FOUZIA SULTANA, et al, International Journal of Computers Electrical and Advanced Communication Engineering [IJCEACE]™
Volume 1, Issue 10, PP: 138 - 148, AUG - DEC' 2016.

SURVEY OF GLOBAL CONTENT FILTERING BY SIMILARITY SEARCH QUERY

FOUZIA SULTANA, SUMAYYA SAMREEN

Department of Computer Science & Engineering
Lords Institute of Engineering and Technology, Himayatsagar, Hyderabad-500 091.

Email-Id: fouzia.sultana114@gmail.com,sumayyasamreen01@gmail.com

ABSTRACT: Closeness joins are crucial operations with a wide extent of employments. In this paper, we consider the issue of vector closeness join size estimation (VSJ). It is a hypothesis of the in advance focused on set closeness join size estimation (SSJ) issue and can manage all the additionally interesting cases, for instance, TF-IDF vectors. One of the key challenges in resemblance join size estimation is that the join size can change definitely depending upon the data likeness edge. We propose a looking at based estimation those usages Locality Sensitive-Hashing (LSH). The proposed count LSH-SS uses a LSH document to engage powerful examining even at high edges. We differentiate the proposed methodology and sporadic analyzing and the best in class strategy for SSJ (conformed to VSJ) and show LSH-SS offers more correct evaluations all through the closeness edge degree and little change using genuine data sets.

Keywords: Term Frequency, Join Size Estimation, Locality Sensitive-Hashing, vector closeness.

1. INTRODUCTION

Given a likeness measure and a base comparability edge, a similarity join is to find all courses of action of articles whose closeness is not more little than quite far.

The article in a proportionality join is reliably a vector. For instance, a report can be tended to by a vector of words in the report, or a photograph can be tended to by a vector from its shading histogram. In this paper, we concentrate on the vector representation of things and study the running with issue.

Definition 1 (The VSJ Problem). Given a party of veritable respected vectors V = {v1 , ..., vn } and a state of imprisonment on a closeness measure sim, gage the measure of sets J = $|\{(u, v) : u, v \in V, sim(u, v) \geq \tau, u = v\}|$.

Similitude joins have a clearing degree of employments including close copy report affirmation and end, Endorsement to make advanced or printed adjustments of all or some fragment of this work for individual or classroom use is allowed without expense gave that duplicates are not made or scattered for advantage or business advantage and that duplicates bear this notice and the full reference on the primary page. To duplicate generally, to republish, to show on servers or on redistribute to records, requires earlier specific consent

and/or a charge. Articles from this volume were welcome to demonstrate their outcomes at The37th International Conference on Very Large Data Bases, August 29th – September third 2011, Seattle, Washington. Methodology of the VLDB Endowment, Vol. 4, No. 6

Demand refinement for web search for, coalition range, and information cleaning outlines [17, 2]. As necessities be, similarity joins have beginning late gotten much thought. Chaudhuri et al. perceived a likeness join operation as a primitive administrator in database structures [6]. To effectively interlace similarity join operations in database frameworks, it is crucial that we have attempted and genuine size estimation system for them. The request enhancer needs correct size estimations to pass on an upgraded ask for game-plan. Subsequently, in this paper, we concentrate on the size estimation of vector closeness joins.

In the composed work, the closeness join size estimation issue has been portrayed utilizing sets as takes after:

Definition 2 (The SSJ Problem). Given a get-together of real respected sets S = {s1 , ..., sn } and a most remote point τ on a similarity measure sim, gage the measure of sets J = |{(r, s) : r, s ∈ S, sim(r, s) ≥ τ, r = s}|.

Note that our organizing of closeness joins with vectors is more broad and can oversee more supportive applications. For example, while in the SSJ issue a record is basically a game-plan of words in the report, in the VSJ issue a document can be displayed with a vector of words with TF-IDF weights. It can besides manage multi set semantics with events. Truly, a large portion of the studies on resemblance joins first plan the issue with sets and after that extend it with TF-IDF weights, which is in sureness a vector comparability join.

The SSJ issue has been starting now considered by Lee et al. A prompt development of SSJ frameworks for the VSJ issue is to insert a vector into a set space. We change over a vector into a set by seeing an estimation as a section and repeating the portion the same number of times as the estimation respect, utilizing standard altering structures if qualities are not basic. In every practical sense, by the by, this embeddings can effectsly influence execution, exactness or required assets. Naturally, a set is a remarkable event of a parallel vector and is not more difficult to handle than a vector.

For instance, Bayardo et al. portray the vector comparability join issue and fuse excellent improvements that are conceivable precisely when vectors are twofold vectors (sets).

In our VSJ issue, we consider cosine likeness as the closeness measure sim since it has been satisfactorily utilized over two or three districts. Let u[i] mean the i-th estimation of vector u. Cosine comparability is portrayed as cos(u, v) = 2u•v/u v , where u•v = i u[i]•v[i] and u =i u[i] .

We concentrate on self-joins and talk about expansions to general takes an interest in Appendix B.2. One of the key difficulties in closeness join size estimation is that the join size can change basically subordinate upon. the information closeness limit. While the join size can be close n2 at low edges where n is the database size, it can be little at high edges. For instance, in the DBLP data set, the join selectivity is just around 0.00001 %at $\tau = 0.9$. While various reviewing counts have been proposed for the (equi-) join size estimation, their insurances misfire in such a high selectivity range, e.g. Actually, it is not utilitarian to apply essential unpredictable testing when the selectivity is high. This is risky since likeness confines some place around 0.5 and 0.9 are routinely used. Note that the join size in that degree may be adequately significant to influence question change in light of the extensive cross thing measure. Additionally, as saw in [13], join size bumbles multiply. That is, paying little heed to the way that the primary botches are pretty much nothing, their transitive impact can obliterate.

In this paper, we propose investigating based frameworks that attempt the Locality Sensitive Hashing (LSH) plan, which has been successfully associated in closeness looks across over various spaces. LSH creates hash tables such the similar things will likely be in the same bucket. Our key believed is that notwithstanding the way that examining a couple satisfying a high farthest point is astoundingly troublesome, it is for the most part easy to test the pair using the LSH arrangement since it clusters practically identical challenges together.

We exhibit that the proposed count LSH-SS gives awesome evaluations all through the comparability hres hold range with a test size of $\Omega(n)$ sets of vectors (i.e. $\Omega(n)$ tuples from each join association in an equi-join) with probabilistic certifications. The proposed course of action simply needs unimportant development to the current LSH record and as needs be is immediately suitable to various similarity look applications. As an outline, we make the going with responsibilities:

We show two standard techniques in Section 3. We consider self-assertive testing and modify Lattice Counting(LC) [14] which is proposed for the SSJ issue. We extend the LSH record to reinforce equivalence join size estimation in Section 4. We furthermore propose LSH-S which relies on upon a LSH limit examination.

We portray a stratified testing computation LSH-SS that misuse the LSH document in Section 5. We apply diverse testing philosophy for the two packages prompted by a LSH document: sets of vectors that are in the same can and those that are unquestionably not. We differentiate the proposed courses of action and subjective investigating and LC using certifiable data sets as a piece of Section 6. The exploratory results show that LSH-SS is the most correct with little contrast.

Time expect a critical part in any information space. Besides, it has been

focused on in a couple ranges like information recuperation, question answering, and outline. By using common information recuperation customer will get the bona fide scattering of documents of related scattering. Case I, Consider the catchphrase [Train Accident] for this there will be different related results. Thusly, if the customer is not sure about the date of occasion of event then the customer must go for the navigational endeavors that will prompts information over-troubling.

The outline shows that without saying the time variable for a class of request called time tricky inquiries the precise result for a specific inquiry can't be gotten. Generally the vital time interims ought to be determined unequivocally or certainly.

2. TIME SERIES ANALYSIS

A period plan involves a gathering of data in dynamic time orders and with uniform breaks. this time game plan examination is used to show the transient changes in the data. Time tricky inquiry auto completing system models the entire inquiry history by time course of action and guesses what's to come reputation as necessities be. Likewise, it uses most understood completion procedure in which it considers the totaled repeat of request over past journey logs for finding the future unmistakable quality of a request. Past work on auto realization is of two classes judicious auto fulfillment sentence fulfillment in perspective of lexion estimations of accumulations.

2.1TIME-BASEDQUERY CLASSIFICATION ALGORITHM

This computation contemplates the inquiry data arranged at different positions of the time turn from web question sign to recognize the case of the inquiry.

Occasional inquiries are one kind of time fragile questions, In 2011 M. Shokouhi[3] showed an approach for Detecting Seasonal inquiries. General request goes over in reliably like Christmas and Halloween. It is key for interest engines to successfully recognize the general inquiries and to driving force it fleetingly. A period game plan disintegration framework can be used for recognizing general inquiries. Moreover, these inquiries can be gathered by changing the inquiry repeat history into time-game plan model Recorded collections and over web diaries are enormous information resources, it is required to spare the chronicled setting of far reaching bits on the web, other than that the information can be used for the examination purposes too regarding re making a past periods. Getting to these information resources are incredibly troublesome and require more thought. In 2011 ZeynepPehlivan, Anne Doucet, Stephane Gancarski[11] brought a paper considering the getting to systems for web narratives. For getting to the web records more intricate request with time estimation is required. In this paper a piece based technique is used with the ultimate objective of information recuperation, pages are apparently segmented into semantic squares. The pieces are gained as the result of page

division. A model with common estimation is then used for recouping the web archive data. Time estimation can be procured from the web that is the common expression present in the substance of the webpage page or from the inquiries. Related manages the web recorded access and the square based are the course and full substance interest are proposed by web reporting exercises.

3. Global Regular Expression Print Tools

Searching for a case in a substance record is a basic errand. All front line content supervisors give convenience for searching substance for a case. Regardless, plan organizing moreover is important past substance modifying. Customers of UNIX found that they constantly use the ed charge g/re/p, which infers comprehensive mission for the general expression and print the organizing lines, where re is a general expression, that an alternate framework was made with the same name, grep. Fundamentally, grep transformed into an acronym for Global Regular Expression.

It is assumed that grep is without a doubt the most used instrument as a piece of venture insight. Observational studies show that item builds use grep extensively in their step by step bolster assignments [Singer and Lethbridge 1997]. Regardless of the way that it is all things considered recognized that, at any rate for vital typical expression planning, grep is definitely not hard to learn, easy to use, and its hindrances are unquestionably knew, there is a talk about among subjective experts for the bona fide purposes for grep's flourishing. More specifically, it is a non-piddling undertaking to pick up from the accomplishment of grep in laying out and making new program recognition gadgets.

In spite of the way that significant and viable, the primary UNIX grep [IEEE and Open Group 2003b] was assuredly not satisfactory to satisfy all needs. On occasions, customers found part and upgrades that could be intertwined into grep, for instance, the GNU developments found in GNU grep [GNU 2002], and on others, they found better approaches to manage illustration planning that in a general sense changed the behavior of grep, for instance, setting grep [Clarke and Cormack 1996], sorted out grep [Jaakkola moreover, Kilpeläinen 1996], nondeterministic reverse grep [Navarro 2001], and deduced grep [Wu and Manber 1992a]. This paper diagrams all the critical comprehensively valuable overall ordinary 2 expression print mechanical assemblies—varieties of grep or grep-like instruments. For each instrument, we present a brief audit of parts and give a use outline. Whatever is left of the paper is sorted out as takes after. Zone II shows a brief history of grep. Portion III gives an audit of various grep gadgets. Zone IV completes up the paper with our recognitions.

4. Nondeterministic Reverse Grep: nrgrep

Nondeterministic reverse grep [Navarro 2001] is the most up to date expansion to the grep instruments. It was produced to encourage looking regular dialect content. Nrgrep is the primary example coordinating apparatus that uses the bit-parallel reproduction of a nondeterministic addition machine [Navarro and Raffinot1998]. At the abnormal state, nrgrep gives comparable

usefulness as agrep. In any case, its execution is totally diverse. Its utilization of a solitary and uniform algorithmic idea, as restricted numerous particular calculations as in agrep and GNU grep, permits it to perform straightforward, refined, and surmised design coordinating with an effectiveness that corrupts easily as the unpredictability of the example increments. Ngrep additionally fuses a portion of the helpful alternatives found in GNU grep, for example, Nrgrep arranges designs into three classes: basic examples, broadened examples, and normal expressions. Basic examples permit the use of character classes in altered strings. Developed examples permit some character classes to be discretionary or dreary. Normal expressions are the most broad and take into account consideration of unfilled strings, connection, union, and reiteration. This characterization is utilized as a part of developing upgraded rendition of the calculation for every class. Further, it takes into consideration subpattern streamlining. The outcome is an instrument with practically identical execution to grep and agrep for short and correct examples and an observably predominant execution for long examples and for estimated designs.

One of the restriction of ngrep is settled size cradled handling of records. The span of the cushion is indicated ahead of time and the all every record must fit in the cradle for device to create precise results.

Case. To discover all events of "calculation" including toward the start of a sentence and take into consideration up to 3 erasure and substitution mistakes in the record readme.txt:nrgrep - k 3ds "[Aa]lgorithm" readme.txt The inspiration driving component coreference1 is to pick if unmistakable notification of formal individuals, spots or things insinuate the same genuine substance. A notification is an occasion of a name in a record, a site page, et cetera. Case in point, in two reports (e.g., news articles), two or more notice of the name James Henderson may exist and a component co reference count can reply if they really recognize the same certifiable person. The component co reference errand is trying basically in light of two general edges: how to discover setting information for each notification and how to utilize the association in a reasonable way. On one hand, we need to assemble setting information for those differentiation ent notice. We can assemble the association from the chronicles where the notification happen.

The Internet can be another hotspot for finding association information. Of course, it is genuinely vital to utilize the setting appropriately. There are distinctive conditions that can trick the substance co reference comes about. Name assortments, the use of shortened forms, and wrong spellings would all have the capacity to expect a section in the last results [Bilenko et al. 2003]. As well, the accumulated data may start from heterogeneous sources and may not be done. For instance, two news articles may depict various parts of James Henderson. One article may determine the

name and association while the other one can fuse his name, date of birth, email location, et cetera. Besides, may be clatters in the data gave.

For example, some date information is fused into the association for James Henderson moreover, it is managed as his date of birth; regardless, that date information could simply be the date of a party that this James Henderson went to. A component co reference computation ought to have the ability to oversee such issues and troubles. Component co reference in the Semantic Web [Berners-Lee et al. 2001] is used to perceive break even with mysticism events. In the Semantic Web, a reasoning is an unequivocal and formal specific of a conceptualization, formally depicting a space of talk. A cosmology involves a course of action of terms (classes) and the associations (class chains of significance likewise, predicates) between these terms. RDF is an outline based data model for depicting resources and their associations and it is a W3C proposal for addressing being developed in the Web2. Two resources are related by method for one or more predicates in the sort of triple. A triple, <s, p, o>, involves three areas: subject, predicate and challenge. The subject is an identifier (e.g., a URI) and the thing can either be an identifier or a demanding worth, for instance, strings, numbers, dates, et cetera. A URI that expect the subject position in one triple can be the thing in another; along these lines, the triples themselves shape a graph, the RDF graph. In a RDF diagram, a mysticism event is addressed by a URI; how-ever, phonetically specific URIs could truly address the same certified component.

For instance, a man can have distinctive URI identifiers in bibliographic databases for instance, DBLP [Ley 2002] and Cite Seer [Giles et al. 1998] however such URIs address the same individual; thusly they are co referent. In the Semantic Web, co referent cases are associated with each other with the owl:sameAs predicate and after that such co reference data can be further utilized by various parts of Semantic Web related examination, for example, Semantic Web based request answering, information coordination, etc. There has been different examination for interfacing reasoning cases in the Semantic Web. Associated Data3 [Bizer et al. 2009] is one of the principle attempts around there. Accord-ing to the latest statistics4, there are at this moment 207 datasets (from various spaces, e.g., media, topography, disseminations, et cetera.) in the Linked Open Data (LOD) Cloud with more than 28 billion triples and around 395 million associations transversely over different datasets. How-ever, one issue of these current owl:sameAs associations is that they were delivered with figurings that are not adequately correct. Starting late reported by Halpin et al. [2010], simply half (± 21%) of the owl:sameAs associations are correct. Along these lines, there rises the ought to have the ability to actually recognize splendid owl:sameAs joins between cosmology events from heterogeneous datasets.

In this paper, we display a novel substance co reference estimation to perceive co referent transcendentalism cases. With everything taken into account, given a few events of comparative classes, our calculation reprimands on the chance that they are co referent, i.e., suggest the same certifiable component, for instance, the same individual, dispersion, et cetera. In our count, for a given event, we find its neighborhood graph from the entire RDF outline through an augmentation strategy and we wind up having a course of action of routes starting from this case and conclusion on another center in the RDF chart. Each way is made out of a couple triples. Next, we figure the discrim-failure of each triple, checking its predicate. Such discriminability is then discounted as showed by the triple's detachment to the root center (the cosmology event). With such a partition based decreasing methodology and the triple discriminability, we figure the greatness of each path in the region chart (the setting) of a philosophy illustration. Appeared differently in relation to structures that solitary consolidate a subset of these parts, our proposed count finishes the best execution on four sorts of theory case from two unmistakable datasets. In addition, our structure beats best in class frameworks when associated with three benchmark datasets for cosmology case planning. At long last, we take a gander at the flexibility of our proposed structure and grasp one preselect particle strategy to improve its versatility.

Researchers have been wearing down component co reference and equivalent subjects for a long time. To deal with the name disambiguation issue, researchers have developed a grouping of string planning computations [Bilenko et al. 2003; Cohen et al. 2003], have endeavored to dis-vagueness tantamount names by abusing the resemblance of their associations [Pedersen et al. 2005], and have examined applying critical methods to recognize name equivalences in cutting edge libraries [Feitelson 2004]. A couple of examiners have been tackling substance co reference in free substance. Bagga and Baldwin [1998] use a vector space model to do cross-document substance co reference on individual notification in free substance. They first use an in-document co reference system to manufacture co reference chains inside each record. A particular chain contains name notification and pronouns that are co referent. By then, for cross-report co reference, they utilize all the critical sentences to a particular notification as setting. The relevant sentences are those where a notification or its in-report co referent notification happen. How-ever, this technique relies on upon the in-record component co reference structure to give extraordinary results in order to assemble superior to anything normal association information. Gooi and Allan [2004] use three models, the incremental/agglomerative vector space models and KL contrast, for component co reference on individual notification in free substance. They use a window size of 55 words concentrated on a notification to assemble its association

information in light of the way that their examinations exhibited that the best results were expert with this window size. Mann and Yarowsky [2003] use unsupervised grouping over a segment space to do co reference. They evacuate association information for each notification from site pages. The guideline change ence is that they endeavor to focus some more illustrative information from the website pages, for instance, narrative information, marriage, gatekeeper/adolescent associations hence on. Han et al. [2004] pass on two particular models, the Naive Bayes classifier and the SVM, to disambiguate maker names in references. Given a reference, their figuring predicts if it is made by some particular maker. They laid out a game plan of components to fit into the classifiers; nevertheless, such parts may not have any kind of effect todiverse spaces. Some outline based philosophies have been used additionally to disambiguate notice in social net-works [Bekkerman and McCallum 2005] and messages [Minkov et al. 2006]. Other than faultless free substance, Wikipedia and Encyclopedic have similarly been used to find association data [Bunescu and Pasca 2006; Cucerzan 2007]. Exceptional sorts of Wikipedia pages (e.g., disambiguation pages) and the embedded hyperlinks have been used for ex-ploiting setting information. Named component affirmation [David and Satoshi 2007] can be managed as a preprocessing wander for substance co reference. It sees different sorts of notification, for instance, individual, affiliation, et cetera. This technique is out of the degree of this paper. Word sense disambiguation (WSD) and duplicate record area in databases are two immovably related topics to substance co reference. A word can have various ramifications while the endeavor of WSD is to pick the most reasonable one based upon the word's setting [Yarowsky 1995; Zhang and Heflin 2010], for instance, a touch of free substance. Copy record disclosure is to recognize duplicate tuples and remove redundancies from databases [Elmagarmid et al. 2007]. Differing database records can give the same information however are unmistakable in their representations. Case in point, various records can address a man's name in a surprising route, in the sorts of full name or first basic other than family name. Dong et al. [2005] proposed a substance co reference estimation that misuse the associations between different substances to upgrade system execution. They all things considered resolution substances of various sorts by method for social affirmation spread in dependence outlines. They associated the estimation to various genuine datasets and demonstrated its practicality. Kalashnikov and Mehrotra [2006] proposed RELDC (Social based Data Cleaning) to recognize co referent substances by separating component connections. The components and their associations are seen as a graph where edges address the associations between components.

Conclusion:-

It is fundamental to consider the time estimation for looking for over an unlimited assembling of records. Since Searchers often

don't know watchful time or date of an event happened. Planning the transient information alongside the subject equivalence overhauls the information recuperation. We have shown different outlines and circumstances where common information can be amazingly useful for gaining the critical file related with a customer question.

REFERENCES

[1] R. Baeza-Yates and G. Gonnet. A fast algorithm on average for all-against-all sequence matching. In Proceedings of String Processing and Information Retrieval Symposium (SPIRE'99), pages 16–23, 1999.

[2] T. Bozkaya and Z. M. Ozsoyoglu. Distance based indexing for high dimensional metric spaces. In Proceedings of the 1997 ACM SIGMOD Conference on Management of Data, pages 357–368,1997.

[3] S. Brin. Near neighbor search in large metric spaces. In Proceedings of the 21st International Conference on Very Large Databases (VLDB'95), pages 574–584, 1995.

[4] A. Cobbs. Fast approximate matching using suffix trees. In Combinatorial Pattern Matching, 6th Annual Symposium (CPM'95), pages 41–54, 1995.

[5] W. Cohen. Integration of heterogeneous databases without common domains using queries based on textual similarity. In Proceedings of the 1998 ACM SIGMOD Conference on Management of Data, pages 201–212, 1998.

[6] D. J. DeWitt, J. F. Naughton, and D. A. Schneider. An evaluation of non-equijoin algorithms.In Proceedings of the 17th International Conference on Very Large Databases (VLDB'91), pages 443–452, 1991.

[7] D. A. Grossman, O. Frieder, D. O. Holmes, and D. C. Roberts. Integrating structured data and text: A relational approach. In Journal of the American Society for Information Science (JASIS), 48(2):122– 132, 1997.

[8] C. Lundquist, O. Frieder, D. O. Holmes, and D. A. Grossman.A parallel relational database management system approach to relevance feedback in information retrieval. In Journal of the American Society for Information Science (JASIS), 50(5):413–426, 1999.

[9] U. Manber and S. Wu. GLIMPSE: A tool to search through entire file systems. In Proceedings of USENIX Winter 1994 Technical Conference, pages 23–32, 1994.

[10] G. Navarro. A guided tour to approximate string matching.To appear in ACM Computing Surveys, 2001.

[11] W. Dakka, L. Gravano, and P.G. Ipeirotis, "Answering General Time-Sensitive Queries," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '12), pp. 1437- 1438, 2012.

[12] McMillanShokouchi,Kira Radinsky, "Time Sensitive Query Auto completion" Proc .35th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval,2012.

[13] M. Shokouhi. Detecting seasonal queries by time-series analysis.In Proc.

SIGIR, pages 1171{1172, Beijing, China, 2011.

[14] A. Z. Broder. On the Resemblance and Containment of Documents. In Proc. SEQUENCES, pages 21{29,1997.

[15] M. S. Charikar. Similarity Estimation Techniques
from Rounding Algorithms. In Proc. STOC, pages
380{388, 2002.

[16] S. Chaudhuri, V. Ganti, and R. Kaushik. A Primitive Operator for Similarity Joins in Data Cleaning. In Proc. ICDE, pages 5{16, 2006.

[17] D. A. Grossman, O. Frieder, D. O. Holmes, and D. C. Roberts. Integrating structured data and text: A relational approach. In Journal of the American Society for Information Science (JASIS), 48(2):122– 132, 1997.

[18] C. Lundquist, O. Frieder, D. O. Holmes, and D. A. Grossman.A parallel relational database management system approach to relevance feedback in information retrieval. In Journal of the American Society for Information Science (JASIS), 50(5):413–426, 1999.

[19] U. Manber and S. Wu. GLIMPSE: A tool to search through entire file systems. In Proceedings of USENIX Winter 1994 Technical Conference, pages 23–32, 1994.

[20] G. Navarro. A guided tour to approximate string matching.To appear in ACM Computing Surveys, 2001.

ABOUT AUTHORS:

Fouzia Sultana is currently working as an Assistant Professor in Computer Science and Engineering Department , Lords Institute of Engineering & Technology, Hyderabad-500091.She has knowledge in Information retrieval systems, Data mining and Databases. Her research includes Survey of global content filtering by similarity search query.

Sumayya Samreen is currently pursuing her M.Tech(CSE) in Computer science and engineering Department,Lords institute of engineering and technology.