International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 4, April 2015, PP 93-97
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Indian Language Wikipedias: A Comparison Study

Vasudevan T V

Asst Professor, Department of Computer Applications,MES College of Engineering, Kuttippuram, Kerala, India

ABSTRACT

Wikipedia is a popular, free, publicly editable internet encyclopedia supported by the non-profit Wikimedia Foundation. This paper presents an overview of research in the Indian Language Wikipedias. Different research areas related with Wikipedia are examined first. This is followed by a comparison study of major Indian Language Wikipedias which analyses the fundamental components of Wikipedia such as articles, authors and edits.

Keywords: Wikipedia, Indian Language, Quantitative Analysis, Articles, Authors, Edits

INTRODUCTION

Wikipedia is a free online multilingual encyclopedia that can be edited by anyone. Wikipedia is supported by the non-profit Wikimedia Foundation. It was launched on January 15, 2001 [1]. Presently it contains 35 million articles in 288 languages. The English Edition of Wikipedia itself contains over 4.8 million articles as compared to more than 120,000 articles in the next largest English language encyclopedia, *Encyclopedia Britannica Online* [2]. Wikipedia is interesting to research because of the vastness and open nature of its data. We can analyse various topics such as fundamental components, structure and growth of information, author collaboration etc.

HISTORY OF WIKIPEDIA IN INDIAN LANGUAGES

Assamese Wikipedia, the first Indian Language Wikipedia was started in 2nd June, 2002. However, Tamil Wikipedia was the first one to reach the milestone of 100 articles. It crossed a century of articles in January 2004. Marathi Wikipedia reached 1000 articles in May 2005. Telugu Wikipedia reached 10000 articles in September 2006. By 2006, Wikipedias were started in all major Indian languages. Hindi Wikipedia reached 1,00,000 articles in December 2013. The launch of mobile versions further increased the growth of Indian Language Wikipedias. Hindi Wikipedia, the largest Indian Language Wikipedia presently contains 1,06,045 articles (in February 2015) [3].

WIKIPEDIA RESEARCH AND INDIAN LANGUAGE WIKIPEDIAS

Wikipedia Databases register every single edit performed by Wikipedia authors in any language version. This provides a wonderful opportunity for researchers in Computer Science, Education, Sociology and Linguistics to analyse the project from different perspectives.

The following are the different research areas related with Wikipedia.

Quantitative Analysis: [4] In this popular research area associated with Wikipedia, we analyse the behaviour of the Wikipedia System using quantitative data. Statistical and Data Mining techniques are used for analysing the system. WikiXRay is a python software tool that automates quantitative analysis of all wikipedia language editions [6].

Given below are some research questions related to this area:

- What is the total number of articles / authors / words in Wikipedia?
- What is the total / average size of content in Wikipedia?
- Find out the number of contributions received in a month / year?

Address for correspondence:

vasudevantv@yahoo.co.in

- Compare these quantitative parameters for different language editions of Wikipedia.
- Identify the contribution patterns of Wikipedia authors?
- What are the different types of vandalisms affecting Wikipedia?
- Forecast the future trends of Evolution of Wikipedia.

Analysing Quality of Content: [5] In this area we try to assess the quality of content in Wikipedia. In an open access system such as Wikipedia this is very important. Further, the problem becomes harder due to the huge size and dynamic nature of the Wikipedia project. Researchers are also trying to develop automated systems that measure the quality of content in Wikipedia.

Some research questions related to this area are:

- What is the average quality of articles in Wikipedia?
- Is the average quality improving or deteriorating? Does a typical article improve over its lifetime? How fast? What trends do we see?
- How can we improve the article assessment system?
- What percentage of articles cites no references?
- What policies can we act out to prevent article deterioration?

Social Networking Analysis: [5] Social networking researchers are attracted towards large collaborative network systems such as Wikipedia. They are interested in finding out community behavior patterns of different language editions, content popularity, relationships between popularity of content and total number of contributions.

Few research questions related to this area are:

- What are the most popular articles?
- Which pages are visited together? How close are they related in content?
- How many hits per day does the Wikipedia site receive?
- How many visitors come from Google? Which pages have high Google Page Rank?
- Identify important topics of interest in Wikipedia.

But, Wikipedia research is mainly focussed on English Wikipedia and other active Wikipedias such as German, French and Dutch. Little research has been done on Indian language Wikipedias. The Access to Knowledge team of The Centre for Internet and Society (CIS-A2K) carried out a quantitative analysis to identify trends and growth patterns in Indian Language Wikipedias over the time period from September 2012 to April 2013 [7]. This paper discusses about Researching Indian language Wikipedias.

QUANTITATIVE ANALYSIS OF INDIAN LANGUAGE WIKIPEDIAS

In this section we discuss about the three basic elements of Wikipedia viz. Articles, Authors (Editors) and Edits followed by Quantitative Analysis of these elements in Indian Language Wikipedias. We analyse the trends in Article Count, New Articles per Day, Edits per Month, Active Editors and New Editors during the period July - December 2014. For analysis, we consider Top 10 Indian Language Wikipedias based on Article count - They are Hindi, Tamil, Telugu, Urdu, Marathi, Malayalam, Bengali, Gujarati, Bishnupriya Manipuri and Kannada Wikipedias [3].

Articles

Every article in Wikipedia is referenced via a unique name. One can access it with a URL like http://ml.wikipedia.org/wiki/name-of-article, where the subdomain 'ml' corresponds to Malayalam Language Edition of Wikipedia. We can redirect Synonyms to another article. Articles can be easily edited without any knowledge of HTML using a special Wiki syntax.

Authors (Editors)

Since everyone can contribute, Wikipedia articles can have a large number of authors. An author who writes a Wikipedia article can be a registered user, an anonymous user or a bot.A bot is an automated or a semi-automated program that perform edits in Wikipedia to carry out repetitive and mundane tasks. We have also administrators or sysops who can block user accounts or IP addresses from editing, protect pages from editing etc. to prevent vandalism.

Edits

When an editor changes an article, his edit is recorded and gets listed in the article's version history where differences between selected versions are highlighted.Users can add articles to their watch list to track changes. Users can also observe new contributors to their articles.There is an interesting phenomenon called 'edit war' where two authors revert each other's edits.

Analysis of Article Count

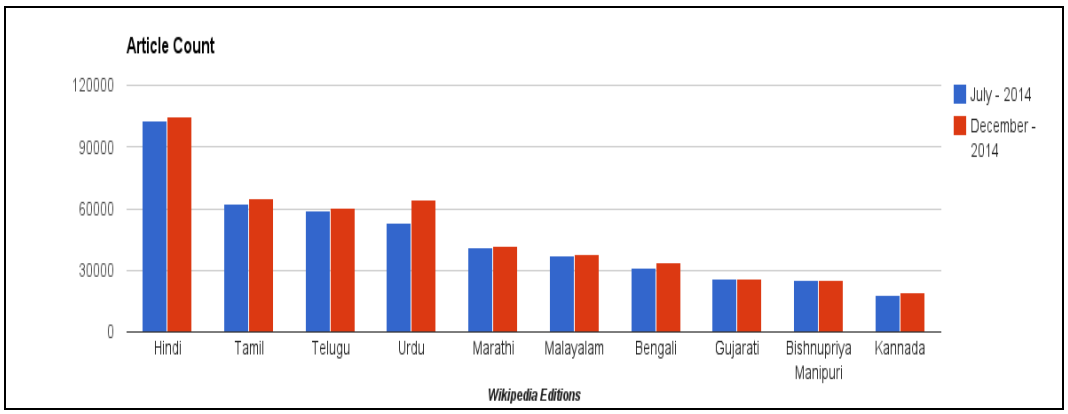

Figure1. *Growth of article count in Indian Language Wikipedia editions from July 2014 to December 2014*

1. Generally, Indian Language Wikipedias show a healthy trend in growth during the period from July 2014 to December 2014.
2. All Wikipedias except Gujarati and Bishnupriya Manipuri Wikipedias increased their article count during this period. The decrease in article count of Gujarati Wikipedia is a cause of concern.
3. Urdu, Bengali, Kannada and Tamil Wikipedias have a growth rate of 21%, 8%, 8% and 5% respectively. These were the top four wikipedias regarding growth rate during this period.
4. In terms of absolute number of articles, Urdu, Tamil, Bengali and Hindi Wikipedias have grown by about 11000, 3000, 2531 and 2000 articles respectively.
5. Given the small size of Bishnupriya Manipuri Wikipedia community, the article count achieved by them is commendable. However, we need to strengthen the community to ensure that this momentum continues.

Analysis of New Articles per Day

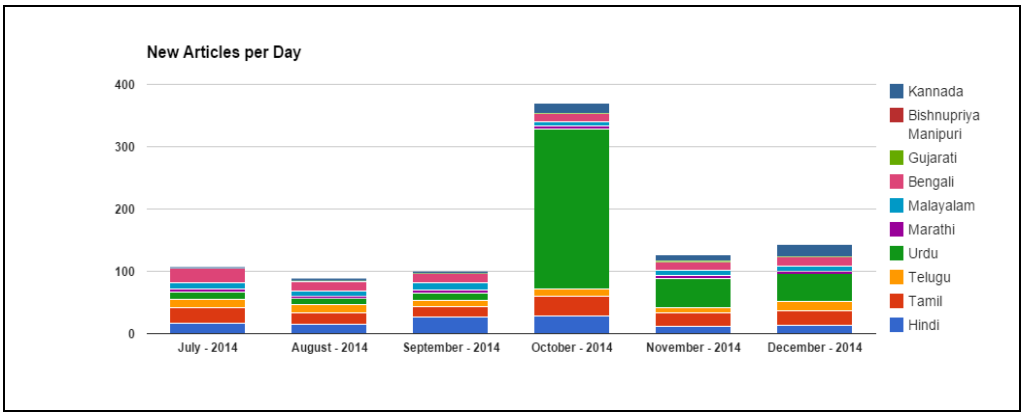

Figure2. *New Articles per Day in Indian Language Wikipedia editions from July 2014 to December 2014*

1. On an average 16 new articles have been created every day in Indian Language Wikipedias during this period.
2. Hindi, Tamil, Urdu and Bengali Wikipedias have consistently contributed more than 10 articles every day.
3. In Urdu Wikipedia there was hectic activity during October where 258 new articles joined on an average every day.
4. At least 100 new articles has been created every day altogether in Indian Language Wikipedias except the month of August.

5. Bishnupriya Manipuri Wikipedia has not created a single new article during this period which is a cause of concern.

Analysis of Edits per Month

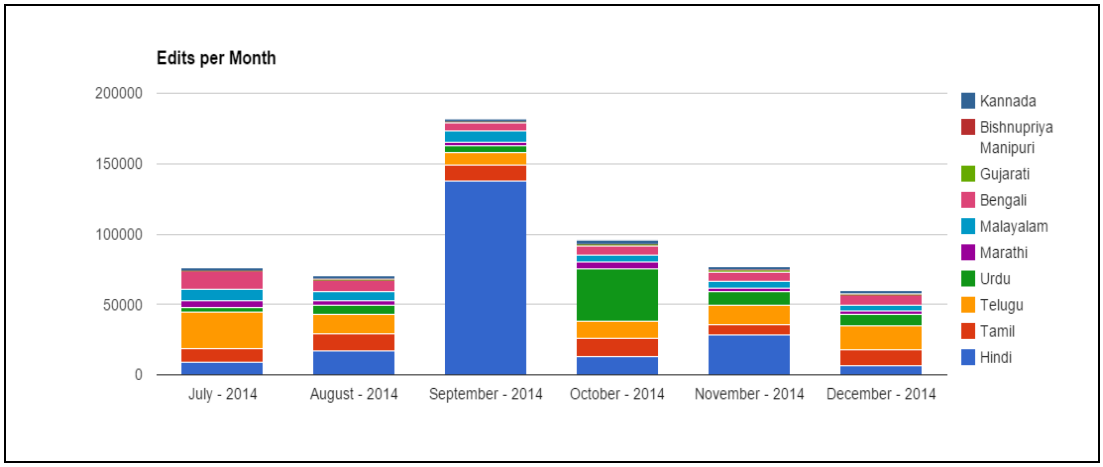

Figure3. *Edits per Month in Indian Language Wikipedia editions from July 2014 to December 2014*

1. On an average 9350 edits were performed every month in Indian Language Wikipedias from July 2014 to December 2014.
2. Hindi Wikipedia has contributed more than 2,00,000 edits during this period which was the maximum among Indian Language Wikipedias.
3. Hindi Wikipedia has performed 1,38,000 edits in September 2014 which was the maximum edits in a single month during this period.
4. More than 50000 edits were performed every month altogether in Indian Language Wikipedias during this period.
5. Only 156 edits were performed in Bishnupriya Manipuri Wikipedia during this period which is the minimum among Indian Language Wikipedias.

Analysis of Active Editors

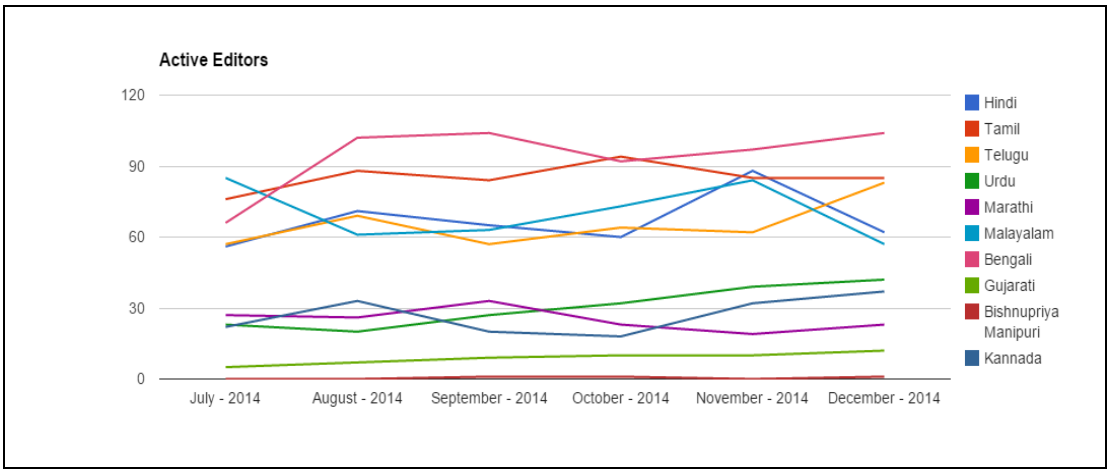

Figure4. *Active Editors in Indian Language Wikipedia editions from July 2014 to December 2014*

1. There is a fluctuation in the number of active editors except for Gujarati Wikipedia which has seen consistent growth during this period.
2. Bengali Wikipedia has crossed the 100 mark on a number of occasions, making it the only Indian Language Wikipedia to achieve this in this period.
3. Looking at the current trends, Tamil Wikipedia may soon achieve the 100 count for active editors.
4. We need to strengthen the Bishnupriya Manipuri Wikipedia which had a maximum of 1 active editor during this period.
5. The number of active editors in Malayalam Wikipedia has come down from 84 in November to 57 in December, which is a cause of concern.

Analysis of New Editors

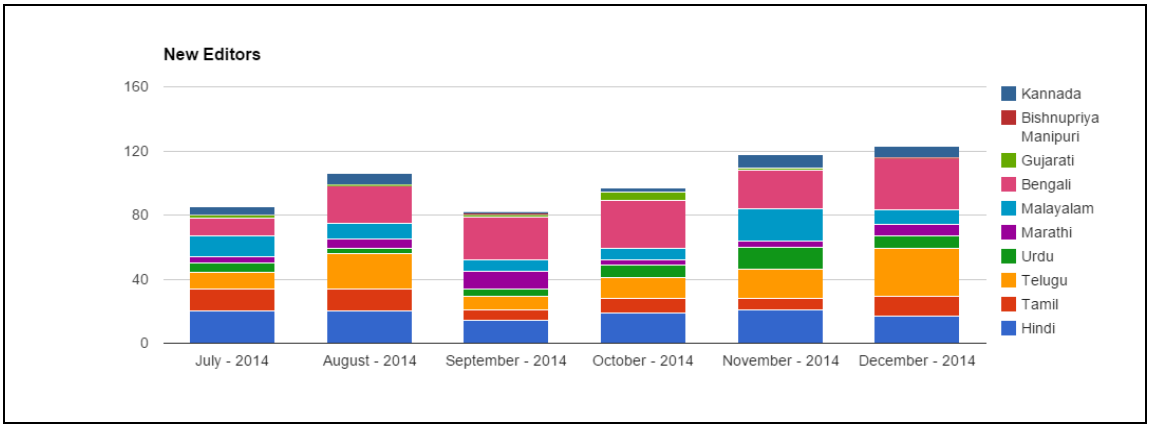

Figure5. *New Editors in Indian Language Wikipedia editions from July 2014 to December 2014*

1. On an average, 10 new editors have joined Indian Language Wikipedias every month during this period.
2. Overall, 611 new editors have joined Indian Language Wikipedias during this period.
3. Bengali Wikipedia has contributed 147 new editors, which is the maximum among Indian Language Wikipedias in this period.
4. Hindi and Bengali Wikipedias have consistently seen more than 10 new editors joining every month.
5. The conversion rate of new editors into active editors is a challenge among all Indian Language Wikipedias.

CONCLUSION

This paper performed a comparison study of major Indian Language Wikipedias.After looking at the history of Indian Language Wikipedias, we discussed about different research areas related with Wikipedia. Finally we performed Quantitative Analysis of Indian Language Wikipedias by analysing the trends in article count, new articles per day, edits per month, active editors and new editors for the period July - December 2014. There are several possibilities for further analysis regarding content popularity, content quality etc.

REFERENCES

[1] https://en.wikipedia.org/wiki/Wikipedia
[2] https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons
[3] stats.wikimedia.org/EN/Sitemap
[4] Jos´ Felipe Ortega Soto , " Wikipedia: A Quantitative Analysis ", *Doctoral Thesis, Madrid, 2009,* p 28
[5] Jos´ Felipe Ortega Soto , " Wikipedia: A Quantitative Analysis " , *Doctoral Thesis, Madrid, 2009,* p 29
[6] Jos´ Felipe Ortega Soto , " Wikipedia: A Quantitative Analysis ", *Doctoral Thesis, Madrid, 2009,* pp 53-59
[7] http://cis-india.org/a2k/blogs/indian-language-wikipedia-statistics