
Leila Zia

Research Showcase
2017-12-13
Hi! It’s been a while!
Let’s rewind.
2015
English Wikipedia (950,277)
Native Speakers 527M
Russian Wikipedia (298,215)
Native Speakers 254M

Spanish Wikipedia (261,495)
Native Speakers 389M

Portuguese Wikipedia (185,133)
Native Speakers 193M
Arabic Wikipedia (87,017)
Native Speakers 467M
All Wikipedia languages

Insights

- We have a lot of missing articles across languages
- Not all of these missing articles are available in at least one Wikipedia language

⇒ Wikipedia alone won’t be the solution for all knowledge gaps.
Goal

Increase article coverage in terms of the number of articles in different languages and the contents of the articles within a language by identifying and prioritizing missing content and routing attention where it’s needed.
Growing Wikipedia Across Languages by Article Creation recommendations

The system

Find → Rank → Recommend

Articles

1
2
3
...

Articles

Editors
Experiment

Articles

100K

100K

100K

personalized

random

held-back

Editors

6K

6K

6K
Results

- 3.2x growth in article creation rate with no loss in article quality

- 2x growth in activation rate if the recommended article is personalized for the editor based on their interests
On the ground

- GapFinder
  https://www.mediawiki.org/wiki/GapFinder
- ContentTranslation Suggestions
- Wherever else you’re using the API :)
GapFinder

Wikipedia GapFinder

agriculture

Environmental impact
agriculture's impact on the environment
5641 recent views

Crop rotation

Monoculture

Intensive animal farming

Agricultural wastewater

Integrated pest management

Renewable resource
a natural resource which can replenish with the passage of time,
12915 recent views

Environmental impact
Crop rotation is the practice of growing a series of dissimilar or different types of crops in the same area in sequenced seasons. It is done so that the soil of farms is not used for only one set of nutrients. It helps in reducing soil erosion and increases soil fertility and crop yield.

Growing the same crop in the same place for many years in a row disproportionately depletes the soil of certain nutrients. With rotation, a crop that leaches the soil of one kind of nutrient is followed during the next growing season by a dissimilar crop that returns that nutrient to the soil or draws a different ratio of nutrients. In addition, crop rotation mitigates the buildup of pathogens and pests that often occurs when one species is continuously cropped, and can also improve soil structure and fertility by increasing biomass from varied root structures.

Crop rotation is used in both conventional and organic farming systems.
Climate change in Malagasy
2016 to now
A series of revelations
Stubs

We were reminded, over and over, that there are many articles already in Wikipedia that are stubs.

“We have enough low-quality stub articles that need human effort to improve and we're not really interested in more unless either (a) they demonstrably combat some of the systematic biases we're struggling with or (b) they demonstrably attract new cohorts users to do that improvement.” -- stuart
Onboarding newcomers

- We learned about The Africa Destubathon project.
- We reached out to User:Anthere to see what kind of research would help them in this kind of project. We learned:

  ⇒ A lot of manual work is going into building templates to help newcomers learn how to expand an already existing article in Wikipedia
Existing initiatives
The numbers

- **Stubs:** 37% of enwiki articles have a stub template
- **Good or better:** Only 1% of enwiki articles have quality class label of Good or better
- **Inconsistency:** 80% of the sections created in enwiki are used only in one article
- **Potential:** 14K accounts get created every month
Sections count for: all articles vs. only high quality (FA, GA and A)

- **High Quality**
- **All articles**

Normalized frequency

Sections count (Average)

0 2 4 6 8 10 12 14

0.00 0.05 0.10 0.15 0.20
Take-away

- We have a lot of missing content in Wikipedia articles.
- Editathon organizers are spending a lot of time on manually creating templates that can help onboard newcomers.
- There are initiatives that have identified these needs and are working on this problem and the challenge for them is how to scale the recommendations beyond very specific topics/types of articles.
Let’s imagine!
Imagine

... we have a system that breaks down the article structure into different pieces (lead paragraph, infobox, sections, images, links, references, ...), provides recommendation for each of these pieces to the editors, and provides more in-depth information about the pieces (such as: “early life section in biographies in enwiki on average contains x characters.”).
Let’s start with sections
Growing Wikipedia Across Languages by Section Recommendations

https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_articles_across_languages, Tiziano Piccardi, Michele Catasta, Diego Saez-Trumper, Robert West, Leila Zia
Intra-language Information extraction
Inter-language Information extraction
Inter-language recommendations
Problem statement

Given an article A in language L, recommend the list of sections to be added to the article considering similar articles to article A in language L.
How to define similarity?
Article-based recommendations

Stanford, California

INPUTS

1 Geography
2 Demographics
3 Politics
4 Education
5 Notable People
6 See also
7 References

Text content

OUTPUTS

Matrix Factorization for Collaborative Filtering

Topic models (LDA) trained on Wikipedia corpus

computation of probability distributions of sections over similar articles
Category-based recommendations
Evaluation

![Graph showing automatic evaluation results](image1)

**Automatic**

![Graph showing human evaluation results](image2)

**Human**
Conclusion

- Introduced the section recommendation problem
- Explored several methods using
  - features derived from the raw input article
  - Wikipedia’s category network
- Showed that the category centric count-based approach does best (precision @10 at 80%) for a Wikipedia language with a large category network such as enwiki.
- Learned that enwiki’s category network is key in offering useful recommendations, but the network needs pruning (we developed a methodology for doing that)
Discussion

- **What about non-enwiki?**
  - The methods used are language independent though they (will most likely) rely on the size of the category network and the number of articles in these categories in a language.
  - We expect the methodology to work in languages with large enough category network. We will test it next in frwiki.
  - What if the language is small?
Discussion

- **Improving section recommendations**
  - Sourcing signal from other languages (inter-language)
  - Semantically related sections
  - Providing more in-depth information
  - The order of sections
  - Including less frequent and the long tail of sections

- **Entry point**
  - Article versus category
What’s next?

● More languages
  ○ Build the tool/API for testing further
  ○ Test the models in frwiki: are they good?
  ○ What if the language is much smaller than fr?
    ■ Test in a smaller language
    ■ Develop a different set of technologies for small languages
  ○ Ma Commune

● Experiment with less-experienced editors
What’s next?

● Other types of recommendations
  ○ Image recommendation
  ○ Info-box recommendation (a simpler problem)
  ○ ....
Thanks! :)  

Documentation at  
https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_articles_across_languages