User Awareness Semantic Search Based on Approximate Methodology in Massive Storage System

Kommu Anusha & P. Lalitha Surya Kumari
M.Tech (Computer Science & Engineering) Bharat Institute of Engg & Tech, Ibrahimpatnam(M) R.R Dist.
M.Tech, (Ph.D) Associate Professor, Dept of CSE Bharat Institute of Engg & Tech, Ibrahimpatnam(M) R.R Dist.

Abstract:*Textual documents produced and distributed on the Internet are ever changing in numerous forms. Furthermost of prevailing works aredevoted to topic modeling and the progression of individual topics, while sequential relations of topics in successive documents publishedby a specific user are ignored. Inthis paper we explore a semantic Search method for processingthe large scale data volumes in the cloud. We use the hashingalgorithms and flat structured addressing schemes for theretrieval of the data by using the semantic queries. Here the datais processed by using the caching techniques and the data isretrieved by using the semantic query. This techniques reducesthe time delay for the retrieval of the data from the large scalestorage systems.*

Keywords-Personalization, user modeling, document streams, rare events,

I. INTRODUCTION

The emergent semantic-based knowledge technologies [Berners-Lee 2001], also labeled “the semantic web” a term which is falling short today for the breadth and reach of this area – aim at endowing softwaresystems with a deeper insight into the meaning of the data they manipulate, create, store, and exchange.Functional areas where the capabilities enabled by such advanced semantic representations are beingenvisioned include, among others, sharing (exchange and integration), retrieval (search, filtering, browsing,recommendation), and presentation (visualization, navigation, composition) of application-domainknowledge. More recently, a growing interest is being raised for the potential of semantically richdescriptions to achieve improvements in the area of personalization technologies [Gauch 2004].In the context of software systems and applications, personalization refers, in general terms, to the development of models and systems that represent and capture user preferences, goals, needs, knowledge, demographicinformation, environment, device, mood, capabilities, disabilities, etc., and use this information in the system tobetter meet user needs and expectations, and help users achieve their tasks and goals more efficiently [Kobsa2001]. Personalization technologies gained significance in the 90’s, with the boost of large-scale computingnetworks which enabled the deployment of services to massive, heterogeneous, and less predictable and consumer audiences. As the number of services and the volume of content (text and multimedia; public, commercial and personal) in these networks keep growing, personalization is more than ever a critical enabler inhelping consumers manage capacity and complexity, and help vendors (content providers, managers, brokers,distributors, technology providers) reach their target audience and attain a competitive edge. Semantic-based techniques enable to infuse software systems with a more precise understanding ofapplication-domain knowledge, and henceforth, provide better means to define user needs, preferences, andactivities within or with regards to the system. Moreover, they can be used for a richer representation of user-related information itself. In this paper, we describe a personalization system that has been developed in thisperspective. The system is part of a wider framework being developed in the aceMedia project [Kompatsiaris2004]. aceMedia aims at integrating knowledge, semantics and multimedia technologies to solve userproblems via intelligent content and applications. Using automatic content analysis tools, the aceMediasystem augments multimedia items with self-descriptive metadata, thereby building up an understanding ofthe meaning of contents along different dimensions: application-domain concepts, visual semantics, mediaproperties, formats, etc. aceMedia exploits this knowledge to provide new or better content services, such asintelligent search, or advanced browsing and navigation facilities [Grira 2004]

In the view to growth theoverall performance of processing the massive amount of information thefollowing issues which can be related to the statistics research wantto be addressed.Increased get admission to latency: The accessing of the information may also takea large quantity of time because of the extended quantity ofrequests which may additionally create the bottleneck in the cloud serversthe response to the requests may additionally take time in view that the prevailingprocedures to an unordered search of the statistics and analysisin particular relays at the device based lumps of information documents and thefeatures related to the multimedia primarily based images [3]. If we usethe method which relays on the precise content it is able to producethe extended amounts of auxiliary information which may additionally growththe bottleneck of the device.

High Energy Consumption: Due to the bottleneck created inthe cloud servers.The reaction to the requests may also postpone dueto the put off inside the reaction time electricity intake will behigh Hence the reaction time want to be reduced to lessen thepower intake. The bugs within the statistics want to be correctedto lessen the electricity consumption which may reduce theneed for digital servers.

Data Authentication: The Cloud servers need to offerauthorization to the users so that best the legal andrequested users can get entry to the data traffic can be created inthe cloud which may additionally lessen the processing velocity.

High query cost: In order to get admission to the facts in the cloud,processing of the queries are within the high call for. The studiesbased on the facts of the cloud may additionally devour ample systemssources inclusive of the memory area, I/O bandwidth, High-performance multicore processors [4]. The foremost wrongdoer for theexpanded quantity of useful resource charges is the bottleneck caused bythe high-overall performance query operations.

In order to overcome the above problems the followingstrategies can be used together with Flat Structured addressing [5]Algorithms which includes the locality touchy algorithms [6]cuckoo based totally hashing algorithms may be used. In order tomixture the semantically correlated files SANE [7] methodcan be used to combine the correlated files into flat andpossible businesses to achieve expanded processing of thesemantic queries.

II. RELATED WORKS

The real-time and cost-efficient scheme that's known as theSmartEye is used inside the cloud-supported catastropheEnvironments. The major concept of the SmartEye is that itincorporates the improved great of provider inside the networkdeduplication scheme for the networks that are softwaredistinct. The idea of the SmartEye is that it aggregates thenetwork flows which contain the identical capabilities through the usage ofthe semantic hashing and presents the widely recognized communication services for all of the flows which are aggregated,here the SmartEye isn't associated with a single float it mainlyrelays at the aggregated flows. To acquire this SmartEye usesthe following optimization strategies called the semanticprimarily based hashing and the gap-efficient filters. Efficient sharingof the image is used to detect the disaster and popularity ofthe scene[8].

The accelerated use of the smartphones that are prepared withthe digicam and drugs had led the users to capture the massiveamount of films and images. In 2011 the worldwidepurchaser digital storage wishes had grown to 329 exabytes andinside the year 2016, it had grown to 4.1 zettabytes [9]. To deal witha large range of features that are extracted from thepictures a locality sensitive hashing algorithms are used to vicinity the nearby descriptors in the index. This technique presents theview to approximate the similarity in the queries. Whichallows examining simplest small fraction in the database. Evenalthough locality sensitive hashing scheme has a higheroverall performance that's related to theoretical view, a sensibleimplementation is very sluggish [10]. The local binary pattern is usedfor the face image reputation here the face image is dividedinto numerous elements by means of applying the neighborhood binary pattern functionthe divided parts are extracted and concatenated to apply as facedescriptor. This technique is applied to understand the facebeneath various demanding situations [11].

Further SmartStore is used which incorporates the metadatasemantics of files to mixture correlated documents into semanticbased groups and retrieval gear to retrieve the statistics. Toimprove the device scalability and to reduce the query latencythe decentralized layout techniques may be used for thecomplicated queries which can be the higher technique for constructing thesemantic associated caching. Smart shop limits the complexity forsearching the queries for the unmarried or the semanticallyaggregated corporations and it limits the use of incorporating thebrute force seek in the device [12]. Other strategies can beutilized by extracting the awesome capabilities from the images andaggregating it right into an unmarried characteristic. These extracted capabilitiesmay be matched with the high chance for huge databasefunctions from many images [13]. Due to the increased increaseand complexity, data volumes had led the very high demandfor efficient searching of the statistics inside the cloud.

A present garage device inside the cloud doesn't offer a wellfunctionality for the data analytics associated with the actual time.Because the precise value and the real worth of the data dependsheavily on how the data analytics must be carried within the actualtime. Since the large fractions of the statistics terminates with theirdata being misplaced and notably reduced due to the statisticsstaleness.

To address the above hassle a cost-efficient method referred to as-isthe FAST is carried out for searchable analytics of the information.Hence the primary concept of the FAST is to have a look at and analyzethe semantic correlation among the datasets by using thecorrelation primarily based hashing and feasible flat structuredaddressing to extremely lessen the processing Latency even asincorporating small loss inside the data search and correctness.

The idea of the FAST is to swiftly perceive the correlated documentsand lowering the wideness of the statistics to be processed [14]. Inthe present method some other technique used is to hash thepoints from the database via confirming that the possibility ofthe collisions is lesser for the deviceswhich might be located at a bigdistance apart from the items which are located closed to everyother.This technique has experimental proof which affords angreen development within the runtime in comparison to differenttechniques for searchable excessive dimensional areas by usinghierarchical tree decomposition [15]. A locality sensitivehashing scheme is used for approximating the closestthe neighbor hassle which is underneath the L_p norm, primarily based on thesolid distributions this scheme improves the going for walks time ofthe algorithm, this set of rules finds an appropriate nearest neighborin *O(log n)* time for satisfying the positive bounded increasesituation[16].

III. APPROACH

In the view to increase the accessing capability of the data inthe cloud storage systems the following techniques are usedsuch as the hashing algorithms are used in this paper. Thefollowing Fig.1 shows how the data is placed in the specificmanner.

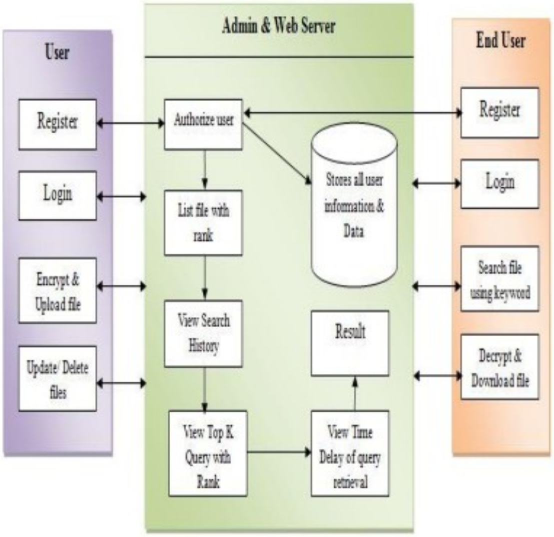

Fig. 1 System Model for Data Storage

In this approach initially the user registers to the cloud serverafter the registration the user login's to the cloud server laterthe user can upload the files in to the cloud server byencrypting the data here the user can add n number of files andcan update or delete the files which is been added to the cloudserver. By retrieving the files from the cloud server databasewhich is been added by the user the admin lists all the fileswith rank, views the search history of the previous users andmakes the lists of top k queries in rank, views the results of thetime delay and stores the updated information in the database,the end user can retrieve the file by login to the cloud serverand can search the file by using the keyword, decrypts the filesand the user can download the required file.

The following Fig. 2 shows the flow chart diagram of how theend user retrieves file by using the semantic search, here thedata user registers to the server, if the user is already beenregistered the data user needs to login to the server or elseneeds to register to the webServer after successful registrationthe user uploads the file to the server.

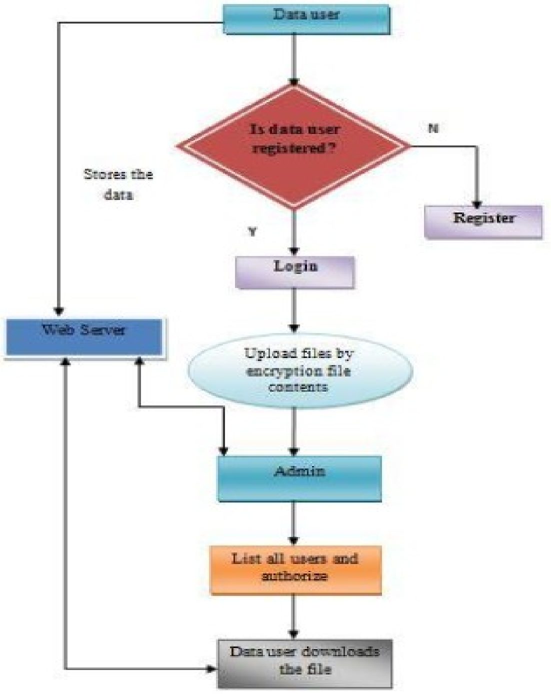
Data user
Is data user registered ?
N
Register
Stores the data
Y
Login
Web Server
Upload files by encryption file contents
Admin
List all users and authorize
Data user downloads the file

Fig. 2 Flow chart for data retrieval by the user

Here the files which is been uploaded by the user is retrievedby the admin, the admin stores all the files in the cloud serverbased on the file ranks, the files are placed by using Cachingtechniques and the locality sensitive hashing algorithms whichhas the complexity of *O(1)*. Locality sensitive hashingalgorithm is used to search and aggregate identical files intothe correlation based groups. This provides the retrieval to benarrowed to the one or the limited number of groups byincorporating correlation awareness. Later when the userrequests for the specific file, the admin uses bloom filters forthe searching of the files. Bloom filters has the features ofsimplicity and easy to use. In bloom filters the large sizevectors of files is hashed effectively to Identify similar files inthe real time manner. Bloom filter uses the method based onmultiple identical vectors, if two files contain identical vectorsit maintains the list of the memberships of the vectors andmakes the lists of the similar files. By using this bloom filtersthe admin searches the file requested by the user, and the userdownloads the requested file. If the requested file is notavailable in the server database, the admin lists the correlatedand similar files. Here the user searches the files by usingsemantic keywords. All the user transactions such as therequest for the files, the files which are downloaded by theuser, files has been searched by the user, files uploaded andother user information is stored in the server database.

IV. CONCLUSION

In this paper we had discovered the various techniques used toincrease the get into capability in the existing cloud storagesystems and how to access the data in the cloud servers, Thedifficultieshappened due to the storage of large amount ofdata. Here we had explored how data need to be processedbefore it is used in any explicit approach. And we hadexplored numerous hashing algorithms such as the LocalitySensitive hashing algorithm for hashing purpose and also hadexplored the bloom filters for filtering purpose to access thedata through the use of semantic queries. By using thesetechniques we can decrease the time delay experienced forsearching of the specific file and their retrieval from the largescale storage systems.

REFERENCES

[1] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, “Frequent patternmining with uncertain data,” in Proc. ACM SIGKDD’09, 2009, pp.29–38.

[2] R. Agrawal and R. Srikant, “Mining sequential patterns,” in Proc.IEEE ICDE’95, 1995, pp. 3–14

[3]Real-time Semantic Search using Approximate Methodology for Largescale Storage SystemsYu Hua, Senior Member, IEEE, Hong Jiang, Fellow, IEEE, Dan Feng,Member, IEEE.

[4]D. Zhan, H. Jiang, and S. C. Seth, CLU: Co-optimizing Locality andUtility in Thread-Aware Capacity Management for Shared Last LevelCaches, IEEE Transactions on Computers, vol. 63, no. 7, pp. 1656– 1667,2014.

[5] R. Pagh and F. Rodler, Cuckoo hashing, Proc. ESA, pp. 121–133, 2001.

[6]P. Indyk and R. Motwani, ―Approximate nearest neighbors: towardsremoving the curse of dimensionality, Proc. STOC, pp. 604–613, 1998.

[7]Y. Hua, H. Jiang, Y. Zhu, D. Feng, and L. Xu, SANE: Semantic-AwareNamespace in Ultra-large-scale File

Systems, IEEE Transactions on Paralleland Distributed Systems (TPDS), vol. 25, no. 5, pp. 1328–1338, 2014.

[8]SmartEye: Real-time and Efficient Cloud Image Sharing for DisasterEnvironments

[9]Storage Newsletter, 7% of consumer content in cloud storage in 2011,36% in 2016, 2012.

[10]Y. Ke, R. Sukthankar, and L. Huston, Efficient near-duplicate detectionand sub-image retrieval, Proc. ACM Multimedia, 2004.

[11]T. Ahonen, A. Hadid, and M. Pietikainen, Face description with localbinary patterns: Application to face recognition, IEEE Transactions onPattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037– 2041,2006.

[12]Y. Hua, H. Jiang, Y. Zhu, D. Feng, and L. Tian, SmartStore: A NewMetadata Organization Paradigm with Semantic-Awareness for NextGeneration File Systems, Proc. SC, 2009.

[13]D. Lowe, Distinctive image features from scale-invariant keypoints,International Journal of Computer Vision, vol. 60, no. 2, pp.

[14]Y. Hua, H. Jiang, and D. Feng, FAST: Near Real-time Searchable DataAnalytics for the Cloud, Proceedings of the International Conference forHigh Performance Computing, Networking, Storage and Analysis (SC), 2014.91–110, 2004.

[15]A. Gionis, P. Indyk, and R. Motwani, Similarity search in highdimensions via hashing, VLDB, pp. 518–529, 1999.

[16] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, Locality-sensitivehashing scheme based on p-stable distributions,Proc. Annual Sympo- siumon Computational Geometry, 2004.