Improving Wikipedia Semantic Data by adding Category-Based Data

Bhavya Agarwal[1] and Mrinal Priyadarshi[2]

Abstract— **This project attempts at improving the semantic content of wikipedia by suggesting additional candidate categories which could be added to the page in order to improve their classification. This can also be used in improving the semantic search by introducing additional ontologies for the pages. We define inter-category and intra-category links for all the categories and try to evaluate the 'tightness' measure for the determining the quality of the category, and distance between categories in order to determine the inter-category links.**

I. INTRODUCTION

Wikipedia is the world's largest free encyclopedia, in which articles are collaboratively written and maintained by large number of on line volunteers(77 thousand). One of the most interesting aspect of Wikipedia articles is categorization and linkage in between the articles of same category and linkage between articles of different category.

Here we assume that more tightly knit category carry more semantic information about the page than categories that are widespread, hence possibly covering various different domains. This can help a lot in differentiating between various categories in a page by introducing an order of relevance and using those results, we attempt to introduce more candidate categories for the page which can be added to the page by users/automatically.

We can determine the tightness of a category by computing the distance between every pair of nodes. Since it is very inefficient to compute the hopping distance between every pair of nodes in a category, we select the page with highest Pagerank as our center of the category and compute the distance relative to the center. The measured distances are then used as input to view and analyze the connectivity in the category and compare it across various categories to determine the quality of content.[9]

For the inter-category links, we use the measures given by Volkel et al[8] which gives a quantitative measure of the relevance between 2 categories. Those distances are calculated and the results are then used to suggest some additional categories for the page. The technique presented in this paper can be used in the framework of Semantic data also in order to improve the ontology information for a page.

Computer Science Department,Stony Brook University

[1]Bhavya Agarwal is Graduate Student of Computer Science, State University of New York, Stony Brook bhavya.agarwal@cs.stonybrook.edu

[2]Mrinal Priayadarshi is Graduate Student of Computer Science, State University of New York, Stony Brook mrinal.priyadarshi@cs.stonybrook.edu

II. PRIOR WORK

Calculating distance between pages has been part of many major research areas based on Wikipedia data. The basis of our motivation is that quality of category is directly proportional to its connectivity within its pages. The existing algorithms used to determine the quality and cooperation of data in wikipedia articles use these techniques:

1) Length of Article [4]: The more the length of the article, the article would be more relevant as content determines the quality of article.
2) Number of edits in article [3]: If the article has been edited number of times, then it's more relevant and up -to-date.
3) Number of outgoing and incoming links [6]: The importance of an article is determined by connectedness of the articles with other articles in Wikipedia. If the article is cited or referred by lots of other articles, this enhances the relevancy of the article.
4) Number of contributors to the article [3]: Number of contributors are directly proportional to the number and frequency of updates on the article.

Secondly, the inter-category distance is the index of the closeness between two categories, which in turn can be used enhance the search mechanism of Wikipedia.

For developing the semantic relationship among the wikipedia articles, ontological graph was created from

the Wikipedia articles [7]. These ontological graphs specify the semantic relationship of the form 'is instance of' in between the articles. Like the paper by Volkel et al. , We will not determine the exact semantic relationship but only if the two categories are semantically related or not.

III. Algorithm

A. Data Processing

Wikipedia provides its data dump for all the languages: We used the Wikipedia data only for English language (http://dumps.wikimedia.org/enwiki/). The data used was for 2008 in order to fit it in the computer's main memory and do computations in the real time. However, the algorithm is generic and can be used on any semantically-connected dataset. The following databases were used -

1) enwiki-20080103-page.sql: this is the data dump of all the Wikipedia pages, their ids and other properties of the page. (Size : 1 GB)
2) enwiki-20080103-pagelinks.sql: this is the data dump of the links among the articles/pages. (Size : 6.6 GB)
3) enwiki-20080103-categorylinks.sql: this is data dump of category association of articles pages in Wikipedia. (Size: 1.5 GB)

These sql files expand even further when dumped in the sql database. 2008 dump of wikipedia had around 4.6 million pages and 320,000 categories. Also around 213 million links between pages were present.

IV. Algorithm

The figure 1 below shows the flowchart of the presented algorithm.

Data cleansing was the first step taken by us in order to filter out data with little/no relevant information. Wikipedia has a special namespace 0 defined for all its articles. However, wikipedia contains many other pages like talk-pages, user-account-pages, some discussion-pages etc. which were not very relevant to our analysis. We cleaned the tables in order to make make the computation faster.

After that we moved to Pageranking the whole input data in order to determine the most relevant pages in the category. The basic implementation of Page Rank was:

1) Creating an adjacency list with outlinks with each row having the outlinks for that page. This is taken from the database dumped in a txt file.
2) Create an inlinks-based table in order to compute PageRank
3) Run the pagerank algorithm and store the results.There are some configurable parameters in the algorithm. The value of variables we used were:Epsilon = 0.0001 which defines the point where iteration needs to stop and Prob = 0.2 ,i.e. the probability that the surfer will jump to a random page.

After that we find the Inter-Category distance. The basic idea behind the inter-category distance is to find out similar categories. Similar categories can be used to more relevant categories for wikipedia pages thus suggest categories for future dataset. The following was used for the input-

1) Pages : Page Id and Page Title
2) Pagelinks: Page Ids(Source) and Page Title(Destination)
3) Category: Category Id and Category title
4) Categorylinks: Category Id and Page Title

Category tightness determines how closely knitted are pages in a certain category. If the category is having more tightness, then the quality of the category is assumed to better.

Radius of category = Sum of distances of all the articles in the category from the highest rank page/ Total number of articles

Since an $O(n^2)$ would had been impossible to calculate, we use the concept of Erdos number, i.e. we find the distance of every node in the category with the most significant node in that category, i.e. the one with highest PageRank. Connectivity of the category is inversely proportional to the radius of the category.

But the closeness of the category is also dependent on the size of the category. To make more closer observation,we have segregated the categories in 5 groups :

1) Categories with less than 5 pages
2) Categories with pages between 10 to 50
3) Categories with pages between 50 to 100
4) Categories with pages between 200 to 500
5) Categories with more than 500 pages

Below we show the most tightly knit category for each of these classes. We found out that the results are nicely separated for all the categories which have a high number of pages. We move to the next step of finding inter-category links. We use the following 2 different measures to find the inter-category links between 2 categories A and B-

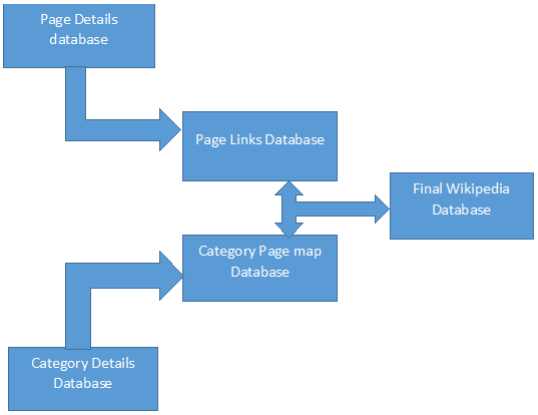

Fig. 1. Various Databases and their combination

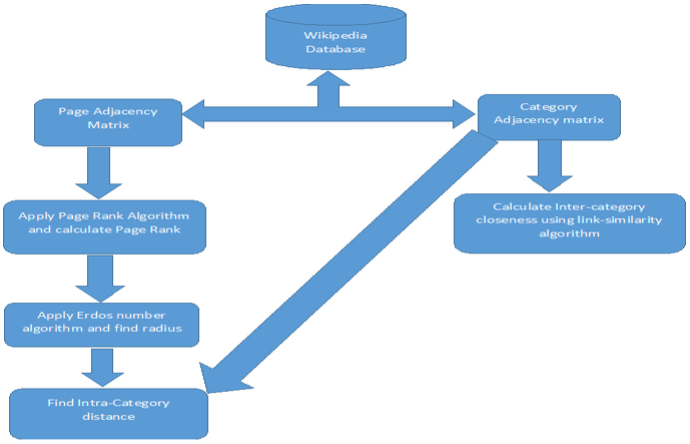

Fig. 2. Flow of the algorithm

1) For the first measure, we compute the intersection of 2 categories and find the measure by dividing the number of pages in intersection set by number of pages in the category A, i.e. similarity = $|(A \cap B)| \,/\, |A|$.
2) Here we make a set by computing the outlinks of all the pages in category A and for a set C. The similarity is calculated by $|(A \cap B)| \,/\, |A|$.

The steps for calculating the inter-category links were -

1) Mapping of category ID and Page ID: We created a map of category ID and page ID by joining Page table(Page Id to Page title Map) and CategoryLinks table(Category Id to Page Title Map).
2) Category Adjacency list: On the similar lines of Page Adjacency List, we calculated the Category Adjacency List. Category A will be part of adjacency list of Category B, if any of the pages in Category A is having the outlink in any page from Category B.
3) Category Set calculation: For obtaining the inter-category distance, we used the categories which are having high tightness as this determines quality of category. The idea behind using this set is to find the similarity on the best quality categories of Wikipedia and find a correlation between the inter and intra category distance.
4) Similarity index calculation on Category Set: We used this category set for calculation of similarity index.

Since, this is again an $O(n^2)$ and the number of categories is huge, we ran this algorithm only on categories with high tightness and evaluated and present the results in the next section. Here we try to find if these final categories obtained can be suggested as candidate categories for pages in which they feature as

Wikipedia Page Id	Page Rank	Page Topic
5302153	0.002138361	United_States
84707	0.001350467	2007
88822	0.001302155	2008
1921890	0.001287489	Geographic_coordinate_system
5300058	0.000958918	United_Kingdom
81615	0.000823325	2006
1804986	0.000705178	France
5535280	0.000665401	Wikimedia_Commons
5308545	0.000663389	United_States_postal_abbreviations
687324	0.000651346	Biography

TABLE I
PAGERANK DATA

Fig. 3. PageRank

Category	Number of Pages	Page Tightness
14th_Lok_Sabha_Members	540	1.49074
1st_millennium	1011	1.80119
1984_video_games	199	1.80402
1960s_music_groups	696	1.91523
13th_century_Roman_Catholic_bishops	166	1.92771
1970s_automobiles	360	2
1960s_automobiles	282	2.00709
1913_films	189	2.01587
1982_video_games	158	2.01899
1942_films	203	2.02463

TABLE II
MOST TIGHT CATEGORIES

Fig. 4. Most Tight Categories

one of the most tight category.

V. RESULTS

Since the data of 2013 was too big to be sequentially dumped and processed on the machine, we processed the 2008 data and found the following PageRank results. We present the results we obtained in each of the steps.

Fig.3 shows the PageRank data obtained from the adjacency matrix. We then analysed the tightness of first 23,000 categories (due to time and space constraints). Fig.4 shows the most tight categories obtained. After that we go on to analyse the most loose categories analysed as shown in Fig.5. Then we show the most tight categories obtained for each class mentioned above in figure 6 and finally figure 7 shows the prediction for candidate categories for some of the most tight categories.

As it can be seen, the algorithm fairly predicts some of the categories that can be predicted as the candidate categories. We have the api which can give us the most similar categories for each category. Here we only present some selected results. We made the table of results for many top results and after seeing the results found that the related categories are very similar. The results couldn't be quantitatively evaluated due to the recent wikipedia dumps being too huge in size to be analysed in realtime. However, after observing the results and having a short survey, we can say that more than 50 percent of categories could be added while some were ambiguous. We suggest coupled with a user-based survey on wikipedia, we can make many additions to the wikipedia semantic data.

VI. CONCLUSIONS

This algorithm is a novel way to improve the quality of wikipedia content in terms of ontological relationship of pages and categories by suggesting the candidate categories for pages.Firstly using the intra category distance could be used to find the categories which are not of good quality and which are of bad

1781_births	226	3.81416
1940s_drama_film_stubs	254	3.83858
1990s_hip_hop_album_stubs	243	3.88889
2000s_album_stubs	1595	3.92915
1990s_album_stubs	862	3.9652
2000s_heavy_metal_album_stubs	757	3.97622
1980s_album_stubs	328	4.00915
1981_albums	600	4.02333
1990s_pop_album_stubs	264	4.05303

TABLE III

MOST LOOSE CATEGORIES

Fig. 5. Most loose categories

Category with highest tightness	Page Range	Number of pages	Tightness
1946_births	>1500	3692	2.10428
14th_Lok_Sabha_Members	500-1500	540	2
1970s_automobiles	200-500	360	1.49074
1984_Summer_Olympics_stubs	100-200	115	0.991304
1936_Summer_Olympics_events	0-100	21	0.952381

TABLE IV

MOST TIGHT CATEGORIES FOR EACH CLASS

Fig. 6. Most Tight Categories for each class

Original Page	Similar Page 1	Similar Page 2
14th_Lok_Sabha_Members	Living People (0.957407)	Indian_politician_stubs(0.433334)
1984_video_games	Commodore_64_games(0.957407)	ZX_Spectrum_games(0.433334)
1st_millennium	1st_century Centuries (0.957407)	All_articles_with_unsourced_statement(0.957407)
13th_century_Roman_Catholic_bishops	12th_century_births(0.162651)	13th_century_births (0.162651)
1960s_music_groups	N_Sync_members(0.05747)	Til_Tuesday_members (0.01437)

TABLE V

PREDICTIONS FOR THE GIVEN CATEGORIES

Fig. 7. Prediction for few categories

quality.

The candidate categories for the page can be found by analysing the inter-category links and finding the similar categories. The semantic relationship of pages in the categories which are less tight can be improved by suggesting/adding similar categories in order of their similarity index with category of target page.

We believe that our algorithm can be used in various applications for semantic web and it will an step closer to semantically richer Wikipedia.

VII. ENHANCEMENTS AND FUTURE WORK

Since we were limited by resources, we could enhance the algorithm used for finding the tightness of the category.Rather than choosing the highest page rank to determine the principal component of the category, we could have used the top 'n' pages and find the distance of all the pages from top few pages.

For inter-category distance,instead of connecting two categories only if one of the page is linked to pages of other category, we could have used the length of

the links from pages of one category to pages of other category as an input in calculation of similarity index.

For better evaluation of the results, we could have used automated the evaluation techniques so that we can how much actual candidate categories are different from our results. However, the next complete dataset properly available was from 2012. After 2008, the wikipedia data saw its biggest rise, and hence it became nearly impossible to fit these databases in main memory and analyse it. The huge time spent in cleansing and preparing the data for this algorithm, could have been used in improving it further to give better results.

ACKNOWLEDGMENT

We want to thank Professor Leman Akoglu for her motivation and guidance throughout the course. Also, we would like to thank the Wikimedia Foundation for keeping the Data freely available on the web.

VIII. REFERENCES

1) Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).
2) Grossman, Jerrold W., and Patrick DF Ion. "On a portion of the well-known collaboration graph." Congressus Numerantium (1995): 129-132.
3) Hu, Meiqun, et al. "Measuring article quality in Wikipedia: models and evaluation." 2007
4) Blumenstock, Joshua E. "Size matters: word count as a measure of quality on wikipedia.2008
5) Chernov, Sergey, et al. "Extracting Semantics Relationships between Wikipedia Categories." SemWiki 206 (2006).
6) A. Lih. Wikipedia as participatory journalism: Reliable sources metrics for evaluating collaborative media as a news resource. In Proc. of the 5th International Symposium on Online Journalism, April 2004.
7) Natalia Kozlova. Automatic Ontology Extraction for Document Classification. Master?s thesis, Saarland University, Germany, February 2005
8) Max Volkel, Markus Krotzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic Wikipedia. In Proceedings of the 15th international conference on World Wide Web,Edinburgh, Scotland, 2006.
9) Bhavya Agarwal, Mrinal Priayadarshi , Improving Category-Based Semantics in Wikipedia, CSE590 Project Milestone Report, Fall 2013, Stony Brook University