Hierarchical Supervised Topic Selection for Indian Knowledge Resource Building

Team Zero - Arghya B. , Neha M., Sanjana S., Sathvik B.

October 2019

Identifying topics/pages to be included in Indian Language Wikipedia

1 Abstract

' There is an increase in usage of Wikipedia in Indian languages. Currently, they are not very well-developed and identification of new topics to be included becomes important.

In this deliverable, our focus has been on understanding what could be considered appropriate topics for Hindi Wikipedia, extraction of possible topics from various relevant resources and using these set of topic lexicons as candidates for further investigation.

2 Introduction

Hindi Wikipedia is 16 years old (launched July 2003) and has 133,660 articles. As of 2019, the Hindi Wikipedia has 55 thousand unique categories and 55.68 % of them do not have an appropriate page in the category namespace. The average article in this language version has 5 categories, while the ratio of number of unique categories per articles is 0.42. The largest number of articles belongs to the Nature (27%) and Science (16%) category. In Hindi Wikipedia, articles related to Business and People has the highest average quality. Content about Law is read more often and has the highest user interest on average.

In this deliverable, we come up with three kinds of possible improvements that can be made to any Encyclopedia. We also mine resources based on cues from the improvement techniques. In light of the collected resources, we make a rough plan of action for selecting topics from the resources collected for output of the next deliverable.

3 Methodologies

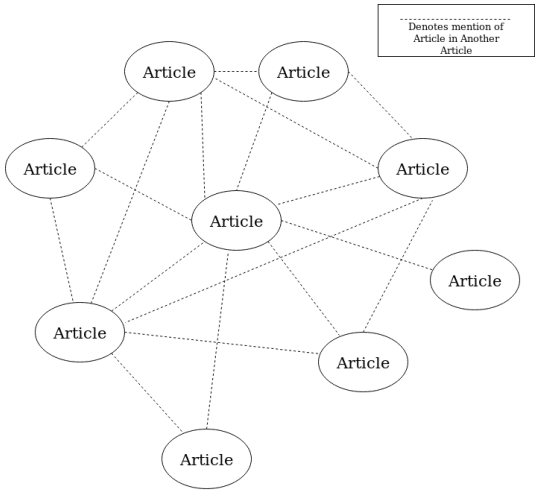
The problem of identifying topics for Indian Wikipedia has the following constraints - the topic should be relevant to the reader of an encyclopedia and there should be a set of topics which are specific to the language in which the encyclopedia exists.

We propose to make 3 kinds improvements to the existing Hindi Wikipedia:

- Improve the inter-connectivity of topics, by adding more link enriching topics.
- Expand it's horizontal breadth with respect to the topics, i.e introduce topics that can be added as a category in upper levels of the taxonomy of categorizing the topics.
- Expand it's vertical depth with respect to the topics. i.e introduce more topics that already fall into categories of upper levels of taxonomy.

3.1 Improving inter-connectivity of topics

Each of the Wikipedia articles can be viewed as a set of links to other Wikipedia articles along with it's article specific information. So since Hindi Wikipedia is currently only 1/10th the size of English Wikipedia, we hypothesize that a lot of topics can be found by filtering relevant words from the articles currently in Wikipedia itself.



Interconnectedness

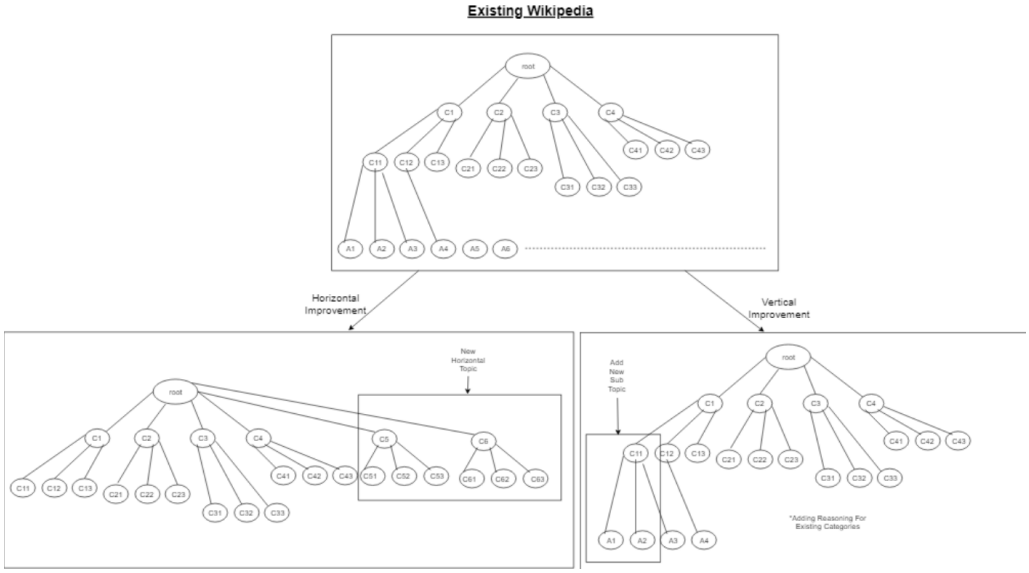
Several Wikipedia articles were parsed for their content and named entity recognition was performed on them. These named entities are topics that are related to the currently existing scraped page, and are hence potential candidates for new topics if they aren't already part of the Hindi Wikipedia. This would increase the number of references in the articles making it convenient for the readers to get information on related topics which in turn will also improve the inter-connectivity of articles.

3.2 Improving breadth of topics

One improvement to Wikipedia can be done by adding new topics of emerging interest. A primary resource for this type of expansion would be news. Popular news articles were scraped and keywords were extracted from them. Analysis of current rising trends was also done.

We have a built a scraper to scrape BBC Hindi News. The rationale for using BBC was twofold. One was that it had news separated into various categories like sports, entertainment, international, science, social and India. Other Indian Newspapers in Hindi weren't so straightforward to scrape. In BBC Hindi, starting from the home pages for each category, we scraped articles from each page, and obtained new links by following recommendations from each page.

Now that the scraper is built and tested, we will improve our data, by scraping more (Indian) news sites online.



3.3 Vertical expansion of topics

Wikipedia already has a predefined taxonomy for it's articles and we'll be using that as our basis to get more topics in Hindi Encyclopedia. There are some other resources which provide content in Hindi but they are not as organized as Wikipedia. The scraper function scrapes the data present on that page but due to different in-built structures, it fails to get data from all other hyperlinks at once and hence the node is initialized a couple of times.

4 Link to code

code and dataset is here. Code for each of the three types of expansion is present along with scope documents and dataset. The structure of the git repo is as follows

```
ire-major-project
├── data
├── docs
│   └── Scope_Document.pdf
└── src
    ├── Scrapy_Crawler
    ├── Vertical_Expansion
    └── Wiki_NER
```

- Scrapy_Crawler contains the BBC Scraper for horizontal expansion. Inside the directory is a file named output.csv containing a sample of the scraped news articles.
- Vertical_Expansion contains code for scraping articles to expand domains.The data/ folder contains this dataset.
- Wiki_NER contains the code for scraping Wikipedia pages and performing named entity recognition.

5 Findings and Analysis

Expansion of related Wikipedia articles was majorly based on how well named entity recognition was performed on the article. While it gave good results for a lot of cases, there were some shortcomings, for example, fictional characters being discussed in a movie plot that don't really require a Wikipedia page of their own but would still be recognised as a named entity.

For horizontal expansion of Hindi Wikipedia, we would need new categories or topics. Scraping different types of regional newspapers could help identify new topic. One example would be important events, which can be a topic in itself.

Vertical expansion of Wikipedia articles helped in finding information relevant to social aspects and society topic of Wikipedia.

6 Difference from scope document

In the scope document we had hypothesized to use a classification based approach, but more research into the problem exposed a fundamental flaw. Curating a list of good and bad topics need not necessarily be a good idea, because being a broad encyclopedia, adding topics that aren't particularly relevant to the Indian context is not an issue. Hence, given this constraint it becomes evident that any machine learning based Classifier wouldn't perform well.

Hence, since the aim of this project is to output possible article names that can be added to Hindi Wikipedia, we had to resort to a different strategy.

6.1 New Timeline

Our updated strategy consists of 2 ways to output topics:

- Finding action words from Hindi Wikipedia and performing Topic Modelling on them to identify representative ones.
- Perform Unsupervised clustering to get topics from articles of news and other domain data we collect.

In the next deliverable we will be presenting our analysis of the data we are scraping and the topics we have retrieved.