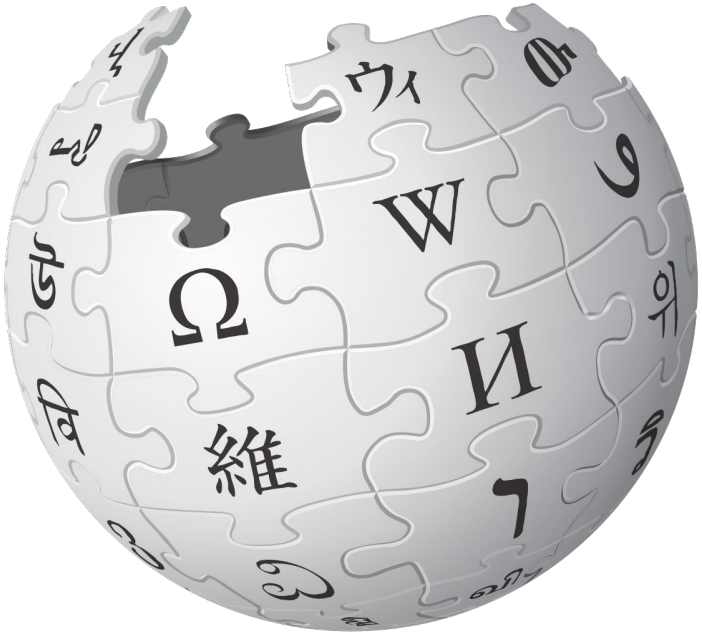
Structuring Wikipedia Articles with Section Recommendations

Tiziano Piccardi, Michele Catasta, Leila Zia, Robert West



Wikipedia Facts

5th most visited website

48M articles in 300+ languages

2.4B edits

78.8M editors

Not integrated

Poor structure

Limited content

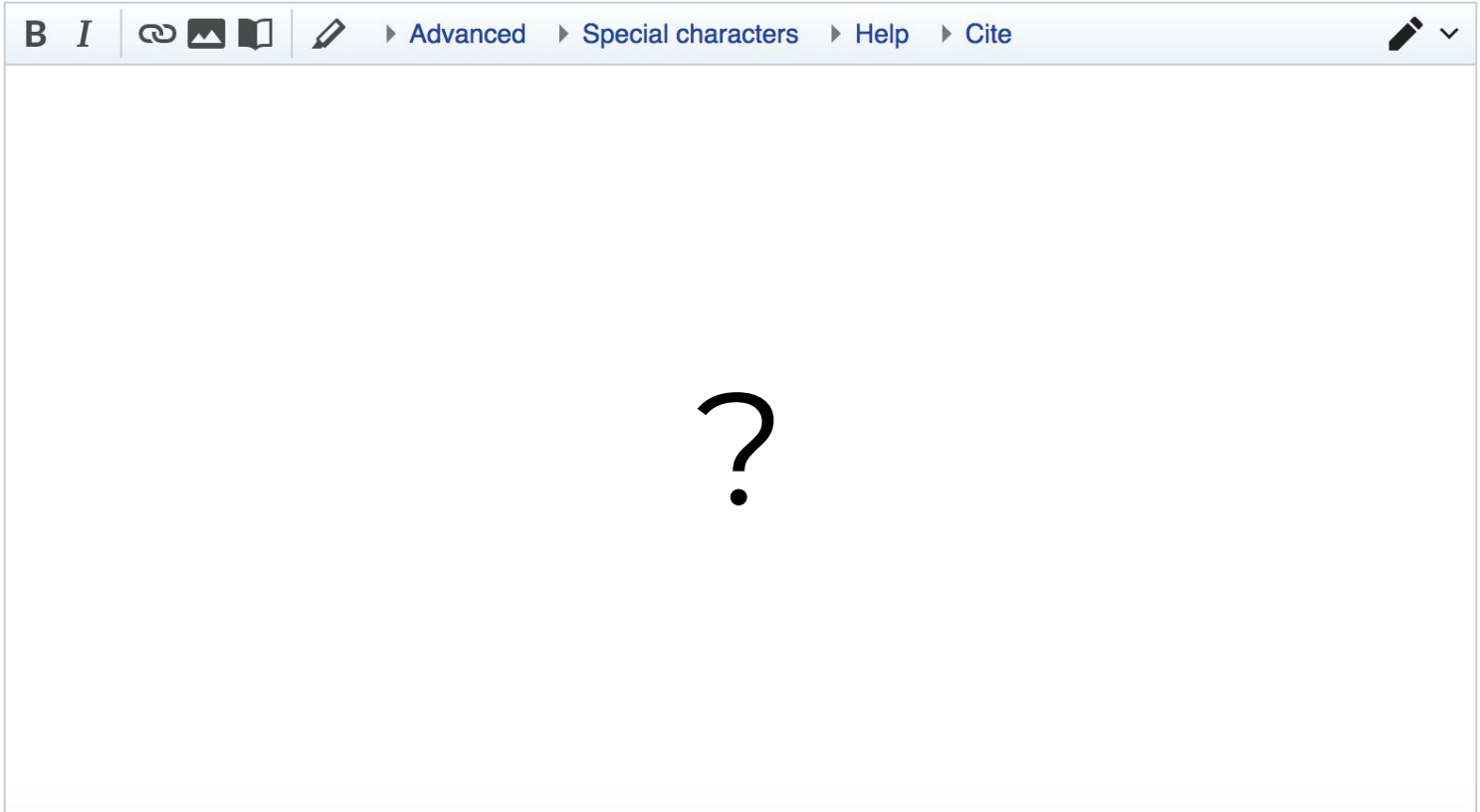
More than 2M+ (37%) articles in English Wikipedia are stubs

No references

Not clear

Editing SIGIR 2018

Content that violates any copyrights will be deleted. Encyclopedic content must be verifiable. Work submitted to Wikipedia can be edited, used, and redistributed—by anyone—subject to certain terms and conditions.



Leonardo DiCaprio

From Wikipedia, the free encyclopedia

Leonardo Wilhelm DiCaprio (/dɪˈkæpriˌoʊ/; born November 11, 1974) is an American actor and film producer.

Contents [hide]

Help to structure the content

Article-based

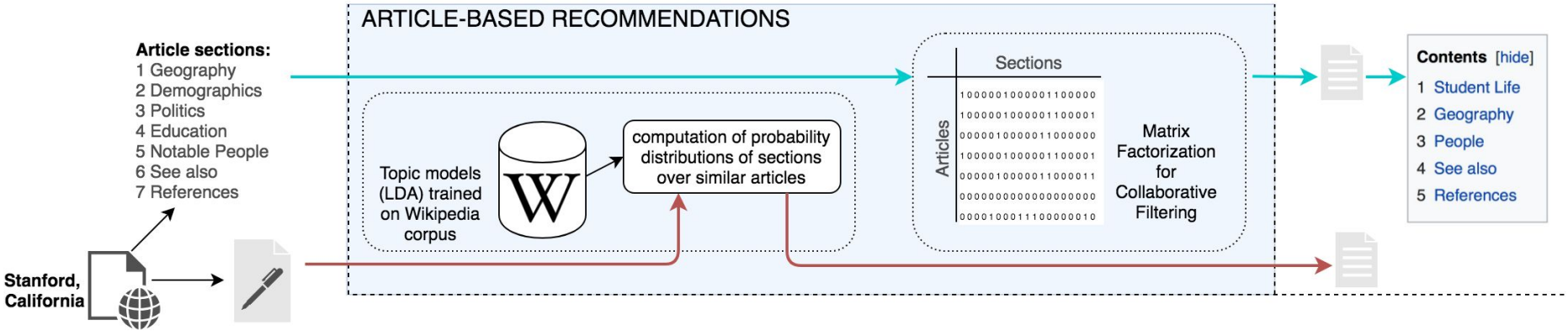
We use the content of the article to generate recommendations

---

Category-based

We use the category network to generate recommendations

Article-based



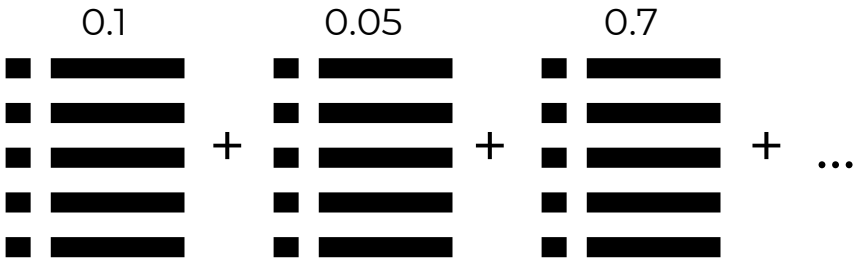
Article-based

Topic Modeling

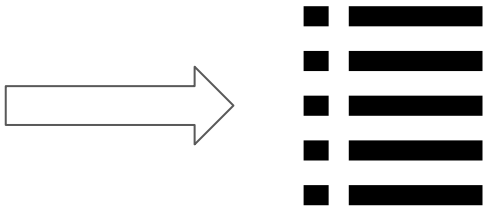
Topic 1 (0.1)
Topic 2 (0.05)
Topic 3 (0.7)
...

Latent Dirichlet Allocation (LDA)
200 topics

Article topics



Final recommendation



Article-based

Articles \ Sections						
	1	0	1	0	0	0
	0	0	0	1	0	0
	1	0	1	0	1	0
	0	1	0	1	0	1
	0	0	0	1	0	0
	1	1	1	0	0	0

Collaborative Filtering

Based on matrix factorization with Alternating Least Squares

One row per article and one column per section

1 if the section S appears in the article A

Limitation:

The article-based approach cannot generate recommendations for new articles!

Category-based

Leonardo DiCaprio

From Wikipedia, the free encyclopedia

Leonardo Wilhelm DiCaprio (/dɪˈkæprioʊ/; born November 11, 1974) is an American actor and film producer.

DiCaprio began his career by appearing in television commercials in the late 1980s. He next had recurring roles in various television series, such as the soap opera *Santa Barbara* and the sitcom *Growing Pains*.

He debuted in his film career by starring as Josh in *Critters 3* (1991). He starred in the film adaptation of the memoir *This Boy's Life* (1993), and was praised for his supporting role in *What's Eating Gilbert Grape* (1993). He gained public recognition with leading roles in *The Basketball Diaries* (1995) and the romantic drama *Romeo + Juliet* (1996). He achieved international fame as a star in James Cameron's epic romance *Titanic* (1997), which became the highest-grossing film of all time to that point.

Leonardo DiCaprio

Since 2000, DiCaprio has received c
DiCaprio's subsequent films include
drama *Catch Me If You Can* (2002), a
which marked his first of many collab
acclaimed for his performances in the
drama *The Departed* (both 2006), the
Revolutionary Road (2008), the psyc
thriller *Inception* (2010), the biograph
(2012), and the period drama *The Gr*

Categories: Leonardo DiCaprio | 1974 births | Living people | 20th-century American male actors | 21st-century American male actors | American environmentalists | American film producers | American male child actors | American male film actors | American male soap opera actors | American male television actors | Best Actor AACTA Award winners | Best Actor AACTA International Award winners | Best Actor Academy Award winners | Best Actor BAFTA Award winners | Best Drama Actor Golden Globe (film) winners | Best Musical or Comedy Actor Golden Globe (film) winners | Contestants on American game shows | Silver Bear for Best Actor winners | Outstanding Performance by a Male Actor in a Leading Role Screen Actors Guild Award winners | American agnostics | Film producers from California | Male actors from Hollywood, Los Angeles | Male actors from Palm Springs, California | People from Echo Park, Los Angeles

Avatar (2009 film)

From Wikipedia, the free encyclopedia

Contents [hide]

1 Plot
2 Cast
3 Production
4 Marketing
5 Release
6 Sequels
7 Stage adaptation
8 See also
9 References
10 Further reading
11 External links

Theatrical release poster

Directed by James Cameron

Produced by James Cameron
Jon Landau

Plot

By 2154, humans have depleted Earth's natural resources, leading to a severe energy crisis. The

Categories: 2009 films | English-language films | Avatar (2009 film)
2000s 3D films | 2000s action films | American adventure films
Science fiction adventure films | 2000s adventure films
20th Century Fox films | 22nd century in fiction | American epic films

300 (film)

From Wikipedia, the free encyclopedia

Contents [hide]

1 Plot
2 Cast
3 Production
4 Soundtrack
5 Promotion and release
6 Reception
7 Depictions of Persians and Iran's reaction
8 In popular culture
9 Sequel
10 See also
11 References
12 External links

Theatrical release poster

Directed by Zack Snyder

Produced by Mark Canton
Bernie Goldmann
Gianni Nunnari
Jeffrey Silver

Plot [edit]

In 479 BC, one year after the famed Battle of Thermopylae, Dilios, a hoplite in the Spartan

Categories: 2006 films | English-language films | 2000s action films
2000s adventure films | 2000s fantasy films | 2000s war films
American action films | American epic films | American war films
American fantasy adventure films | American films

Intuition:

Articles in the same category share similar sections

We can use the categories to generate templates for the editors

Category:American epic films

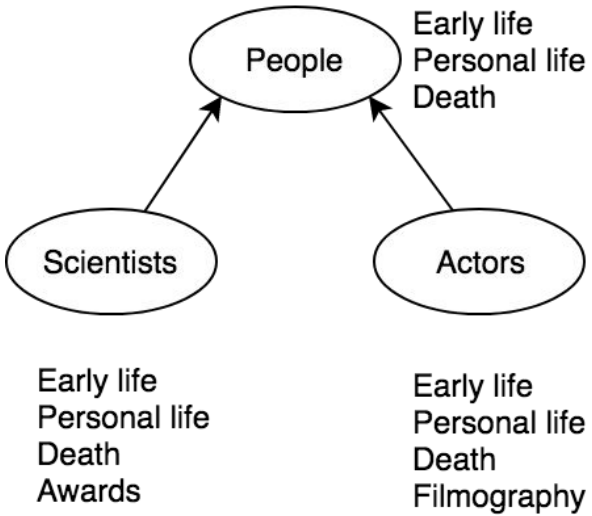
▼

{*Plot, Cast, Production*}

Taxonomic assumption

Categories are organised in a hierarchical structure

Frequent sections on the children may be relevant for the parent

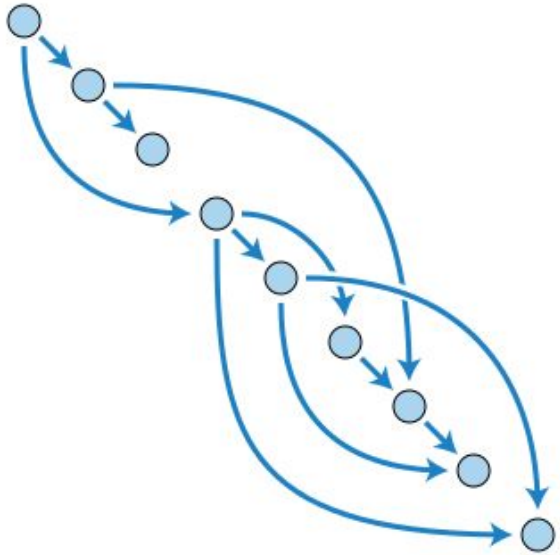


Wait, it’s not so easy...

The category network has loops!

Government ➜ Public administration ➜ Public economics ➜ Economic policy ➜ Government

---

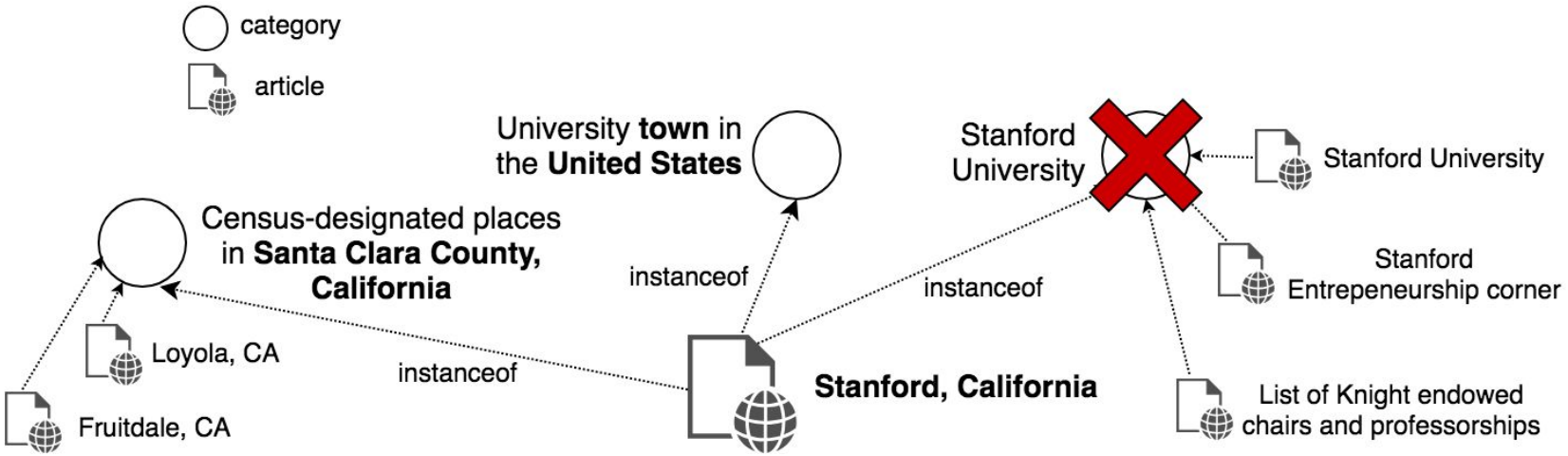


A fast and effective heuristic for the feedback arc set problem

Peter Eades, Xuemin Lin, W.F. Smyth

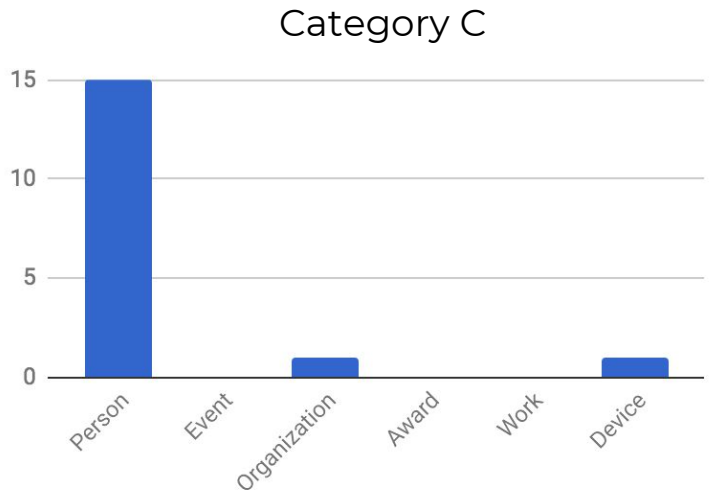
Removed 4k edges out of ~4M

IS-A relation is not always respected!



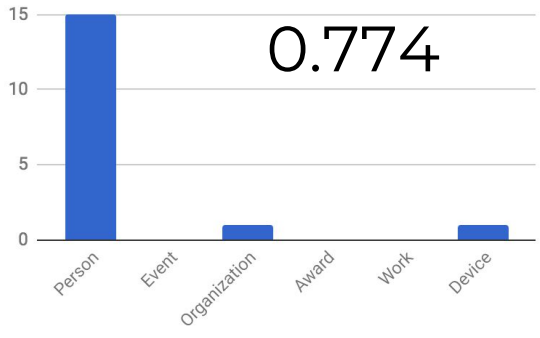
Categories with heterogeneous articles must be removed

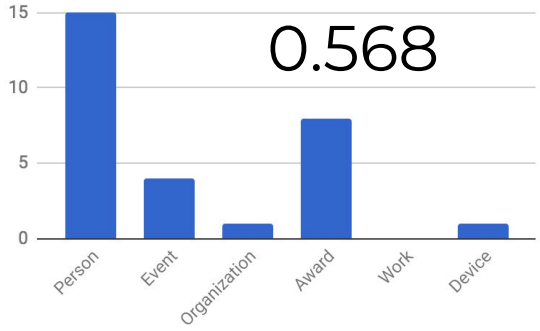
We assigned 55 top level types to the articles



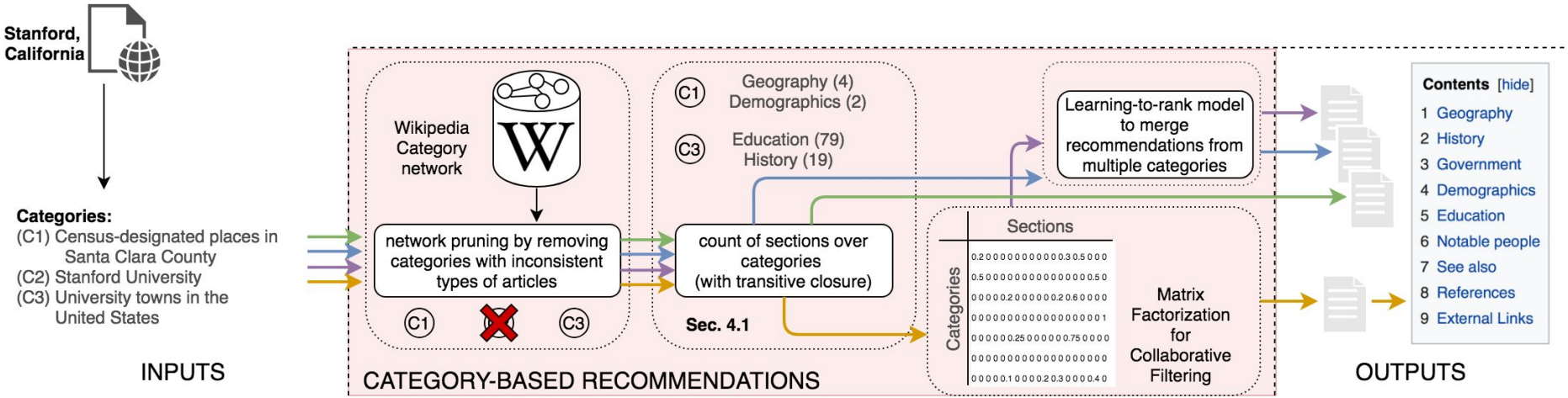
Distribution of the article types in a category

Gini coefficient
to select the categories to keep in the network

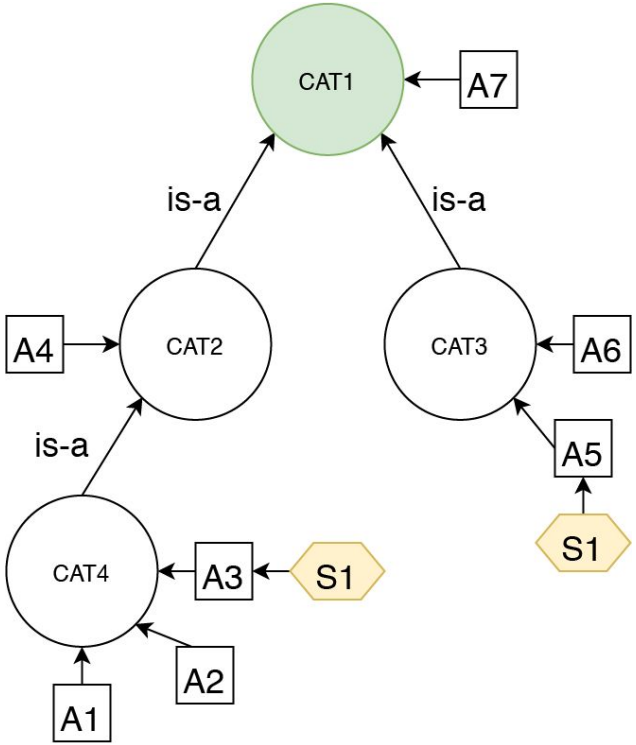




Category-based



Category-based



P(S1 | CAT1) = 2/7

Category–Section counts

Probability P(S | C) of observing section S in category C

Example:

American_male_film_actors
Filmography: 0.59
Career: 0.47
Personal life: 0.38
...

Filmography appears in 59% of the articles in the category "*American_male_film_actors*"

Category-based

Categories \ Sections						
	7/10	0	3/10	0	0	0
	0	0	0	1	0	0
	2/8	0	1/8	0	5/8	0
	0	2/8	0	2/8	0	4/8
	0	0	0	1	0	0
	2/9	3/9	4/9	0	0	0

Collaborative Filtering

Based on matrix factorization with Alternating Least Squares

One row per category and one column per section

Ratings defined as P(S1 | CAT1)

Merging phase

Category-based

Categories_of(*Leonardo DiCaprio*) =
{*American_male_film_actors*,
American_film_producers
Living_people}

American_male_film_actors
Filmography: 0.59
Career: 0.47
Personal life: 0.38
...

American_film_producers
Filmography: 0.39
Career: 0.34
Personal life: 0.26
...

Living_people
Career: 0.25
Personal life: 0.17
Biography: 0.13
...

$$\sum_{ij}(P(S_i|C_j)) \quad C_j \in Categories\_of(Leonardo\ DiCaprio)$$

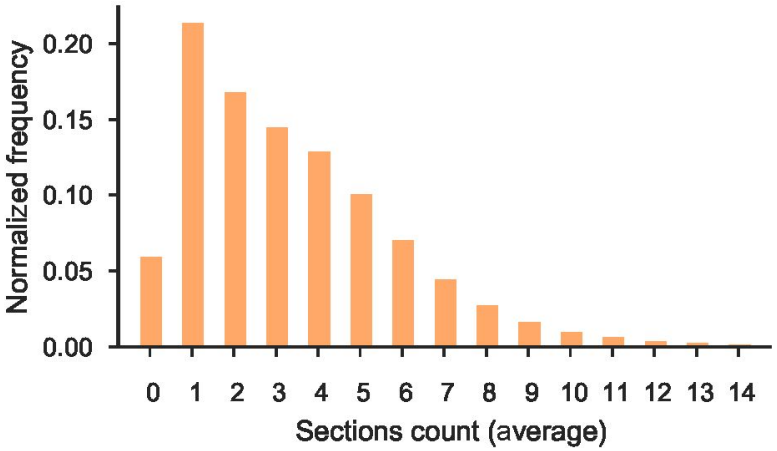
Learning2Rank:
weighted sum

sort

Leonardo DiCaprio
Career: 9.98
Filmography: 9.51
Personal life: 9.48
Early life: 8.40
Awards: 2.45
...

Evaluation

English Wikipedia - September 2017
5.5M articles
300K sections



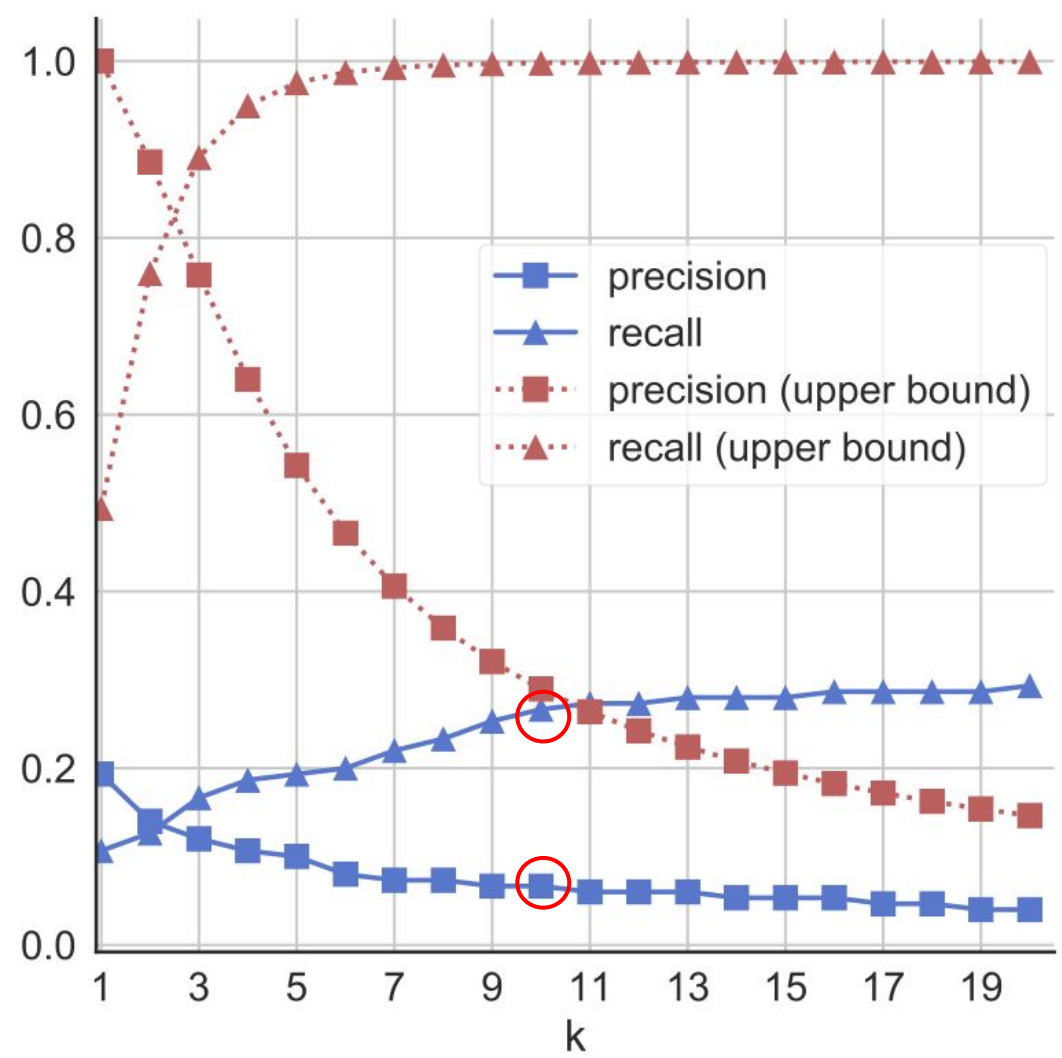
Cold start problem: in average 3.4 sections

Article-based

Collaborative filtering

Precision < 0.2%

Recall < 1.5%



Article-based

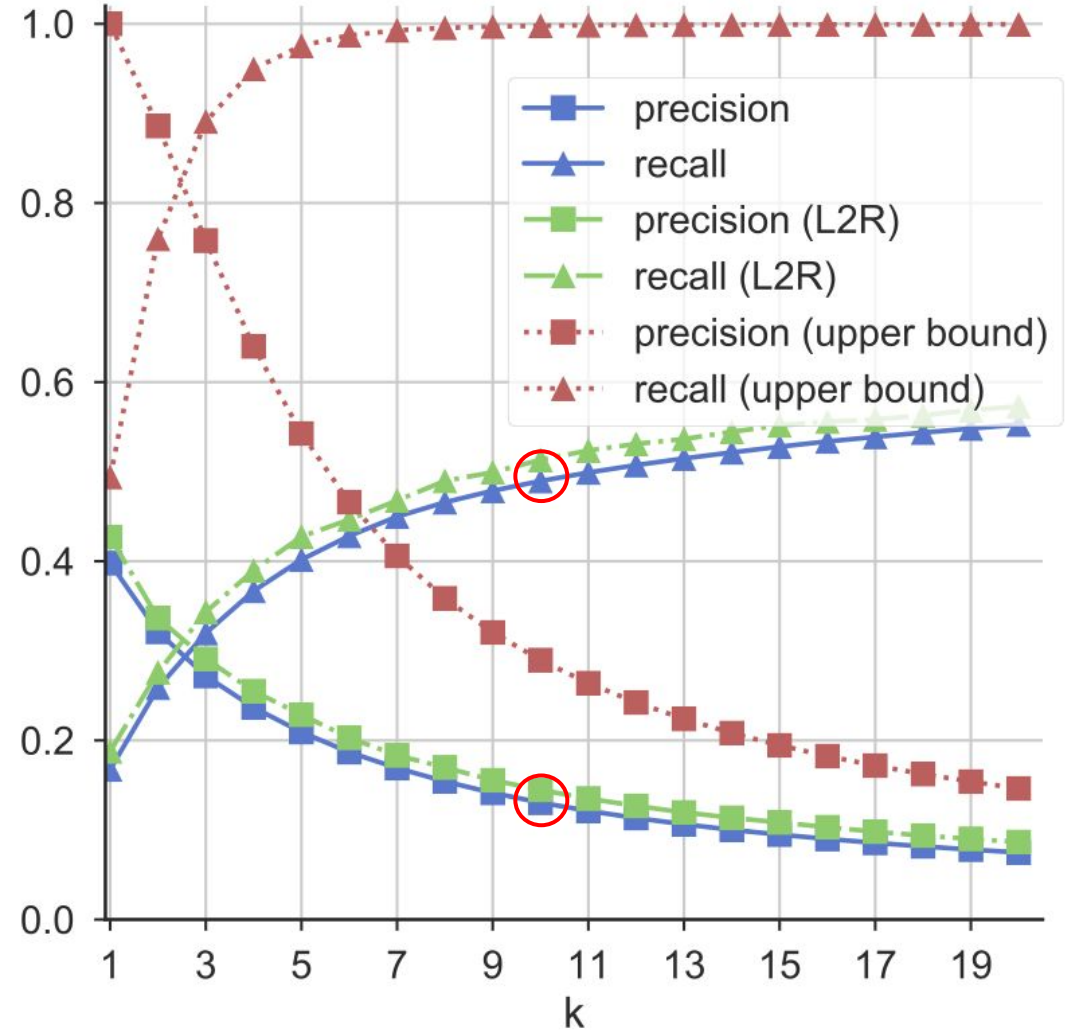
Topic modeling

Precision@10 = 6%

(upper bound 28%)

Recall@10 = 26%

(upper bound 98%)



Category-based

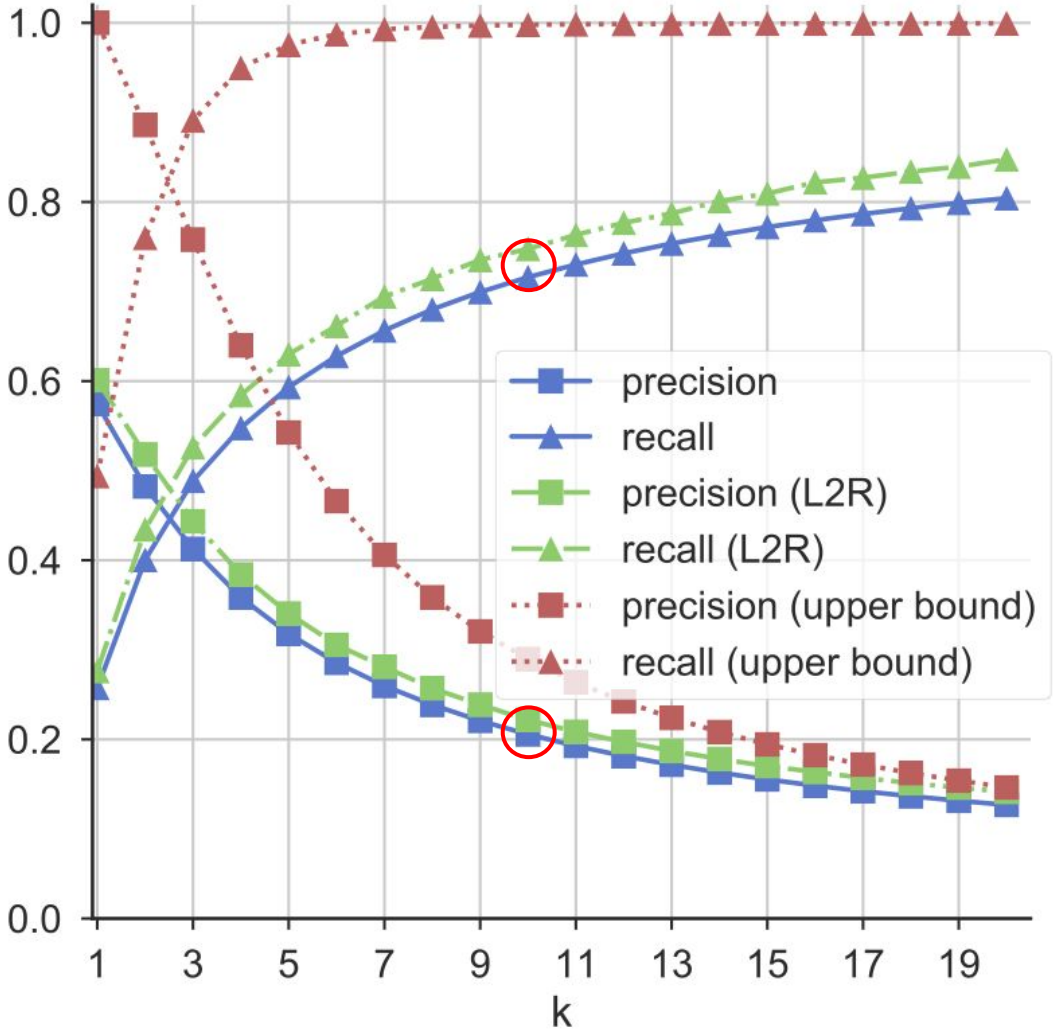
Collaborative filtering

Precision@10 = 13%

(upper bound 28%)

Recall@10 = 49%

(upper bound 98%)



Category-based

Category–Section counts

Precision@10 = 20%

(upper bound 28%)

Recall@10 = 72%

(upper bound 98%)

Automatic evaluation has limitations!

The testing set contains articles with the problem we want to solve

few sections | inconsistent | different syntax

Human evaluation

Wikipedia editors

Crowd-workers

Instructions

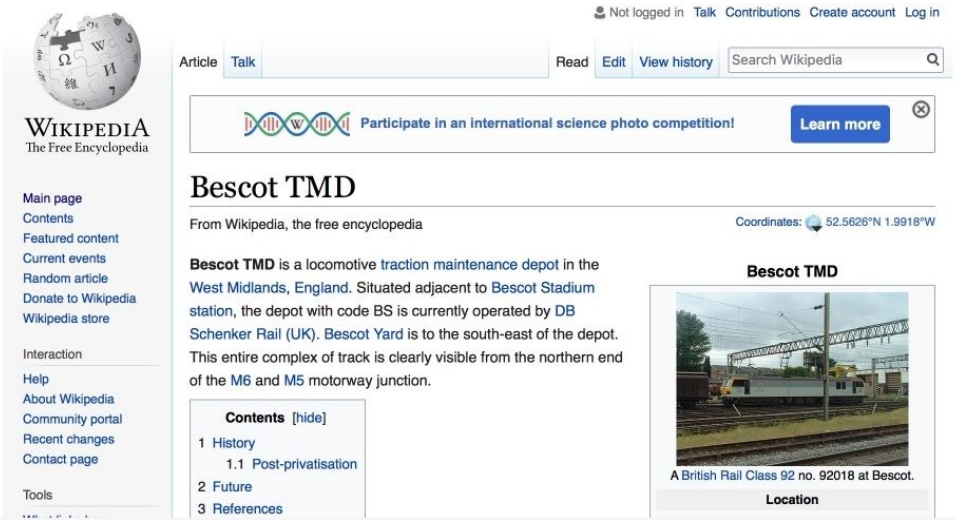
Tell us if the article section

Opening

should be part of the Wikipedia article below.
Don't mind the duplicates -- if the section is already part of the Table of Content (or a very similar one is) you should still select "Relevant section"!

Relevant section (S) | NOT relevant section (K)

You have already evaluated 2 out of 17 sections for this article.
Note that we register your responses only **after** you evaluated all the 17 sections.
Use the keyboard shortcut **S** and **K** if you want to be more efficient!



Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Participate in an international science photo competition! Learn more

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Bescot TMD

From Wikipedia, the free encyclopedia

Coordinates: 52.5626°N 1.9918°W

Bescot TMD is a locomotive traction maintenance depot in the West Midlands, England. Situated adjacent to Bescot Stadium station, the depot with code BS is currently operated by DB Schenker Rail (UK). Bescot Yard is to the south-east of the depot. This entire complex of track is clearly visible from the northern end of the M6 and M5 motorway junction.

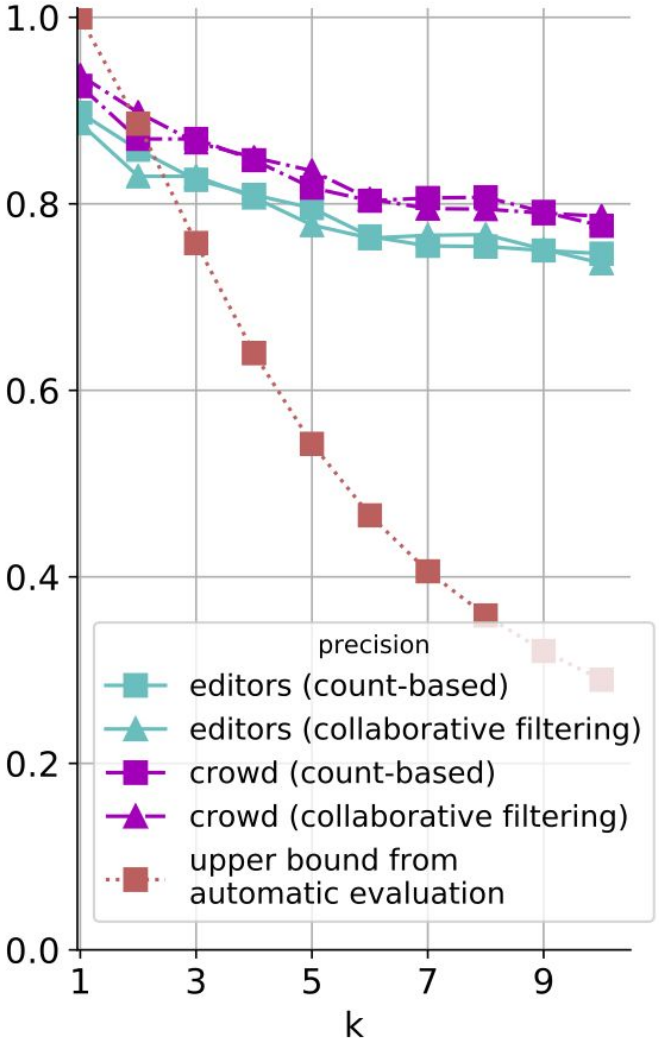
Contents [hide]
1 History
1.1 Post-privatisation
2 Future
3 References

Bescot TMD

A British Rail Class 92 no. 92018 at Bescot.

Location

This website does **not** collect any personal information about the user (e.g., third-party trackers, IP and User-Agent, etc.) We only install an anonymous cookie in the browser to collate all the answers coming from the same user. **Thanks for helping us improve Wikipedia!**



Category-based

Category–Section counts

Human evaluation

Wikipedia editors:
Precision@10 = 72%

Crowd-workers:
Precision@10 = 81%

Conclusion

- Introduced the section recommendation problem
- Explored several methods using
  - features derived from the raw input article
  - Wikipedia's category network
- Learned that category network is key in offering useful recommendations
- We developed a methodology to prune the category network

https://meta.wikimedia.org/wiki/Recommendation_API

https://github.com/epfl-dlab/structuring-wikipedia-articles

Thank You

From Wikipedia, the free encyclopedia

"Thank you" is a common expression of gratitude. It often refers to a thank you letter, a letter written to express appreciation.

tiziano.piccardi@epfl.ch

@tizianopiccardi

https://meta.wikimedia.org/wiki/Recommendation_API

https://github.com/epfl-dlab/structuring-wikipedia-articles