Web usage Mining on usage Patterns of Wikipedia Users

T. Suganya[1], Dr. K . Prabha[2]
[1] Ph. D Research Scholar, [2]Assistant Professor,
[1,2]Department of Computer Science, Periyar University PG Extension Centre

ABSTRACT

Web Usage mining part of broad field of Data mining is essential for co-relating the structure and un-structured data. This helps in identifying the interactivity aspects of the users with the websites with the help of logs from different sources. 291 active wikipedias are available as of October 2018 being an excellent source or candidate to explore the research aspects of Web usage mining through the clickstream data available. This is also to be considered as big data due to sheer volume of clickstream logs. This paper would deduce and analyse the activity patterns of users in Wikipedia based on clickstream dataset. The system adopts open source technologies which can analyse these data and visualize for appropriate interpretations

Keywords

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

As we realise the growth of unstructured data is growing incessantly. This contributes to the big data which is sweeping the internet. Based on 451 Research survey of 200 influencers and decision-making executives report the growth of unstructured data by 60 to 80%. Storage becoming cheaper and different types are on rapid growth of rich media including Audio, Video, Images, etc., Cloud has made the storage easier and container-based technologies has simplified scalability demand handled with ease.

Clickstream is event which is logged whenever the site visitor clicks are navigates to the web content. This would help to determine from the user have navigated to the current page.Typically, these events are logged for analysis. This is a valuable contribution to understand the user behaviour. Path analysis can also be performed from this data.

Wikipedia

Wikipedia one of the largest community driven content website is having 1.8 edits per second. Its English version alone contains 5790029 articles averaging 570 new articles per day. Wikipedia houses many other sub projects like Wiktionary, Wikibooks, Wikisource, Commons, Wikinews, Wikidata, Wikiversity, MediaWiki, Wikiquote, Wikispecies, Wikivoyage, and Meta-Wiki.[1]According to Alexaa internet traffic analytics company Wikipedia Ranks in 5th Position in the world with 3.1-page views per visitor out of which 55% contributed from the Internet search[2]. Products like Amazon Alexa uses Wikipedia regularly to provides response to its users. Wikipedia has been used by researchers to get the necessary background information on the topic of research as it provides coverage and convenience at one place. [3]

Wikipedia which started in 2001[4], has immense impact on collaborative content authoring and editing. Web Usage mining of Wikipedia would help determine the usage patterns and trends. It also would help Wikipedia a Not for Profit organization with better strategies in terms of content handling. This would certainly help in the areas such as Content Personalization, Performance Improvement in content delivery, Site content Structuring, Security, Intelligence etc., [5]

Clickstream Analytics

Clickstreams are data points which are collected from various user events during their browsing session over personal computing clients such as Desktop PCs, Laptops or Mobile Apps.[6]

Clickstream analytics is analytics methodology which involves collecting, analysing and reporting the user behaviour in a website based on the navigation he does with the website and its pages. The following are some of the applications of the Clickstream Analytics:

1. Automatic recommendation of content
2. Targeted Advertisement
3. Study of trending topics
4. A/B Testing to statistically study the changes made between Option A and B

Apache Spark And Rdds

This paper intends to use Spark which is powerful analytics computing engine meant for large-scale data processing. Apache Spark is fast and general-purpose computing platform. This is being adopted as it leverages popular MapReduce model to efficiently support different type of computations and interactive queries[7]

The important aspect of Spark is to process data using in-memory processing of data pipelines. Spark is an apache project. It has four modes of data processing operations namely a) Batch b) Streaming c) Iterative and d) Interactive. It can operate on Single, Standalone and Distributed in cluster of machines. [8]. When the spark applications are run on distributed node they would be driven through Driver and Executor combinations as given in the Figure 1:

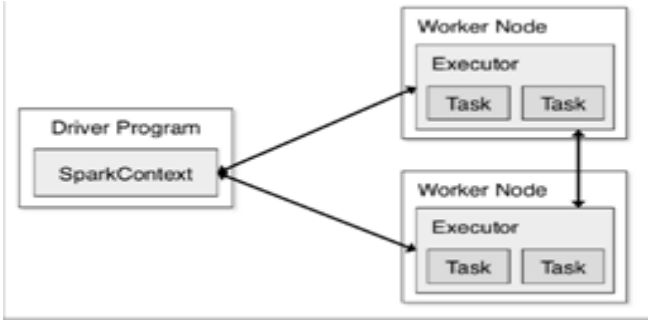


Figure 1: (Source: Learning Spark)

Spark uses the abstraction for working with the data using RDDs (Resilient Distributed Data. It is aimmutable distributed collection of elements. Spark uses RDDs to created, transform and for computational purpose. In the background RDDs are split into multiple partitions, distributed the data available in the RDDs across distributed nodes available in the cluster to perform necessary tasks on them.They are created by loading an external dataset or distributed through the driver program.

To manage the storage and computational capabilities the Spark deals the computations, transformations in lazy way. To save computational time and faster response we can persist the RDD if we need for future reuse, which will be more relevant for big data. RDDs tend to be recomputed whenever any action is performed on them.

Existing Literature

As per new privacy laws and regulations web usage mining would become more challenging. Means organizations are going to be more restrictive in terms of the content being collected from the devices which are being logged. Handling of Privacy preferences(P3P) would be one of the key challenges. There are other challenges like identifying patterns with large volume of data, evaluation of tools available, comparison of performance of these tools, data modelling around the usage data also exists with visualization problems[5]

The paper[9] discusses on the potential impact on user navigation based on the structure of Wikipedia articles. It goes across the details on the user bias happening based on the structure of the Wikipedia article content. This was evaluating more than 56,961,992 clicks during the study.

Random forest which is consuming more processing time in terms of predicting the visitors was discussed in the paper[10] in 2018. Evaluating the log of ecommerce site was at 45.45% and random forest at 95.45%.

Demanded hyperlink prediction discussed in the paper[11] by LaxmiAmulyaGundala and Francesca Spezzano. This was focusing on the solution to estimate the time duration of the search and predicts new hyperlinks in the descending order in terms of duration. This has achieved an AUROC of 0.77 on the Wikipedia dataset. There is a plan to deal with large dataset as the monthly dataset if being released. Attaining Higher value of AUROC at 0.83 was also discussed by limiting to only network-based features.

As we are awestruck with Big data and analytical capabilities leveraging Realtime clickstream data has been used largely in Targeted Digital Advertisements. This requires a good understanding and use of distributed technologies to deliver results with appropriate data pipeline. This also helps to identify the meta data such as Geography, Client used, Timezone aspects etc., Research by RamannaHanamanthrao, S. Thejaswini[12] implemented a Data pipeline with Kibana, ELK Stack along with Apache spark and Kafka.

Apriorialogorithm was used to analyse the level of access to the web pages based on the webcontents which has been consumed by the users in the research by Supriyadi[13]. It was using the Mysql to store the data results based on the analysis done on Shopping cart data. This paper also outlined the steps such as cleaning and data aggregation done on the data source with a size of 632146266 bytes.

To understand the user behaviour with the application/website or products the nature of the users to be modelled appropriately. By appropriately analysing the user and clickstream data we would be able to interpret the semantic information and trends of the captured behaviours. In this paper[6] more than 16000 clickstream traces has been used to implement a unsupervised learning systems to model the user behaviours.

Clickstream data of a website can be very helpful in providing the navigation context for the end users and thereby improving the overall experience[14]. The research concludes on the navigational importance of the content and its relationship with the user behaviour. Also outlines the potential areas of improvement in terms of the content, section segregation etc,

In 2019, the paper [15] by Mohsen discussed various prerequisite relations based on the navigation patterns of users on the Wikipedia using supervised learning approach. It derives a navigation network structure to identify dependencies on such concepts for learning. Conclusively it outlined that human learning goes backward from the advanced topics as per the necessary pre-requisites based on the semantic space of Wikipedia.

Methodology

Data Source

Wikipedia clickstream typically consists of Referrer and resource pairs with weightage. The weightage refers to the frequency in which they navigate from one page to another.Wikipedia provides the dataset it the form of Raw XML Dump, Page Views and Clickstream.Wikipedia clickstream data includes Page counts with 22 TB, Pageviews with 255 MB, Clickstream data with 1.2 GB, Wikipedia – 54GB.

The Wikipedia data is a combination of referent and current pairs. When a web client or mobile client requests a resource through a link or by performing a search the URI of the webpage which links to that resources is included in the HTTP header as referrer. The content would be based on the Wikipedia Namespace which is recognized by the Media wiki software. The representative content is provide in Figure 1.All the dumps of the clickstream available from the Wikimedia Dumps[16].

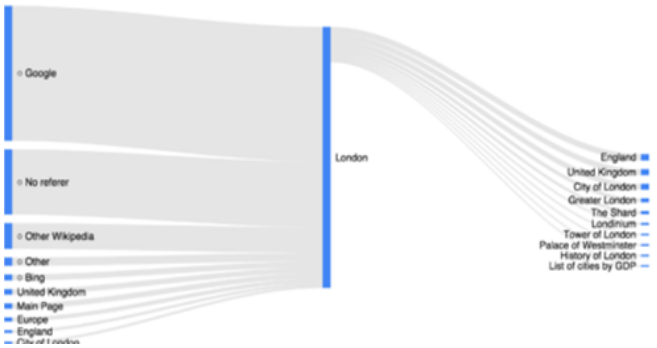


Figure 2: Source : Media Wiki

There are 4 key attributes in this data source:

Sno#	Attribute	Meaning
1	Prev	Referrer – Title from which the user has navigated to the current article.
2	Curr	Current – Typically title of the current article
3	Type	This explains the combination of the Prev andCurr with the following approach: Link – Both referrer and Current are articles in Wikipedia External – Referrer is an external website Other – Spoofed referrer
4	N	Number of occurrences of the Prev, Curr Pair (Weightage)

The following are some of the key interpretations which needs to be analysed from the clickstream data set for the purpose of Referrer and some of these are fixed better analysis:

- Article – Wikipedia main namespace -> the article title
- Page from another Wikimedia project -> other-internal
- Referred from an external search engine -> other-search
- Referred directly from any other external site -> other-external
- Unable to determine the referrer -an empty referrer -> other-empty
- Non-deterministic - anything else -> other-other

Typical existing strategies will adopt the mechanism of the leveraging the polyglot persistence mechanism by leveraging different tools and databases in the context of the below given context diagram approach (Figure 3). This would have problems in terms of performance tuning and different infrastructure dependencies as well. Apache Spark helps to managing this very well by appropriate aggregation strategies.

Implementation And Discussion

January 2019 take has been taken for the purpose of analysis which contains 30.9 million records. This file has been available in TSV (Tab Separated Value) format. This has been processed by loading the TSV with the following Schema:

```
StructType(
List(
StructField("prev", StringType, true),
StructField("curry",StringType,true),
StructField("type",StringType, true),
StructField("occur", IntegerType, true)
))
```

By evaluating the different type of referrers Link contributes user navigation up to 75% (Figure 4) and external referrals go up to 25% for Jan 2019. We were able retrieve this data with in 24s in 3 stages in apache spark. The Spark was run on with single mode on MAC OS(Mojave) with 8 GB RAM and 1.8 GHz Intel Core i5 Processors. The implementation was done using Spark and Scala combination. Using Spark 2.4.1 and Scala version 2.11.12 on 64-bit Server.This was a able to load the TSV file in 1 stage within 0.6 seconds.

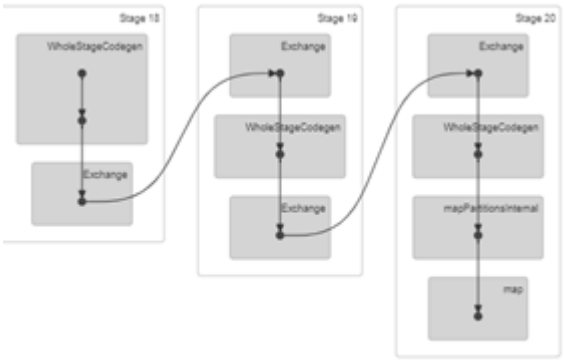


Figure 4 (DAG of extracting Total referrers)

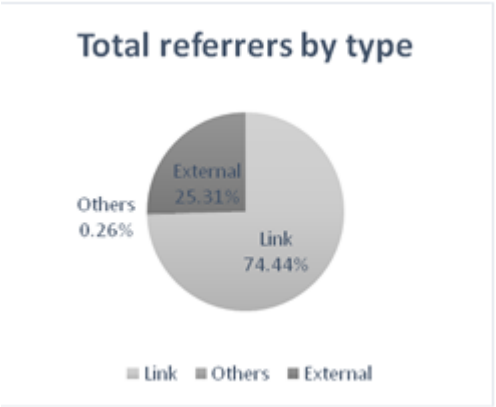


Figure 5 (Referrers by Type)

DAG of this implementation can be found in the representation above DAG available from Apache Spark Execution logs at 3 stages (Figure 4). Similarly the metrics available with respect to Apache Spark such as Scheduler

Delay, Task Deserialization time, Peak Execution Memory, etc., would help to optimize the interactions with the data.

For understanding of website users where do they come from and how they have been redirected to the Wikipedia site, its essential to understand the different set of referral means. By understanding the attribute "Prev" it would be easier to identify means such as referral though Search Engines, External website links, Internal links, etc., To identify different set of referrals the Wikipedia users arrive at Wikipedia articles can be aggregated across through the following figure given below (Figure 6).

This also helps to narrow down the navigational aspects of using search engine rather than leveraging the Wikipedia search and sections available in the wiki articles. Wikipedia being opensource project and supporting many languages key inferences in this area would drive appropriate analogy for improving navigational aspects of accessing the content with ease.

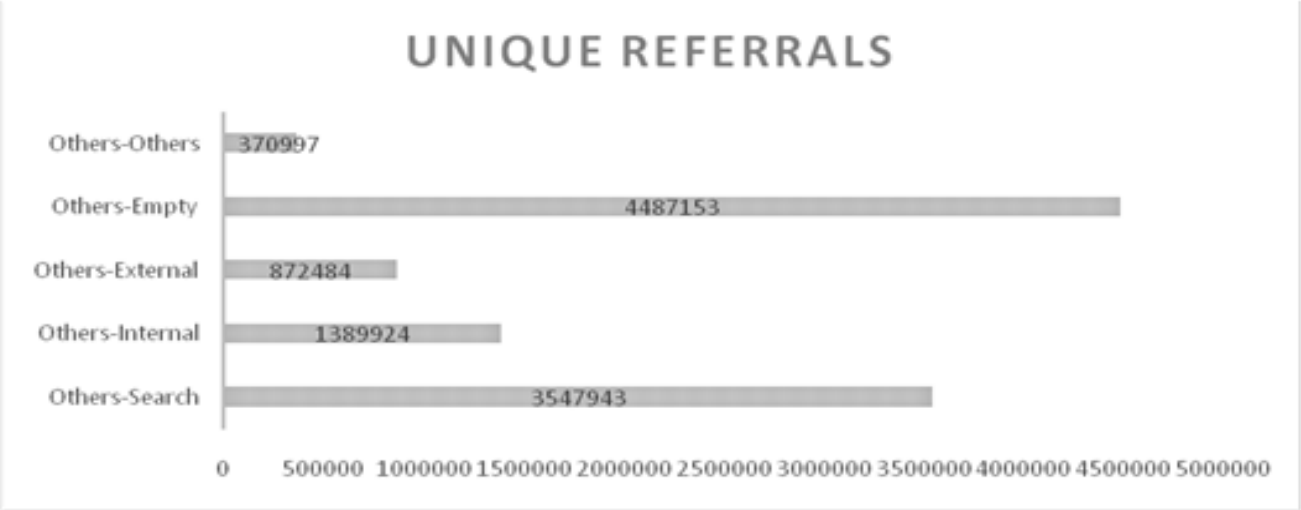
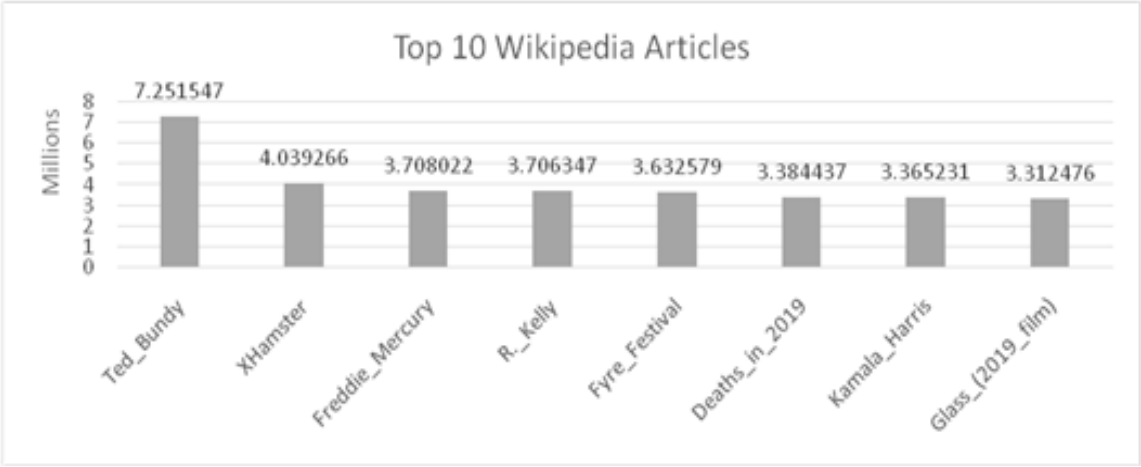


Figure 6 (Different sources which has referred Wikipedia articles)

Based on this its understood that Others-Empty generally refers to https traffic seems higher but on the reality Search engines such as Google, Bing has been the highest referral which sends the most traffic to Wikipedia (~354 million).

The below given provides the details insights on the top 10 Wikipediaarticles for Jan 2019 which is excluding the Main page which is at ~51 billion hits and hyphen-minus page at 4 billion which might skew the chart for appropriate representation.



Apache Spark Jobs took 1381 MB data for processing this jobs and completed extraction in 34 seconds which is pretty impressive on the volume of data it has process.

Conclusion

This paper approached the analysis of Wikipedia clickstream data and it attempted to answer the questions on the navigational patterns from the external and internal sources of Wikipedia. Identified the time taken for the processing the data with the help of Apache Spark. Also helped to identify ways for determining the highly visited links from the clickstream data 30.9 million records. Identified the volume of hits to specific pages to determine how is the popularity of certain articles by period. This also paves way to future research in the direction of co-relating the external happenings which could drive more inquisitiveness in learning about specific subject in Wikipedia for additional content authoring and viewership. here is a potential to extend the research in the areas to derive a specific trends over a six months period by considering multiple month Wikipedia clickstream data. Potential research opportunity exists to predict and protect vandalism behaviour on Wikipedia edit stream.

References

[1] T. LEITCH, Wikipedia U - Knowledge, Authority, and Liberal Education in the Digital Age, 2014: Johns Hopkins University Press, Baltimore.

[2] "Alexa Top 500 Global Sites," Alexa Internet, Inc. 1996 - 2018 , [Online]. Available: https://www.alexa.com/topsites. [Accessed 19 01 2018].

[3] A. J. Head and M. B. Eisenberg, "How Today's College Students Use Wikipedia for Course-Related Research," First Monday, vol. 15, no. 3, p. , 2010.

[4] J. . Broughton, Wikipedia – The Missing Manual, ed., vol. , N. . Barber and P. . Meyers, Eds., , : "O'Reilly Media, Inc.", 2008, p. .

[5] M. Aldekhail, "Application and Significance of Web Usage Mining in the 21st Century: A Literature Review," International Journal of Computer Theory and Engineering, vol. 8, no. 1, pp. 41-47, 2016.

[6] G. . Wang, X. . Zhang, S. . Tang, C. . Wilson, H. . Zheng and B. Y. Zhao, "Clickstream User Behavior Models," ACM Transactions on The Web, vol. 11, no. 4, p. 21, 2017.

[7] A. K. P. W. a. M. Z. Holden Karau, Learning Spark, Sebastopol, CA 95472: O'Reilly Media, 2014.

[8] A. Nandhi, Spark for Python Developers, Brimingham-Mumbai: Packt Publishing, 2015.

[9] D. . Lamprecht, K. . Lerman, D. . Helic and M. . Strohmaier, "How the structure of Wikipedia articles influences user navigation," The New Review of Hypermedia and Multimedia, vol. 23, no. 1, pp. 29-50, 2017.

[10] M. Dr.B.Sateesh Kumar, "CLICKSTREAM DATA PREDICTION USING RANDOM FOREST CLASSIFIER," Indian J.Sci.Res., vol. 17, no. 2, pp. 29 - 32, 2018.

[11] L. A. Gundala and F. . Spezzano, "Readers' Demanded Hyperlink Prediction in Wikipedia," , 2018. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/www2018c.html. [Accessed 1 2 2019].

[12] R. . Hanamanthrao and S. . Thejaswini, "Real-time clickstream data analytics and visualization," , 2017. [Online]. Available: https://semanticscholar.org/paper/real-time-clickstream-data-analytics-and-hanamanthrao-thejaswini/67b35fefcd4ebdac701b75a30580ef701d28284f. [Accessed 4 3 2019].

[13] S. . Supriyadi, Y. . Nurhadryani and A. I. Suroso, "Website Content Analysis Using Clickstream Data and Apriori Algorithm," TELKOMNIKA : Indonesian Journal of Electrical Engineering, vol. 16, no. 5, pp. 2118-2126, 2018.

[14] D. L. a. K. L. a. D. H. a. M. Strohmaier, "How the structure of Wikipedia articles influences user navigation," New Review of Hypermedia and Multimedia, vol. 23, no. 1, pp. 29-50, 2017.

[15] J. G. J. L. A. K. L. Mohsen Sayyadiharikandeh, "Finding Prerequisite Relations using the Wikipedia Clickstream," WWW '19 Companion,, 2019.

[16] "Analytics Datasets: Clickstream," Media Wiki, 4 March 2019. [Online]. Available: https://dumps.wikimedia.org/other/clickstream/. [Accessed 4 March 2019].