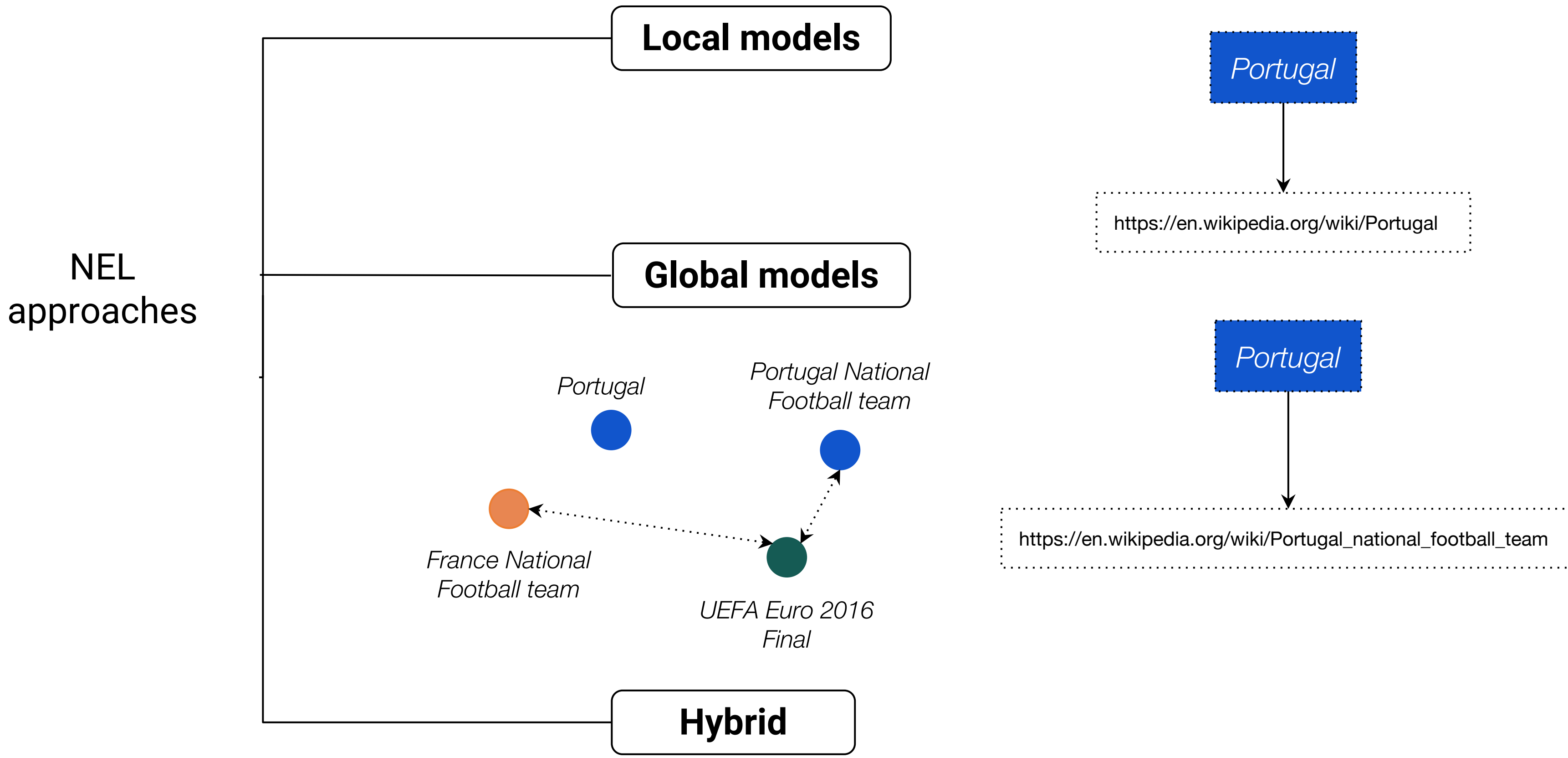
same medical condition

- **Incomplete ontologies/KBs**

Named Entity Linking (NEL)

Types of systems

«Portugal defeated France 1–0 at UEFA Euro 2016 Final»



NER + NEL + Recommender systems

What is the role of Text Mining in Recommender systems?

*«(...) text mining techniques can be exploited for the development of recommender systems (...) can be applied to detect user preferences (**user profiling**) and also to **extract context data.**»* [1]

[1]Y. Betancourt and S. Ilarri, “Use of text mining techniques for recommender systems,” in *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, 2020, vol. 1, no. Iceis, pp. 780–787, doi: 10.5220/0009576507800787.

NER + NEL + Recommender systems

LASIGE driven by excellence

Ref	Field	RS type	Role of NER/NEL	NER/NEL Tool	Ontologies
1	Videos	-	To extract content information from videos	-	-
2	News	Content-based	NER is used in tweets and external sources of news articles for creating better users' profiles, improving the recommendations	-	-
3	Movies	Content-based	To identify most relevant context found in text related to the movies, to create users' and items' profiles	DBpedia Spotlight, Wikipedia Mine, TAGME	DBpedia, Wikipedia
4	Books	Content-based	To identify most relevant context found in text related to the books, to create users' and items' profiles	TAGME	DBpedia

1. Q. Qi and J. Dong, "Named entity recognition in titles of Chinese videos from the web," *Proc. - 2011 IEEE Int. Conf. Comput. Sci. Autom. Eng. CSAE 2011*, vol. 4, pp. 220–224, 2011, doi: 10.1109/CSAE.2011.5952838.
2. F. Abel, Q. Gao, G. J. Houben, and K. Tao, "Analyzing user modeling on Twitter for personalized news recommendations" *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6787 LNCS, pp. 1–12, 2011, doi: 10.1007/978-3-642-22362-4_1.
3. C. Musto, G. Semeraro, P. Lops, and M. de Gemmis, "Combining distributional semantics and entity linking for context-aware content-based recommendation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8538, pp. 381–392, 2014, doi: 10.1007/978-3-319-08786-3_34.
4. P. Basile and C. Musto, "Aggregation strategies for linked open data-enabled recommender systems," in *2014 Europ. Semantic Web Conf. (Linked Open Data-enabled Recomm. Syst. Challenge)*, 2014, pp. 1–5, [Online]. Available: http://2014.eswc-conferences.org/sites/default/files/eswc2014-challenges_rs_submission_16.pdf.

NER + NEL + Recommender systems

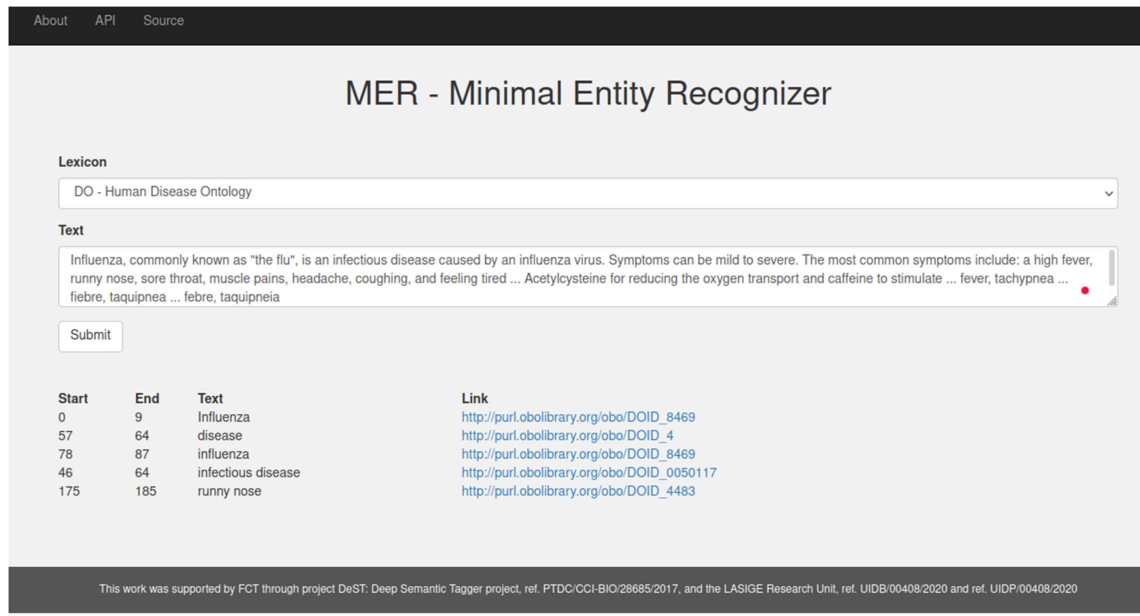
Ref	Field	RS type	Role of NER/NEL	NER/NEL Tool	Ontologies
5	Agro-business web pages	Collaborative-filtering	To extract entities from web pages to enrich the dataset	-	-
6	Agro-business web pages, movies	Collaborative-filtering	To extract entities from web pages to enrich the dataset	REMBRANDT, Stanford NER	Wikipedia
7	Dietary-related web pages, scientific text	-	To extract dietary concepts and recommendations from scientific sources	drNER	-
8	Teaching resources	-	To extract and link entities in the transcript of educational resources	Dandelion NER	DBpedia
9	Movies	Content-based	To find relevant entities mentioned in the user sentence in order to improve a dialog manager	-	Wikidata

5. M. A. Domingues *et al.*, "Applying multi-view based metadata in personalized ranking for recommender systems," in *Proceedings of the ACM Symposium on Applied Computing*, 2015, vol. 13-17-Apri, pp. 1105-1107, doi: 10.1145/2695664.2695955.
6. M. G. Manzato *et al.*, "Mining unstructured content for recommender systems: An ensemble approach," *Inf. Retr. J.*, vol. 19, no. 4, pp. 378–415, 2016, doi: 10.1007/s10791-016-9280-8.
7. T. Eftimov, B. K. Seljak, and P. Korošec, *A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations*, vol. 12, no. 6. 2017.
8. C. Limongelli, M. Lombardi, A. Marani, and D. Taibi, "Enrichment of the Dataset of Joint Educational Entities with the Web of Data," in *Proceedings - IEEE 17th International Conference on Advanced Learning Technologies, ICALT 2017*, 2017, pp. 528–529, doi: 10.1109/ICALT.2017.13.
9. A. Iovine, F. Narducci, and G. Semeraro, "Conversational Recommender Systems and natural language:: A study through the ConveRSE framework," *Decis. Support Syst.*, vol. 131, no. June 2019, 2020, doi: 10.1016/j.dss.2020.113250.

MER

Minimal Named Entity Recognizer[1]

- NER + NEL step
- text processing command-line tools *grep* and *awk*
- inverted recognition technique
- Python implementation: merpy

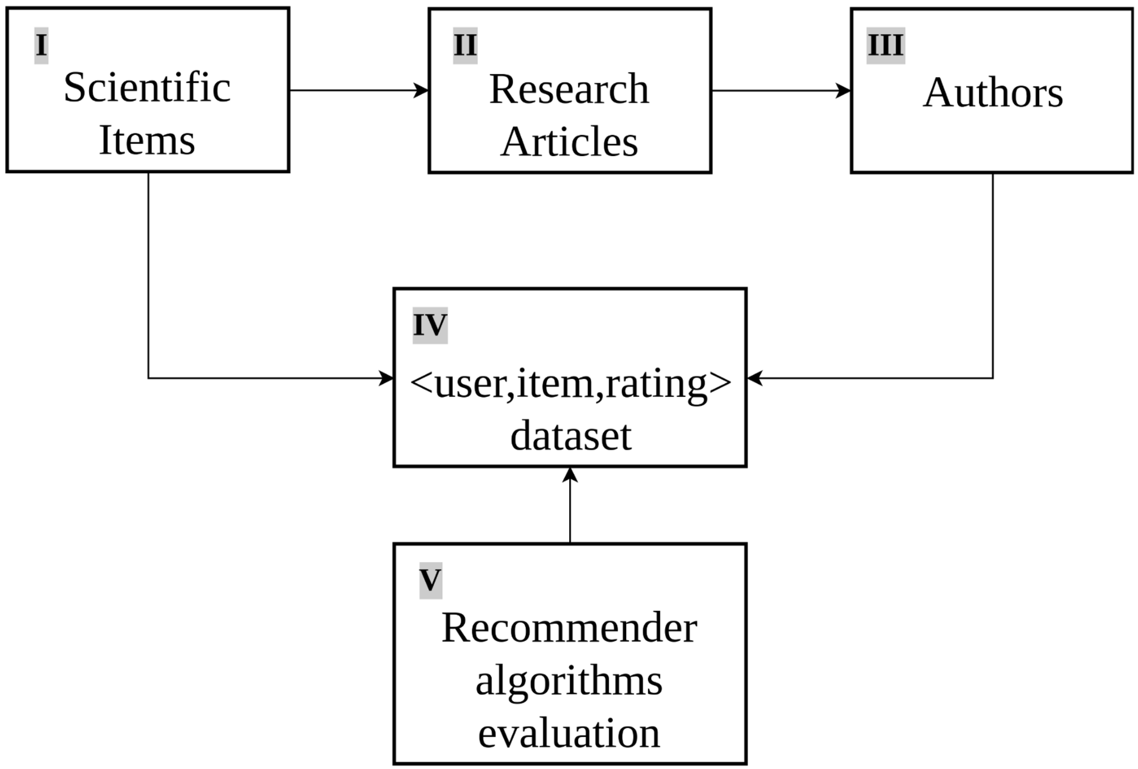


[1] F. M. Couto and A. Lamurias, "MER: a shell script and annotation server for minimal named entity recognition and linking," *J. Cheminform.*, vol. 10, no. 1, p. 58, Dec. 2018, doi: 10.1186/s13321-018-0312-9.

LIterature Based RecommEndaTion of scienTific Items (LIBRETTI)

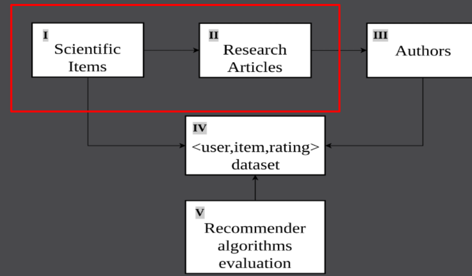
Márcia Barros

Goal: Create a **standard dataset** (user, item, rating) for recommender algorithms by extracting **implicit** information from the **scientific literature**



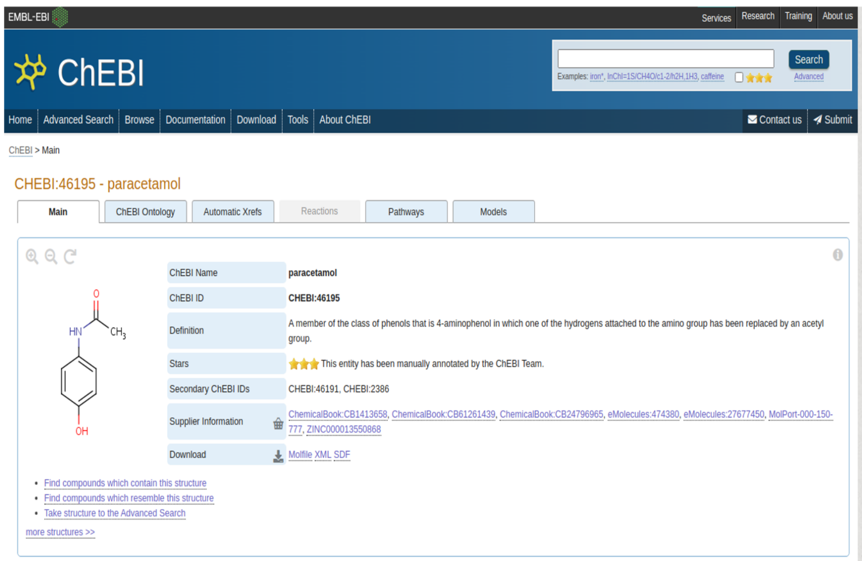
LIBRETTI

knowledge base



driven by excellence

knowledge base with information about research articles mentioning the entities



Citations

Cohen IV, Cirulli ET, Mitchell MW, Jonsson TJ, Yu J, Shah N, Spector TD, Guo L, Venter JC, Telenti A (2018)
Acetaminophen (Paracetamol) Use Modifies the Sulfation of Sex Hormones.
EBioMedicine **28,** *316-323* [PubMed:29398597]
[show Abstract]

Lee WM (2017)
Acetaminophen (APAP) hepatotoxicity-Isn't it time for APAP to go away?
Journal of hepatology **67,** *1324-1331* [PubMed:28734939]
[show Abstract]

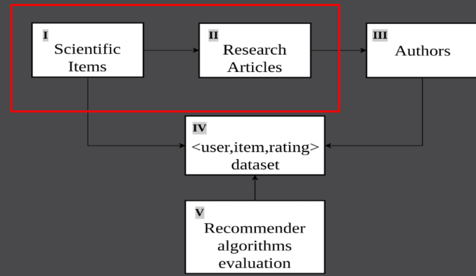
Sawaguchi A, Sasaki K, Miyanaga K, Nakayama M, Nagasue M, Shimoda M (2016)
Rapid absorption of diclofenac and acetaminophen after their oral administration to cattle.
The Journal of veterinary medical science **78,** *1481-1485* [PubMed:27320817]
[show Abstract]

Ciências ULisboa

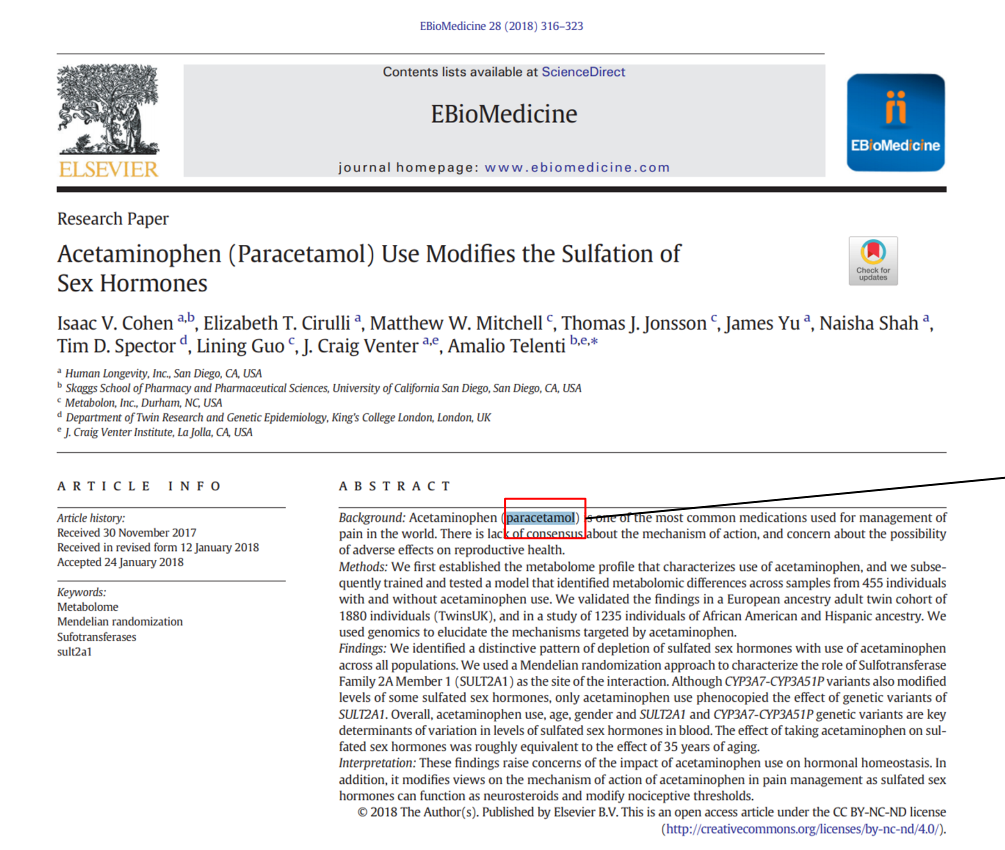
FCT Fundação para a Ciência e a Tecnologia

LIBRETTI

NER + NEL



NER + NEL



EBioMedicine 28 (2018) 316–323

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Acetaminophen (Paracetamol) Use Modifies the Sulfation of Sex Hormones

Isaac V. Cohen [a,b], Elizabeth T. Cirulli [a], Matthew W. Mitchell [c], Thomas J. Jonsson [c], James Yu [a], Naisha Shah [a], Tim D. Spector [d], Lining Guo [c], J. Craig Venter [a,e], Amalio Telenti [b,e,*]

[a] Human Longevity, Inc., San Diego, CA, USA
[b] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, CA, USA
[c] Metabolon, Inc., Durham, NC, USA
[d] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[e] J. Craig Venter Institute, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 30 November 2017
Received in revised form 12 January 2018
Accepted 24 January 2018

Keywords:
Metabolome
Mendelian randomization
Sufotransferases
sult2a1

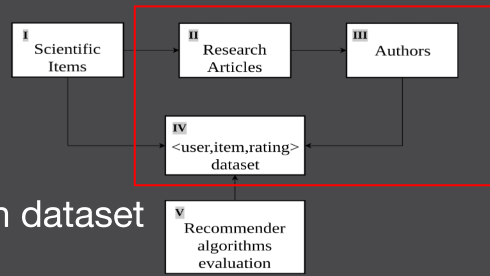
ABSTRACT

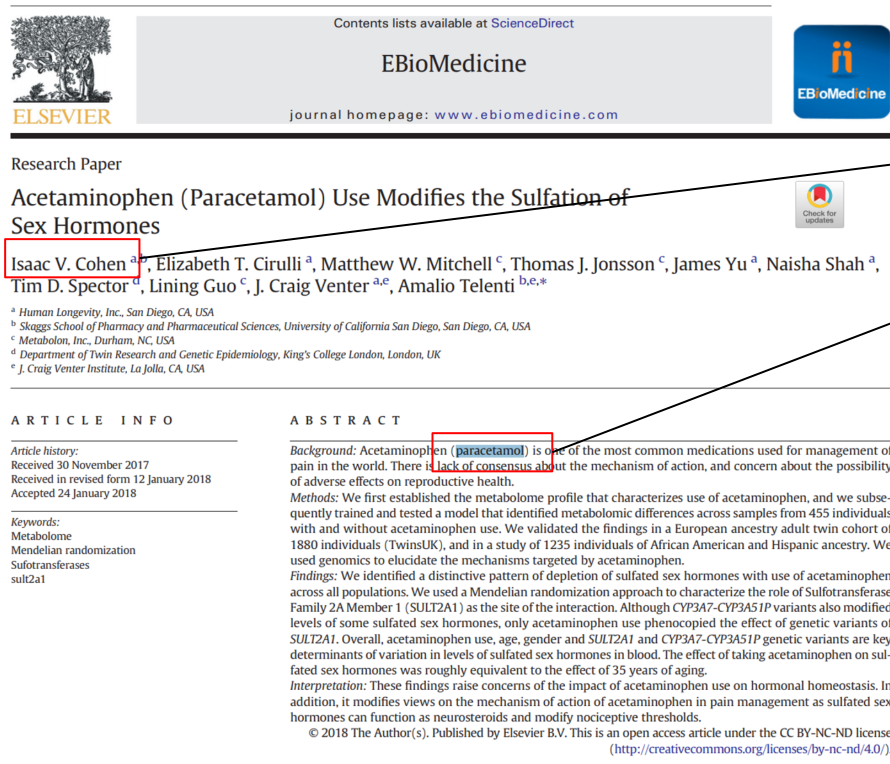
Background: Acetaminophen (paracetamol) is one of the most common medications used for management of pain in the world. There is lack of consensus about the mechanism of action, and concern about the possibility of adverse effects on reproductive health.
Methods: We first established the metabolome profile that characterizes use of acetaminophen, and we subsequently trained and tested a model that identified metabolomic differences across samples from 455 individuals with and without acetaminophen use. We validated the findings in a European ancestry adult twin cohort of 1880 individuals (TwinsUK), and in a study of 1235 individuals of African American and Hispanic ancestry. We used genomics to elucidate the mechanisms targeted by acetaminophen.
Findings: We identified a distinctive pattern of depletion of sulfated sex hormones with use of acetaminophen across all populations. We used a Mendelian randomization approach to characterize the role of Sulfotransferase Family 2A Member 1 (SULT2A1) as the site of the interaction. Although *CYP3A7-CYP3A51P* variants also modified levels of some sulfated sex hormones, only acetaminophen use phenocopied the effect of genetic variants of *SULT2A1*. Overall, acetaminophen use, age, gender and *SULT2A1* and *CYP3A7-CYP3A51P* genetic variants are key determinants of variation in levels of sulfated sex hormones in blood. The effect of taking acetaminophen on sulfated sex hormones was roughly equivalent to the effect of 35 years of aging.
Interpretation: These findings raise concerns of the impact of acetaminophen use on hormonal homeostasis. In addition, it modifies views on the mechanism of action of acetaminophen in pain management as sulfated sex hormones can function as neurosteroids and modify nociceptive thresholds.

Chemical compounds
chEBI ID: **CHEBI:46195**

LIBRETTI

From the articles to the recommendation dataset





Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Acetaminophen (Paracetamol) Use Modifies the Sulfation of Sex Hormones

Isaac V. Cohen [a,b], Elizabeth T. Cirulli [a], Matthew W. Mitchell [c], Thomas J. Jonsson [c], James Yu [a], Naisha Shah [a], Tim D. Spector [d], Lining Guo [c], J. Craig Venter [a,e], Amalio Telenti [b,e,*]

[a] Human Longevity, Inc., San Diego, CA, USA
[b] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, CA, USA
[c] Metabolon, Inc., Durham, NC, USA
[d] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[e] J. Craig Venter Institute, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 30 November 2017
Received in revised form 12 January 2018
Accepted 24 January 2018

Keywords:
Metabolome
Mendelian randomization
Sufotransferases
sult2a1

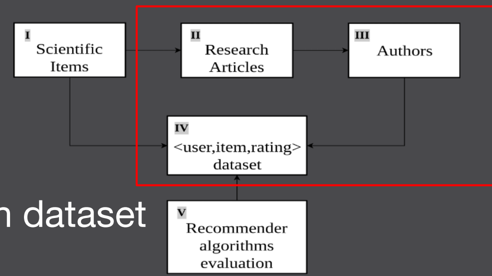
ABSTRACT

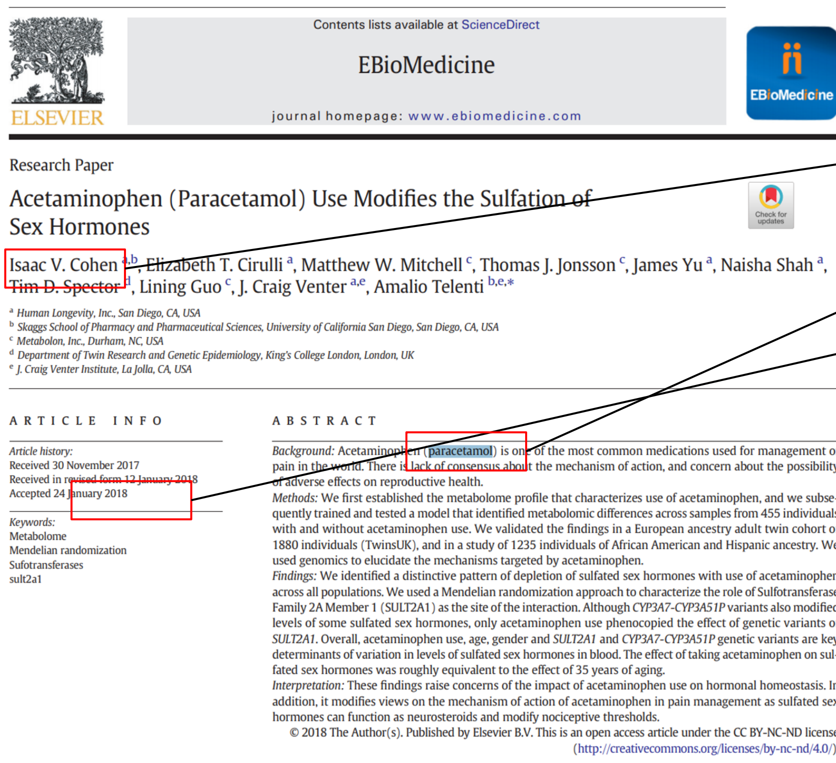
Background: Acetaminophen (paracetamol) is one of the most common medications used for management of pain in the world. There is lack of consensus about the mechanism of action, and concern about the possibility of adverse effects on reproductive health.
Methods: We first established the metabolome profile that characterizes use of acetaminophen, and we subsequently trained and tested a model that identified metabolomic differences across samples from 455 individuals with and without acetaminophen use. We validated the findings in a European ancestry adult twin cohort of 1880 individuals (TwinsUK), and in a study of 1235 individuals of African American and Hispanic ancestry. We used genomics to elucidate the mechanisms targeted by acetaminophen.
Findings: We identified a distinctive pattern of depletion of sulfated sex hormones with use of acetaminophen across all populations. We used a Mendelian randomization approach to characterize the role of Sulfotransferase Family 2A Member 1 (SULT2A1) as the site of the interaction. Although *CYP3A7-CYP3A51P* variants also modified levels of some sulfated sex hormones, only acetaminophen use phenocopied the effect of genetic variants of *SULT2A1*. Overall, acetaminophen use, age, gender and *SULT2A1* and *CYP3A7-CYP3A51P* genetic variants are key determinants of variation in levels of sulfated sex hormones in blood. The effect of taking acetaminophen on sulfated sex hormones was roughly equivalent to the effect of 35 years of aging.
Interpretation: These findings raise concerns of the impact of acetaminophen use on hormonal homeostasis. In addition, it modifies views on the mechanism of action of acetaminophen in pain management as sulfated sex hormones can function as neurosteroids and modify nociceptive thresholds.

User	Item	Rating
Isaac Cohen	Paracetamol	1
Elizabeth Cirulli	Paracetamol	1
Matthew Mitchell	Paracetamol	1
Thomas Jonsson	Paracetamol	1
James Yu	Paracetamol	1
Naisha Shah	Paracetamol	1
Tim Spector	Paracetamol	1
Lining Guo	Paracetamol	1
Craig Venter	Paracetamol	1
Amalio Telenti	Paracetamol	1

LIBRETTI

From the articles to the recommendation dataset





Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Acetaminophen (Paracetamol) Use Modifies the Sulfation of Sex Hormones

Isaac V. Cohen [a,b], Elizabeth T. Cirulli [a], Matthew W. Mitchell [c], Thomas J. Jonsson [c], James Yu [a], Naisha Shah [a], Tim D. Spector [d], Lining Guo [c], J. Craig Venter [a,e], Amalio Telenti [b,e,*]

[a] Human Longevity, Inc., San Diego, CA, USA
[b] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, CA, USA
[c] Metabolon, Inc., Durham, NC, USA
[d] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[e] J. Craig Venter Institute, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 30 November 2017
Received in revised form 12 January 2018
Accepted 24 January 2018

Keywords:
Metabolome
Mendelian randomization
Sufotransferases
sult2a1

ABSTRACT

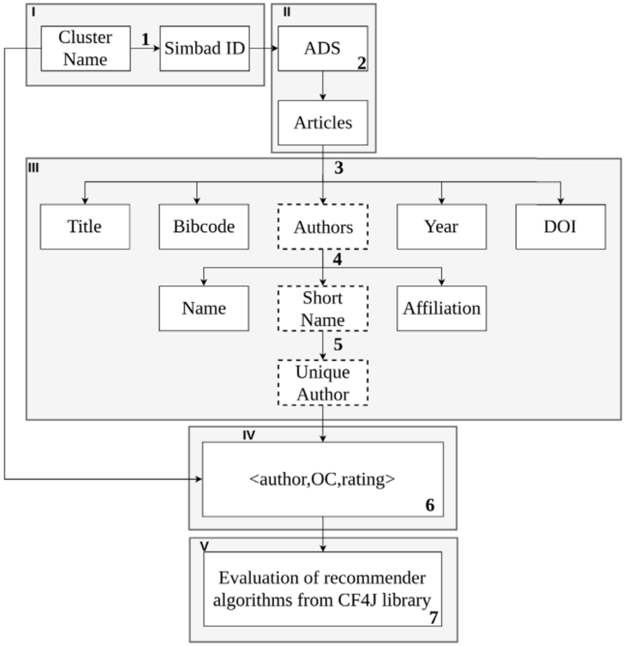
Background: Acetaminophen (paracetamol) is one of the most common medications used for management of pain in the world. There is lack of consensus about the mechanism of action, and concern about the possibility of adverse effects on reproductive health.
Methods: We first established the metabolome profile that characterizes use of acetaminophen, and we subsequently trained and tested a model that identified metabolomic differences across samples from 455 individuals with and without acetaminophen use. We validated the findings in a European ancestry adult twin cohort of 1880 individuals (TwinsUK), and in a study of 1235 individuals of African American and Hispanic ancestry. We used genomics to elucidate the mechanisms targeted by acetaminophen.
Findings: We identified a distinctive pattern of depletion of sulfated sex hormones with use of acetaminophen across all populations. We used a Mendelian randomization approach to characterize the role of Sulfotransferase Family 2A Member 1 (SULT2A1) as the site of the interaction. Although *CYP3A7-CYP3A51P* variants also modified levels of some sulfated sex hormones, only acetaminophen use phenocopied the effect of genetic variants of *SULT2A1*. Overall, acetaminophen use, age, gender and *SULT2A1* and *CYP3A7-CYP3A51P* genetic variants are key determinants of variation in levels of sulfated sex hormones in blood. The effect of taking acetaminophen on sulfated sex hormones was roughly equivalent to the effect of 35 years of aging.
Interpretation: These findings raise concerns of the impact of acetaminophen use on hormonal homeostasis. In addition, it modifies views on the mechanism of action of acetaminophen in pain management as sulfated sex hormones can function as neurosteroids and modify nociceptive thresholds.

User	Item	Rating	year
Isaac Cohen	Paracetamol	1	2018
Elizabeth Cirulli	Paracetamol	1	2018
Matthew Mitchell	Paracetamol	1	2018
Thomas Jonsson	Paracetamol	1	2018
James Yu	Paracetamol	1	2018
Naisha Shah	Paracetamol	1	2018
Tim Spector	Paracetamol	1	2018
Lining Guo	Paracetamol	1	2018
Craig Venter	Paracetamol	1	2018
Amalio Telenti	Paracetamol	1	2018

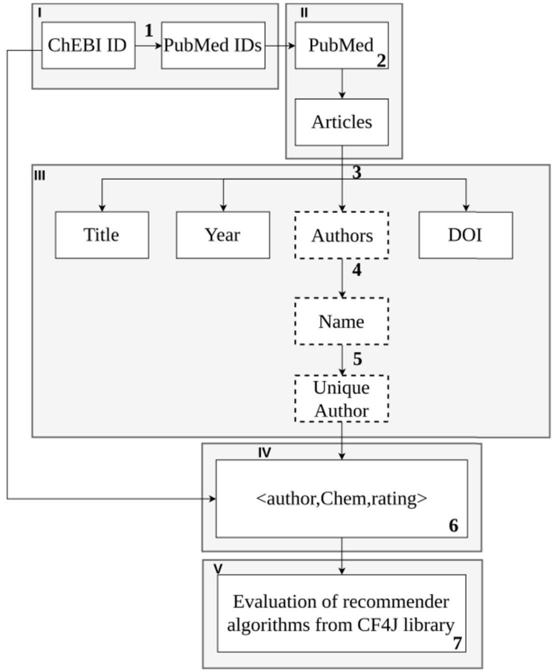
LIBRETTI

Application fields

Astronomy



Chemistry



Barros, Márcia, André Moitinho, and Francisco M. Couto. "Using Research Literature to Generate Datasets of Implicit Feedback for Recommending Scientific Items." IEEE Access 7 (2019): 176668-176680. Q1 Scimago

COVID-19

NER+NEL
CHEBI
Gene Ontology
Disease Ontology
Human Phenotype Ontology

<User, Item, Rating>

CORD-19 dataset

Items from multi scientific fields

Recommendation dataset

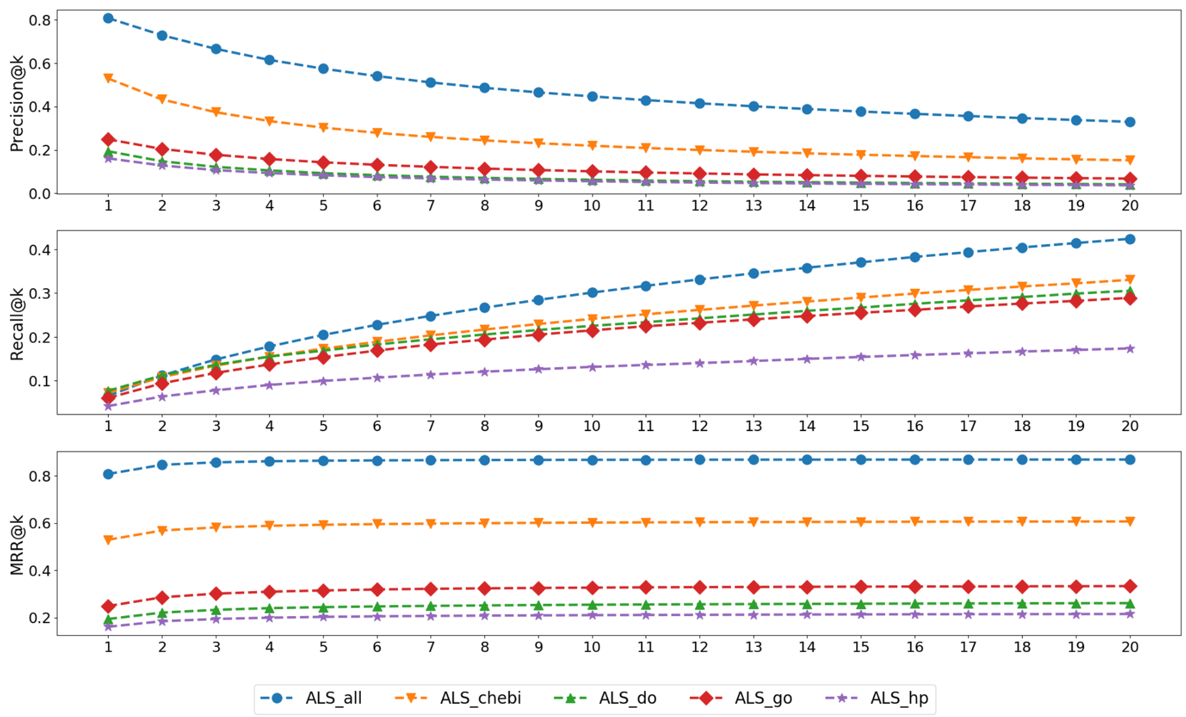
Barros, M. A., Lamurias, A., Sousa, D. F., Ruas, P., & Couto, F. M. (2020). COVID-19: A Semantic-Based Pipeline for Recommending Biomedical Entities. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020

LIBRETTI

Application fields

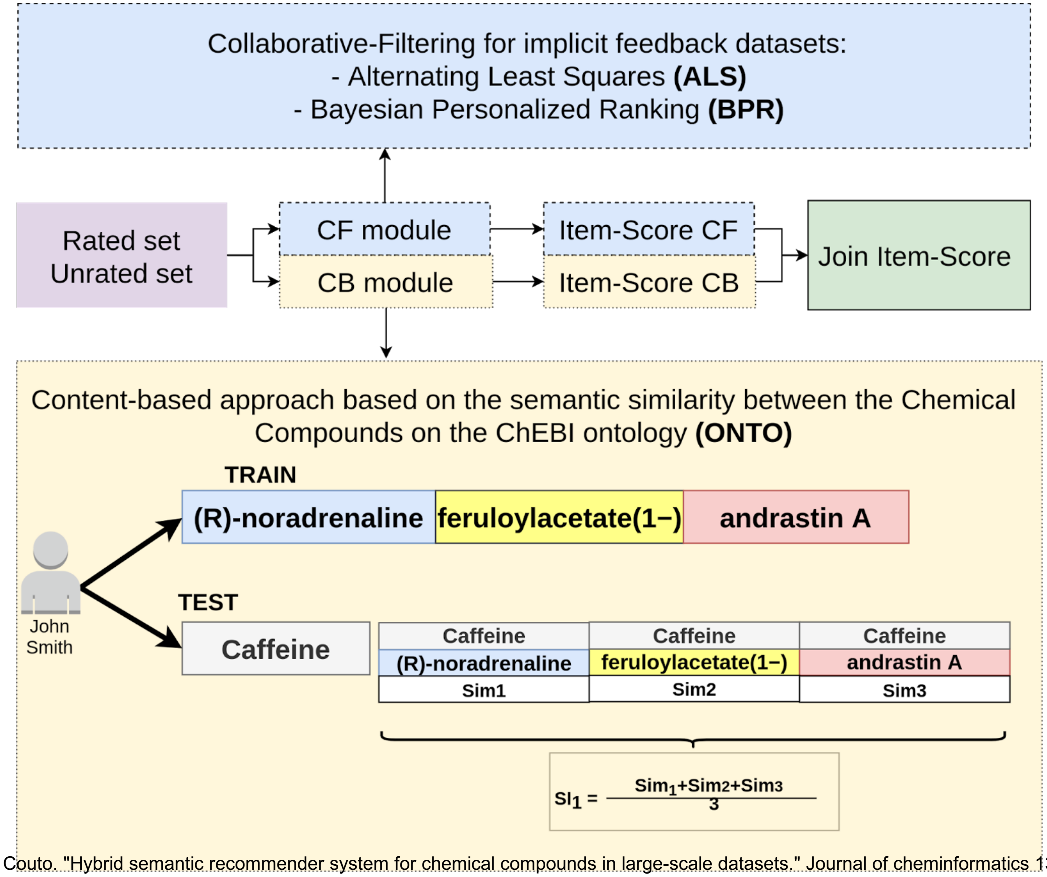
COVID-19

Using ontologies from different fields in the NER phase, we improve the results for state-of-the-art collaborative-filtering recommender systems applied to the dataset created.



Barros, M. A., Lamurias, A., Sousa, D. F., Ruas, P., & Couto, F. M. (2020). COVID-19: A Semantic-Based Pipeline for Recommending Biomedical Entities. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020

ONTO

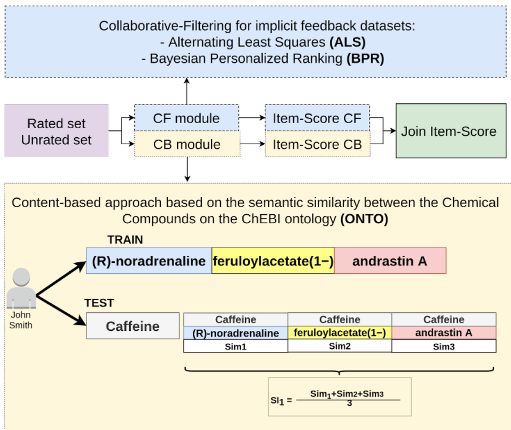


Chemistry

Barros, Marcia, Andre Moitinho, and Francisco M. Couto. "Hybrid semantic recommender system for chemical compounds in large-scale datasets." Journal of cheminformatics 13.1 (2021): 1-18.

LASIGE driven by excellence

ONTO



MRR@k

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ALS	0.56	0.58	0.59	0.60	0.60	0.60	0.60	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
ALS_ONTO_JC_m1	0.61	0.64	0.64	0.65	0.65	0.65	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
ALS_ONTO_JC_m2	0.59	0.63	0.65	0.65	0.66	0.66	0.66	0.66	0.66	0.66	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
ALS_ONTO_LIN_m1	0.60	0.62	0.63	0.64	0.64	0.64	0.64	0.64	0.64	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
ALS_ONTO_LIN_m2	0.63	0.65	0.66	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.68	0.68	0.68	0.68	0.68	0.68
ALS_ONTO_RESNIK_m1	0.60	0.62	0.63	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
ALS_ONTO_RESNIK_m2	0.49	0.53	0.54	0.54	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
BPR	0.39	0.41	0.43	0.43	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
BPR_ONTO_JC_m1	0.48	0.53	0.55	0.55	0.56	0.56	0.56	0.56	0.56	0.56	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
BPR_ONTO_JC_m2	0.42	0.45	0.46	0.47	0.47	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
BPR_ONTO_LIN_m1	0.49	0.53	0.54	0.54	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
BPR_ONTO_LIN_m2	0.47	0.50	0.52	0.52	0.53	0.53	0.53	0.53	0.53	0.53	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.54
BPR_ONTO_RESNIK_m1	0.49	0.53	0.54	0.54	0.55	0.55	0.55	0.55	0.55	0.55	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
BPR_ONTO_RESNIK_m2	0.52	0.56	0.57	0.58	0.58	0.58	0.58	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
ONTO_JC	0.24	0.30	0.33	0.35	0.35	0.36	0.36	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.38	0.38	0.38	0.38
ONTO_LIN	0.30	0.37	0.39	0.40	0.40	0.41	0.41	0.41	0.41	0.41	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
ONTO_RESNIK	0.32	0.38	0.40	0.41	0.41	0.42	0.42	0.42	0.42	0.42	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43

0.68 0.57 0.46 0.35 0.24

Barros, Marcia, Andre Moitinho, and Francisco M. Couto. "Hybrid semantic recommender system for chemical compounds in large-scale datasets." Journal of cheminformatics 13.1 (2021): 1-18.

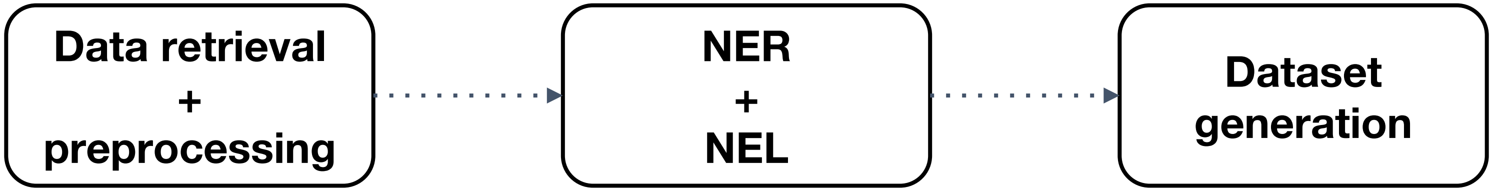
Ciências ULisboa FCT Fundação para a Ciência e a Tecnologia

PART 2

How to build a scientific recommendation dataset?

LASIGE driven by excellence

Tutorial sections



Source: >> **git clone** git@github.com:lasigeBioTM/RecSys.Scifi.tutorial.git

Ciências ULisboa
FCT Fundação para a Ciência e a Tecnologia

THANK YOU!!!