Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 4, April -2017

Efficient Query Processing in Geographic Web Search Engines

[1]Akanksha Singh, [2]Samarth Shekhar, [3]Bhagwati Ghbhane, [4]Prof. Latika Desai,

[1,2,3,4]*Dr.D.Y.Patil institute of engineering and technology,pimpri,Pune,Maharashtra,India*

Abstract — *Microblogging services became among the foremost fashionable services on the net within the previous couple of years. This LED to important increase in information size, speed, and applications. This paper presents Venus; a system that supports period abstraction queries on microblogs. Venus supports its queries on a abstraction boundary R and a temporal boundary T, from that solely the top-k microblogs area unit came back within the question answer supported a spatio-temporal ranking operate. Supporting such queries needs Venus to digest many various period microblogs in main-memory with high rates, yet, it provides low question responses and economical memory utilization. to the present finish, Venus employs: (1) AN economical in-memory spatio-temporal index that digests high rates of incoming microblogs in real time, (2) a ascendible question processor that prune the search area, R and T, effectively to produce low question latency on various things in real time, and (3) a gaggle of memory optimisation techniques that give system directors with completely different choices to save lots of important memory resources whereas keeping the question accuracy nearly excellent. Venus memory optimisation techniques build use of the native arrival rates of microblogs to neatly shed microblogs that area unit sufficiently old to not contribute to any question answer. additionally, Venus will adaptively, in real time, alter its load shedding supported each the abstraction distribution and therefore the parameters of incoming question masses. All Venus elements will accommodate completely different abstraction and temporal ranking functions that area unit ready to capture the importance of every dimension otherwise counting on the applications needs. in depth experimental results supported real Twitter information and actual locations of Bing search queries show that Venus supports high arrival rates of up to sixty four K microblogs/second and average question latency of four unit of time.*

Keywordst; *Microblogs, spatial, location, temporal, performance, efficiency, scalability, memory optimization, social*

I. INTRODUCTION

Social media websites have grabbed huge attention within the last decade because of its growing quality and unprecedentedly giant user base. The new wave of user-interactive microblogging services, e.g., tweets, comments on Facebook or news websites, or Foursquare check-in's, has become the clear frontrunner within the social media race with the most important range of users ever and highest users activity in consistent rates. for instance, Twitter has 288+ Million active users World Health Organization generate 500+ Million daily tweets, whereas Facebook has one.35+ Billion users World Health Organization post three.2+ Billion daily comments [3]. driven by the advances in wireless communication and therefore the quality of GPS-equipped mobile devices, microblogs service suppliers have enabled users to connect location data with their posts. Thus, Facebook additional the choices of location check-ins and close to wherever users will state a close-by location of their standing messages, Twitter mechanically captures the GPS coordinates from mobile devices, per user permission, and Foursquare options area unit all round the location data and therefore the whereabouts of its users. Consequently, a inordinateness of location data is presently on the market in microblogs. we tend to exploit of the supply of location data in microblogs to support spatio-temporal search queries wherever users area unit ready to browse recent microblogs close to their locations in real time. Users of our projected queries embody news agencies (e.g., CNN and Reuters) to possess a first-hand information on events during a sure space, advertising services to serve geo-targeted ads to their customers supported close events, or people World Health Organization need to grasp in progress activities during a sure space. for instance, in Gregorian calendar month 2013, la Times rumored [4] however individuals rush to Twitter for period breaking news concerning Hub of the Universe Marathon explosions. Such users might not apprehend the acceptable keyword or hash tag to go looking for. Instead, they need to grasp the recently announce microblogs during a sure explicit space. Thus, our goal here isn't to switch the normal keyword search in microblogs,but rather to produce another necessary search possibility for localized microblogs. the solution of our spatio-temporal queries may be fed to alternative modules for any process, which can embody event detection, keyword search, entity resolution, sentiment analysis, or image.

II. LITERATURE SURVEY

1. “Efficient processing of window queries in the pyramid data structure,
Authors: W. G. Aref and H. Samet

Window operations function the idea of variety of queries that may be posed during a spatial info. samples of these window-based queries embrace the exist question (i.e., deciding whether or not or not a spatial feature exists within a window) and also the report question, (i.e., reportage the identity of all the options that exist within a window). Algorithms square measure represented for responsive window queries in &amp;Ogr;(n log logT) time for a window of size n x n during a feature area (e.g., associate image) of size T x T (e.g., picture element elements). the importance of this result's that despite the fact that the window contains n2 picture element components, the worst-case time quality of the algorithms is nearly linearly proportional (and not quadratic) to the window diameter, and doesn't rely upon alternative factors. The higher than quality bounds square measure achieved via the introduction of the unfinished pyramid system (a variant of the pyramid information structure) because the underlying illustration to store spatial options and to answer queries on them.

2. “Mercury:A memory-constrained spatio-temporal real-time search on microblogs

Authors: A. Magdy, M. F. Mokbel, S. Elnikety, S. Nath, and Y.

This paper presents Mercury; a system for period support of top-k spatio-temporal queries on microblogs, wherever users square measure ready to browse recent microblogs close to their locations. With high arrival rates of microblogs, Mercury ensures period question response at intervals a good memory-constrained setting. Mercury bounds its search house to incorporate solely those microblogs that have arrived at intervals sure abstraction and temporal boundaries, during which solely the top-k microblogs, in step with a spatio-temporal ranking operate, square measure came within the search results. Mercury employs: (a) a climbable dynamic in-memory index structure that's capable of digesting all incoming microblogs, (b) Associate in Nursing economical question processor that exploits the in-memory index through spatio-temporal pruning techniques that cut back the quantity of visited microblogs to come the ultimate answer, (c) Associate in Nursing index size standardization module that dynamically finds and adjusts the minimum index size to confirm that incoming queries are answered accurately, and (d) a load shedding technique that trades slight decrease in question accuracy for vital storage savings.

3. .logging infrastructure for data analytics at twitter,”

AUTHORS: G. Lee, J. Lin, C. Liu, A. Lorek, and D. V. Ryaboy,

In recent years, there has been a considerable quantity of labor on giant-scale information analytics victimisation Hadoop-based platforms running on large clusters of trade goods machines. A less-explored topic is however those information, dominated by application logs, area unit collected and structured to start with. during this paper, we tend to gift Twitter’s production work infrastructure and its evolution from application-specific work to a unified “client events” log format, wherever messages area unit captured in common, well-formatted, versatile Thrift messages. Since most analytics tasks take into account the user session because the basic unit of study, we tend to pre-materialize “session sequences”, that area unit compact summaries that may answer an outsized category of common queries quickly. the event of this infrastructure has efficient log assortment and information analysis, thereby rising our ability to quickly experiment and ingeminate on numerous aspects of the service.

4. Large-scale machine learning at twitter

AUTHORS: J. Lin and A. Kolcz,

The success of data-driven solutions to troublesome issues, at the side of the dropping prices of storing and process huge amounts of information, has light-emitting diode to growing interest in large-scale machine learning. This paper presents a case study of Twitter’s integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform. we start with an summary of this platform, that handles “traditional” knowledge deposit and business intelligence tasks for the organization. The core of this work lies in recent Pig extensions to supply prognostic analytics capabilities that incorporate machine learning, centered specifically on supervised classification. specifically, we've got known random gradient descent techniques for on-line learning and ensemble strategies as being extremely amenable to scaling bent massive amounts of information. In our deployed resolution, common machine learning tasks like knowledge sampling, feature generation, training, and testing will be accomplished directly in Pig, via fastidiously crafted loaders, storage functions, and user-defined functions.

5. .“Earlybird: Real-time search at twitter,

AUTHORS: M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin,

The web these days is more and more characterised by social and period signals, that we tend to believe represent 2 frontiers in data retrieval. during this paper, we tend to gift Early bird, the core retrieval engine that powers Twitter's period search service. though Early bird builds and maintains inverted indexes like nearly all trendy retrieval engines, its index structures dissent from those designed to support ancient internet search. we tend to describe these variations and gift the explanation behind our style. A key demand of period search is that the ability to ingest content speedily and create it searchable right away, whereas at the same time supporting low-latency, high-throughput question analysis. These demands area unit met with a single-writer, multiple-reader concurrency model and also the targeted use of memory barriers. Early bird represents some extent within the style area of period search engines that has worked well for Twitter's wants. By sharing our experiences, we tend to hope to spur extra interest and innovation during this exciting area.

III. PROPOSED SYSTEM

We propose effective memory improvement techniques: (1) we tend to an analytically develop an index size standardisation technique that achieves important memory savings (up to fifty p.c) while not sacrificing the question answer quality (more than ninety nine percent accuracy). the most plan is to use the range of arrival rates per regions.

The planned system contains following process:

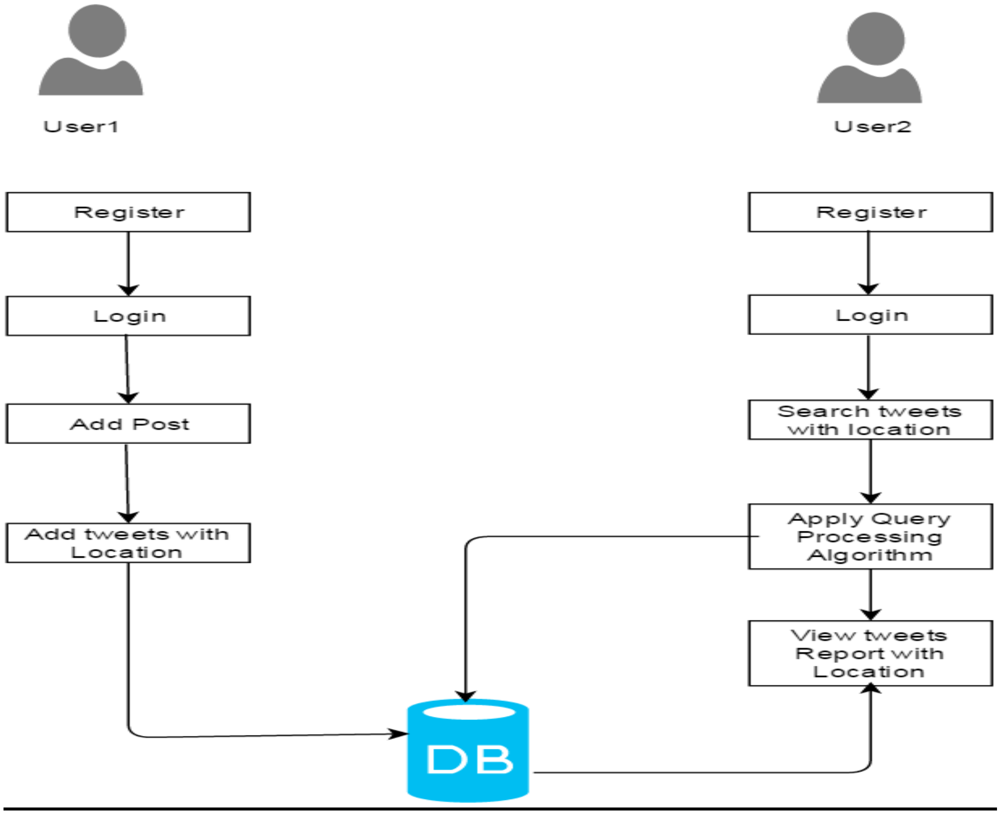
User1
User2
Register
Login
Add Post
Add tweets with Location
Register
Login
Search tweets with location
Apply Query Processing Algorithm
View tweets Report with Location
DB

Figure: Planned System Design

ADVANTAGES OF PLANNED SYSTEM:

- Venus will use completely different ranking functions to be ready to serve needs of various applications.
- These techniques catch the spatial distribution of the incoming queries additionally because the spatial access patterns of the hold on microblogs in order that they create the storage overhead to its least levels (up to eighty p.c less storage) whereas permit to answer queries with virtually good accuracy (more than ninety nine p.c all told cases).

V. CONCLUSION

We have bestowed Venus; a system for time period support of spatio-temporal queries on microblogs, wherever users request a group of recent k microblogs close to their locations. Venus works underneath a difficult setting, wherever microblogs arrive with terribly high arrival rates. Venus employs economical in-memory compartmentalization to support up to sixty four Kmicroblogs/second and spatio-temporal pruning techniques
to provide time period question response of four time unit. additionally, effective load shedding modules ar utilized to well shed the useless knowledge whereas providing virtually excellent question accuracy.

ACKNOWLEDGMENT

We might want to thank the analysts and also distributers for making their assets accessible. We additionally appreciative to commentator for their significant recommendations furthermore thank the school powers for giving the obliged base and backing.

REFRENCES

[1] (2013). Twitter Statistics [Online]. Available: http://expandedramblings. com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

[2] (2014). Twitter Data Grants, 2014 [Online]. Available: https://blog. twitter.com/2014/introducing-twitter-data-grants

[3] (2012). Facebook Statistics [Online]. Available: https://www.facebook. com/business/power-of-advertising

[4] (2013). After Boston Explosions, People Rush to Twitter for Breaking News [Online]. Available: http://www.latimes.com/business/ technology/la-fi-tn-after-boston-explosions-people-rush-to-twitter- for-breaking-news-20130415,0,3729783.story

[5] W. G. Aref and H. Samet, “Efficient processing of window queries in the pyramid data structure,” in Proc. 9th ACM SIGACT-SIGMOD- SIGART Symp. Principles Database Syst., 1990, pp. 265–272.

[6] A. Magdy, M. F. Mokbel, S. Elnikety, S. Nath, and Y. He, “Mercury: A memory-constrained spatio-temporal real-time search on microblogs,” in Proc. IEEE 30th Int. Conf. Data Eng., 2014, pp. 172–183.

[7] G. Lee, J. Lin, C. Liu, A. Lorek, and D. V. Ryaboy, “The unified logging infrastructure for data analytics at twitter,” Proc. Very Large Data Base, vol. 5, no. 12, pp. 1771–1780, 2012.

[8] J. Lin and A. Kolcz, “Large-scale machine learning at twitter,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2012, pp. 793–804.

[9] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin, “Earlybird: Real-time search at twitter,” in Proc. IEEE 28th Int. Conf. Data Eng., 2012, pp. 1360–1369.

[10] C. Chen, F. Li, B. C. Ooi, and S. Wu, “TI: An efficient indexing mechanism for real-time search on tweets,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 649–660.

[11] Analytics: An Intelligent Approach in Clinical Trail Management Ankit Lodha* Analytics Operations Lead, Amgen, Thousand Oaks, California, USA.

[12] Agile: Open Innovation to Revolutionize Pharmaceutical Strategy Ankit Lodha University of Redlands, 333 N Glenoaks Blvd #630, Burbank, CA 91502.

[13] Clinical Analytics – Transforming Clinical Development through Big Data Ankit Lodha University of Redlands, 333 N Glenoaks Blvd #630, Burbank, CA 91502.