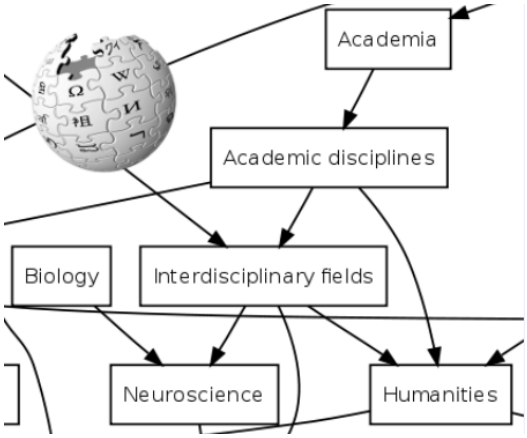

CS 522
ADVANCED DATA MINING

FINAL REPORT
WIKIPEDIA HIERARCHY EXTRACTION

KRISHNA BHARADWAJ(A20398222)
POOJA HEMANTKUMAR PATEL
ANIMESH PATNI

Contents

1 Abstract

Wikipedia as we all know is a huge and well renowned source of information available freely online. As per recent estimates, the English Wikipedia alone has over **5,525,674** articles. These articles are categorized in a dense hierarchy with a few top level hierarchies and bubbles down to the leaf categories. While the structure of the hierarchy is well defined, it is very much subjective. The Subjective nature of the hierarchy defined in Wikipedia is mainly due to the fact that the structure is defined by users who submit the articles. This could mean there's a possibility that two people could define two different hierarchies for the same page. In this project, we aim to take a more data-centric approach towards categorization of wiki articles. We attempt to make use of the dataless approach that has been described in Yangqiu Song et al in their paper 'On Dataless Hierarchical Text Classification'.

1.1 Introduction

Wikipedia hierarchy Extraction is extracting the hierarchy of the article which is input by the user. This is very helpful for the categorization of the articles and also can be used in information browsing well as it will decrease the time required to browse the required information. So in order to achieve such goals it is very necessary to understand the existing hierarchy of the Wikipedia and also there is requirement to fetch the Wikipedia hierarchy. In our project, we will take input from the user as an article title and then extract the hierarchy of the same.

Topics we used for hierarchy extraction:

We chose the major topics which are very common in day to day life:

- Finance
- Biology
- Politics

Structure for 'Wikipedia Hierarchy Extraction':

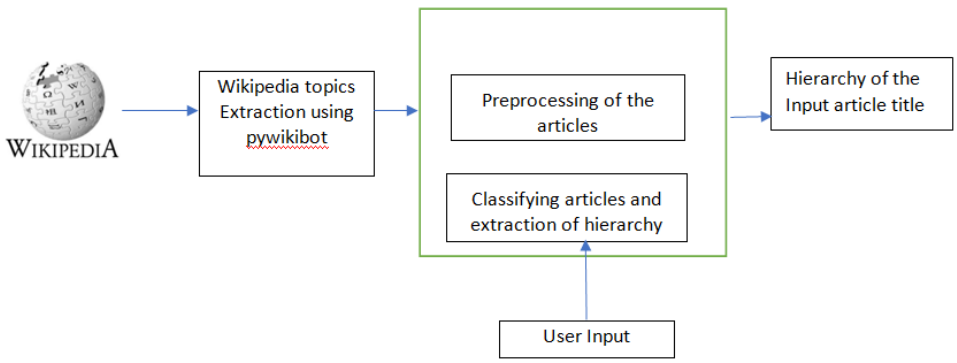

Figure 1 Architecture

The above diagram shows the major steps that are performed in order to extract the hierarchy of the given input from the user. The vital steps include:

- The fetching the existing Wikipedia hierarchies of the top level categories like finance, biology, politics etc.
- Then extracting all the articles related to the above topics thereafter preprocessing all the articles.
- Fetch the hierarchy by getting the most similar page at each level.

2 Dataset

2.1 Data Procurement

For the purpose of this project, Wikipedia dumps were our primary source of data. The process to procure the data is as follows –

1. Read the latest category xml dump file that is available in https://dumps.wikimedia.org/enwiki/latest/
2. The above file only contains information of categories and the hierarchy as defined in Wikipedia
3. We extracted the category tree for the categories that we used to perform our experiment. These categories include 'Finance', 'Politics' and 'Biology'. For our experiments, we consider these categories as the top level categories and hence, they are marked at level 0.
4. The tree was extracted by using pywikibot which is an API written in Python.
5. We restricted the data only to two levels. This was done because as we increased the depth, the categories that were being fetched got irrelevant to the high level topics.
6. Once we procured the trees for each of the top level categories, we then crawl the Wikipedia site for the pages/articles corresponding to the articles in the trees. This becomes our dataset for further experimentation and analysis.

2.2 Data Prepressing

1. The following steps were taken to preprocess the data that we obtained above –
2. Stop Word removal
3. Data Stemming
4. Removal of list articles from the corpus. An example of a list article is 'Lists of organisms by population.

2.3 Feature Engineering

There was not much to be done as part of feature engineering. However, we did add one important column to the existing dataframe called 'Level' which pointed to the current level of the article in the hierarchy as a whole.

	ID	Title	Text	Level	new_column
0	1	biology	biology is the natural science that involves t...	0	biology natural science involves study life li...
1	2	quantum biology	quantum biology refers to applications of quan...	1	quantum biology refers applications quantum me...
3	4	morphology (biology)	morphology is a branch of biology dealing with...	2	morphology is a branch of biology dealing with...
4	6	systems biology	systems biology is the computational and mathe...	2	morphology branch biology dealing study form s...
5	8	paleobiology	paleobiology (uk & canadian english: palaeobio...	2	systems biology computational mathematical mod...
6	10	cell biology	cell biology or cytology, (from the greek κυτο...	2	paleobiology uk canadian english palaeobiology...
7	12	medicine	medicine is the science and practice of the di...	2	cell biology cytology greek kytos vessel branc...
8	14	nutrition	nutrition is the science that interprets the i...	2	medicine science practice diagnosis treatment ...
9	16	astrobiology	astrobiology is the study of the origin, evolu...	2	nutrition science interprets interaction nutri...
10	18	chemical biology	chemical biology is a scientific discipline sp...	2	astrobiology study origin evolution distributi...
11	19	branches of botany	botany is a natural science concerned with the...	2	chemical biology scientific discipline spannin...
12	21	bionics	bionics is the application of biological metho...	2	botany natural science concerned study plants ...
13	22	evolutionary biology	evolutionary biology is the subfield of biolog...	2	bionics application biological methods systems...
14	24	ecology	ecology (from greek: οἶκος, "house", or "envir...	2	evolutionary biology subfield biology studies ...
15	26	mycology	mycology is the branch of biology concerned wi...	2	ecology greek house environment study scientif...
16	28	chronobiology	chronobiology is a field of biology that exami...	2	mycology branch biology concerned study fungi ...
17	30	soil biology	soil biology is the study of microbial and fau...	2	chronobiology field biology examines periodic ...
18	32	physiology	physiology (; from ancient greek φύσις (physis...	2	soil biology study microbial faunal activity e...
19	34	structural biology	structural biology is a branch of molecular bi...	2	physiology ancient greek physis meaning nature...

Figure 2: Corpus after preprocessing

We generated some word clouds in order to visualize the corpus better

2.4 Programming Languages and Packages

Programming Language: Python
Packages: pywikibot, BeautifulSoup, nltk, gensim, pandas, sklearn

3 Experiments

3.1 TF-IDF

TF-IDF(Term Frequency- Inverse Document Frequency) is numerical value that indicates how important the Wikipedia article is in the corpus of wiki articles. The tf-idf value increases proportionally according to the number of times a word appears in the document and is adjusted by word frequency.

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t)$$

Thus tf-idf is Term Frequency times Inverse document frequency. The inverse document frequency is given by:

$$\text{idf}(t) = log\frac{n_d}{1+\text{df}(d,t)}.$$

Where nd is total number of documents, and df(d,t) is the number of documents that contain terms t.

3.2 Bag of Words Representation

We convert the raw data from the corpus into a bag of words representation. This involves vectorizing the text into numerical features. This can be implemented by one of the following strategies –

- **Tokenizing:** Giving an integer id for each possible token, by tokenizing string using white spaces
- **Counting :** Counting the token occurences in each document
- **Normalizing:** weighing the importance of tokens with respect to articles

For our approach, we decided to use gensim.corpora.dictionary.Dictionary.doc2bow function of python gensim class in order to get our bag of words representation

We then convert tokenized documents to vectors.

In essence, vec(Finance) = Vec(all articles that fall under finance category)

3.3 Calculate Similarity

In order to calculate similarity of the documents, we followed two approaches –

1. Bottom up Approach
2. Top Down Approach

3.3.1 Bottom Up Approach

The bottom up approach includes the following tasks –

1. Fetch an article to compare
2. Transform the article into its TF-IDF representation
3. Compare this representation with the summary vectors that were created for the top level categories. This will give us the most similar top level category
4. Once we get the top level hierarchy, fetch the most similar document by comparing the new article with the corpus containing the articles related to the top level hierarchy
5. At this point, we assume that the article obtained above is the parent article
6. Make an intermediate corpus by fetching only those articles that are above the parent article found above

7. Find the article that is most similar to the parent article obtained in step 5
8. We then repeat 5 to 7 until we reach the top level hierarchy

3.3.2 Top Down Approach

The top down approach includes the following tasks –

1. Fetch an article to compare
2. Transform the article into its TF-IDF representation
3. Compare this representation with the summary vectors that were created for the top level categories. This will give us the most similar top level category
4. Fetch the immediate children of the top level category and find the most similar article among these.
5. Repeat step 4 until we reach the leaf or eventually the article in the tree

3.3.3 Word2Vec

1. We also tried the Word2Vec approach to find the similarity amongst the articles. We converted the corpus as well as the new article into a word2vec representation.
2. Computed the Word2Vec using a predefined function Word2Vec under packages "gensim.models".
3. Then after that we converted both the word2vec representation into np vectors.
4. Computed the cosine_similarity, using a predefined function in Python under package: "sklearn.metrics.pairwise".
5. cosine_similarity(vector1, vector2)
6. We were not able to go ahead with this approach as the outcomes were not satisfactory.

Upon application of both the methods, we found the top down approach to be more efficient since we don't have to compare with all the articles from the bottom.

4 Results and Analysis

For our result analysis, we picked 20 articles related to one of the three selected categories from Wikipedia and extracted the hierarchy using **the top down** and **bottom up** approaches.

The results were then compared with the actual hierarchy that was manually extracted for these 20 articles.

The following were the results obtained –

Titles	Actual	Top Down	Bottom Up
Mutual fund	Mutual Fund , Investment funds , financial services , Finance	finance,financial services,investment fund	finance,financial services,investment fund
Hedge fund	Hedge Funds , Investment funds , financial services , Finance	finance,financial risk,venezuela	finance,financial services,investment fund
Bank	Banking , Finance	finance,financial services,bank	finance,financial services,bank
Debt collection	debt collection , Finance	finance,debt,debt collection	Debt collection,immortality,hybrid (biology),biology
Loan	loans , Banking , Finance	finance,debt,loan	Loan,immortality,hybrid (biology),biology
Debt bondage	debt bondage , debt , finance	finance,debt,debt bondage	Debt bondage,taxonomy (biology),philosophy of biology,biology
Corruption	corruption , financial problems , finance	politics,political corruption,corruption	Corruption,biology,natural environment,biology
Deposit account	bank deposits , investment , finance	finance,financial services,deposit account	Deposit account,immortality,hybrid (biology),biology
Quantum Aspects of Life	Quantum Aspects of Life , Quantum biology , biology	biology,mathematical and theoretical biology,history of biology	Quantum Aspects of Life,history of biology,mathematical and theoretical biology,biology
Orchestrated objective reduction	Orchestrated objective reduction , Quantum biology , biology	biology,mathematical and theoretical biology,immortality	Orchestrated objective reduction,immortality,hybrid (biology),biology
Avicide	avicides , biocides , biology	finance,aircraft in fiction,venezuela	Avicide,species,eukaryote,biology
Geographical feature	artificial ecosystems , ecology , natural environment , biology	biology,mathematical and theoretical biology,biologist	Geographical feature,biologist,philosophy of biology,biology
Election	elections , voting, politics	politics,voting,election	Election,biology,natural environment,biology
Political violence	political violence , politics	politics,political violence	Political violence,biology,natural environment,biology
United States presidential debates, 2016	political debates , political events , politics	politics,voting,united states presidential debates, 2016	United States presidential debates, 2016,history of biology,mathematical and theoretical biology,biology
Bankruptcy	Bankruptcy, Corporate finance, Finance	finance,debt,debt collection	Bankruptcy,immortality,hybrid (biology),biology
Bionics	Bionics, Branches of biology, Biology	biology,mathematical and theoretical biology,biology	Bionics,biology,natural environment,biology
Algae bioreactor	Algae bioreactor, Biotechnology, Biology	biology,mathematical and theoretical biology,biology	Algae bioreactor,biology,natural environment,biology
Biomolecule	Biomolecules, Structural Biology, Biology	biology,mathematical and theoretical biology,taxonomy (biology)	Biomolecule,taxonomy (biology),philosophy of biology,biology

As we can see above, the top down approach seems to do better compared to the bottom up approach. However, while making comparisons, it has to be kept in mind that the actual hierarchy that is available in Wikipedia is subjective. That is, it is the view of the user submitting the article. As a result, it cannot be a hard and fast rule that the actual hierarchy that exists on Wikipedia is the only hierarchy that exists. We can also infer this from the fact that an article can have multiple parent articles and conversely, a parent article can have multiple children. Hence, for a given hierarchy, there is always a possibility of having multiple hierarchies for a given article. This is illustrated by taking the example of Election which we have taken above.

An alternate hierarchy could be constructed as follows –

Election →Political Events → Politics

This hierarchy also makes sense when looked at from a human perspective.

5 Conclusion

As summarized above, the top down approach has yielded us acceptable results when it comes to extracting of Wikipedia hierarchies given a particular topic. However, due to the subjective nature of the maintenance of the hierarchies, we can posit that there could be a good chance of having multiple hierarchies for a particular article.

That being said, there are more ways of tuning the above algorithms and discover more suitable techniques to effectively extract any hierarchy given a Wikipedia topic. We would like to explore further into these techniques such as word2vecs and LSA.