UGC Approved Journal

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2017

Nearest Keyword Set Search in Multi- Dimensional Dataset

P. Dhivya[1], S. Priya[2]

Research Scholar, P.G. and Research Department of Computer Science, Swami Vivekananda Arts and Science College, Villupuram, Tamil Nadu, India[1]

Assistant Professor, P.G. and Research Department of Computer Science, Swami Vivekananda Arts and Science College, Villupuram, Tamil Nadu, India[2]

Abstract: A spatial database manages multidimensional objects provides fast access to those objects based on different selection criteria. The importance of spatial databases is reflected by the convenience of modeling entities of reality in a geometric manner. For example, locations of restaurants, hotels, hospitals and so on are often represented as points in a map, while larger extents such as parks, lakes, and landscapes often as a combination of rectangles. Many functionalities of a spatial database are useful in various ways in specific contexts. For instance, in a geography information system, range search can be deployed to find all restaurants in a certain area, while nearest neighbor retrieval can discover the restaurant closest to a given address. Conventional spatial queries, such as range search and nearest neighbor retrieval, involve only conditions on objects' geometric properties. Today, many modern applications call for novel forms of queries that aim to find objects satisfying both a spatial predicate, and a predicate on their associated texts. For example, instead of considering all the restaurants, a nearest neighbor query would instead ask for the restaurant that is the closest among those whose menus contain "steak, spaghetti, brandy" all at the same time. Currently, the best solution to such queries is based on the IR2-tree, which, as shown in this paper, has a few deficiencies that seriously impact its efficiency. Motivated by this, this work develop a new access method called the spatial inverted index that extends the conventional inverted index to cope with multidimensional data and comes with algorithms that can answer nearest neighbor queries with keywords in real time.

Keywords: Spatial database, geometric, spatial queries, IR2-tree, multidimensional data.

I. INTRODUCTION

Objects (e.g., images, chemical compounds, documents, or experts in collaborative networks) are often characterized by a collection of relevant features, and are commonly represented as points in a multi- dimensional feature space. For example, images are represented using color feature vectors, and usually have descriptive text information (e.g., tags or keywords) associated with them. In this paper, here consider multi-dimensional datasets where each data point has a set of keywords. The presence of keywords in feature space allows for the development of new tools to query and explore these multidimensional datasets.
Nearest Keyword set study on content rich different types of data sets. The NKS analysis is an arrangement of catchphrases in vision of theme. Also, the arrangement of the question consolidates ''K'' type of catchphrases as a group and concentrates each and every set which possess data based bunches along with structures in which bunches of multi-dimensional section is created. Each point is labeled with an arrangement of clusters.

An increasing number of uses need the productive execution of nearest neighbor (NN) queries thankful by the properties of the spatial objects. Because of the importance of keyword hunt, especially on the Internet, a considerable lot of these applications permit the client to give a rundown of keywords that the spatial objects (from this time forward alluded to just as objects) ought to contain, in their portrayal or other quality. For instance, online business index permit clients to point to an address and an arrangement of keywords, and return organizations whose portrayal contains these keywords, requested by their separation to the predetermined address area. As another case, land sites permit clients to look for properties with particular keywords in their depiction and rank them as per their separation from a predefined area. We call such queries spatial keyword queries. A spatial keyword query comprises of a query zone and an arrangement of keywords.

The response is a derelict of objects ranked by blend of their division to the query range and the substance of their content depiction to the query keywords. A basic yet well-known variation, which is utilized as a part of our running case, is the separation first spatial keyword query, where objects are ranked by separation and keywords are connected

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2017

as a conjunctive channel to dispose of objects that don't contain them. Which is our running illustration, shows a dataset of imaginary inns with their spatial directions and an arrangement of distinct traits (name, courtesies)? A case of a spatial keyword query is "determine the nearest accommodation to point that enclose keywords web and pool". The top consequence of this query is the inn protest.
NKS queries are useful for many applications, such as photo-sharing in social networks, graph pattern search, geolocation search in GIS systems, and so on.

1.1 Multi-Dimensional Data Sets
One this page you will find some "real world" multi-dimensional data sets. For right now there are two Tiger data sets, extracted from the US Bureau of Census TIGER database by some unknown person (if you know the person please send me email so I can reference appropriately), and a few CFD data sets. This work was partially supported by NSF grant number 9610270.
Only the small data sets are given in ascii format, the rest in binary. Included is a simple (and not very elegant) c program to convert from the binary format to an ascii format. There is just enough documentation at the top to show how to use it.

CFD Data Sets
If you use any of these CFD data sets, please reference this web page and the creator of the data:

2D Point Data
These are vertex data sets from various Computation Fluid Dynamics models. The data sets are for a 2-dimensional problem. A system of equations is used to model the air flows over and around aero-space vehicles. The data sets are for a cross section of a Boeing 737 wing with flaps out in landing configuration at MACH 0.2. The data space consists of a collection of points (nodes) of varying density. Nodes are dense in areas of great change in the solution of the equations and sparse in areas of little change. The location of the points in the data set is HIGHLY skewed.
The format of the data sets is (xminyminxmaxymax) which delimit the lower left hand corner and upper right hand corner of the rectangle. Size these data sets are point data, xmin = xmax and ymin = ymax. The points range from (-20,-20) to (20,20), but I have normalized the data sets to the unit square to facilitate experimental studies that include other data sets. For the 5088 Node Set the original ascii format (not normalized), ascii normalized version, and binary normalized version are all included. For the larger data sets only the normalized binary version is included. If for some reason you need the non-normalized versions contact me. The binary is normal IEEE binary, four floats per point.
The three data sets differ mostly in the number of points. To get a good idea what the data looks like download and preview the postscript picture below for the 5088 Vertex Data Set. The complete normalized shows the whole data set, the blow up is just the region around the centroid.

2D Triangle Data
These data sets are a Delaunay triangulation (with a few flips) of the point data sets above. See the postscript picture included below. These data sets are originally in a space efficient format. Consider first the 9759 Triangle Set below. The original ascii file shows the format. Each line contains the triangle number followed by the point numbers of the three corners of the triangle. To get the coordinates of these triangle vertices, just look up the point in the 5088 original ascii point data set (above). Each triangle of the data set is then bounded by a rectangle as shown in the Non-normalized rectangle-bounded set below. Next, the set of rectangles is normalized to the unit square. Finally the normalized rectangle set is converted to binary. For the larger data sets only the normalized binary data sets are included.

Tiger Data Sets
These are line segment data contains the road maps of Long Beach and Montgomery Counties. If you use these, please reference as: "Extracted from the US Bureau of Census TIGER database".

II. REVIEW OF LITERAURE

Zhishenget. al [1], proposed a geographic query that is made out of query keywords and a location, a geographic search motor recovers documents that are the most textually and spatially pertinent to the query keywords and the location, separately, and ranks the recovered documents as indicated by their joint textual and spatial relevance's to the query. The lack of an effective file that can all the while handle both the textual and spatial parts of the documents makes existing geographic search motors wasteful in noting geographic inquiries. In this paper, we propose an effective record, called IR-tree, that together with a top-k document search algorithm encourages four noteworthy tasks in document searches, to be specific, 1) spatial filtering, 2) textual filtering, 3) relevance computation, and 4) document

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2017

ranking in a completely coordinated way. What's more, IR-tree permits searches to embrace diverse weights on textual and spatial relevance of documents at the runtime and in this way cooks for a wide assortment of utilizations. An arrangement of thorough examinations over an extensive variety of situations has been directed and the trial comes about show that IR-tree beats the cutting edge approaches for geographic document searches.

Christian et. al [2], planned the location-aware keyword query proceeds ranked objects that are almost a query position and that have printed portrayals that match query keywords. This query happens logically in many sorts of versatile and conservative web administrations and applications, e.g., Yellow Pages and Maps administrations. Past work considers the potential consequences of such a query as being autonomous when ranking them. Notwithstanding, a pertinent outcome question with adjacent objects that are likewise applicable to the query is likely to be ideal over an important protest without significant close-by objects.

Christian et. al [3], proposed customary Internet is securing a geo-spatial dimension. Web reports are being geo-labeled, and geo referenced protests, for example, purposes of intrigue are being connected with engaging content records. The subsequent combination of geo-location and reports empowers another kind of top-k query that takes into record both location proximity and content significance. To our knowledge, just local systems exist that is fit for registering a general web information recovery query while additionally taking location into record. This paper proposes another ordering framework for location mindful top-k content recovery. The framework influences the upset document for content recovery and the R-tree for spatial proximity querying.

Chakrabartiet. al[4], refereed the Clients frequently search spatial databases like yellow page information utilizing catchphrases to and organizations close to their flow location. Such searches are progressively being performed from cell phones. Writing the whole question is bulky and inclined to mistakes, particularly from cell phones. We address this subject by presenting type in front search usefulness on spatial databases.

Like watchword explore on spatial information, type-ahead search should be location-aware, i.e., with each letter being typed, it needs to revisit spatial items whose names (or portrayals) are considerable consummations of the question string typed in this way, and which rank most elevated as far as closeness to the client's location and other static scores. Existing answers for type-ahead search can't be utilized specifically as they are not location-aware. We demonstrate that a straight-forward mix of existing systems for performing type-ahead search with those for performing nearness search perform inadequately.

Zhanget. al [5], proposed Mapping concoction are rising Web 2.0 applications in which information objects, for example, sites, photographs and recordings from various sources are combined and set apart in a guide utilizing APIs that are discharged by web based mapping arrangements, for example, Google and Yahoo Maps. These objects are normally connected with an arrangement of labels catching the installed semantic and an arrangement of coordinates showing their geographical locations.

Z. Li et. al[6], Aggregate nearest keyword search in spatial databases,‖ in Asia-Pacific Web Conference, 2010. Keyword search on relational databases is useful and popular for many users without technical background. Recently, aggregate keyword search on relational databases was proposed and has attracted interest. However, two important problems still remain. First, aggregate keyword search can be very costly on large relational databases, partly due to the lack of effcient indexes. Second, the top-k answers to an aggregate keyword query has not been addressed systematically, including both the ranking model and the efficient evaluation methods.

De Felipe et. al[7], Many applications require finding objects closest to a specified location that contains a set of keywords. For example, online yellow pages allow users to specify an address and a set of keywords. In return, the user obtains a list of businesses whose description contains these keywords, ordered by their distance from the specified address. The problems of nearest neighbor search on spatial data and keyword search on text data have been extensively studied separately. However, to the best of our knowledge there is no efficient method to answer spatial keyword queries, that is, queries that specify both a location and a set of keywords.

M. Dataret. al[8], presented a novel Locality-Sensitive Hashing scheme for the Approximate Nearest Neighbor Problem under lp norm, based on stable distributions. Our scheme improves the running time of the earlier algorithm for the case of the l2 norm. It also yields the first known provably efficient approximate NN algorithm for the case $p < 1$. We also show that the algorithm finds the exact near neighbor in O(log n) time for data satisfying certainbounded growth‖ condition. Unlike earlier schemes, our LSH scheme works directly on points in the Euclidean space without embeddings.

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2017

Felipe et.al [9], present an efficient method to answer top-k special keyword queries. To do so, we introduce an indexing structure called IR2-Tree (Information Retrieval R-Tree) which combines an R-Tree with superimposed text signatures. Author presents algorithms that construct and maintain an IR2-Tree and use it to answer top-k spatial keyword queries. its show superior performance and excellent scalability.IR2Tree to rank object from spatial dataset based on combination of their distances to the query location and relevance of their text description to the query keyword. Topk spatial keyword queries which is based on tight integration of data structure and algorithm used in special database search and information retrieval R-Tree(IR2-Tree)which is structure based on the R-Tree at query time and incremental algorithm is employed that uses IR2-Tree which is structure based on the R-Tree at query time and incremental algorithms. is employed that uses IR2-Tree which is structure based on the R-Tree at query time and incremental algorithm.

M. Kleinberg et. al[10], Develop new approach to the nearest-neighbor problem, based on a method for combining randomly chosen one dimensional projections of the underlying point set. Two algorithms are introduce in this first for finding epsilon approximate nearest neighbors and second epsilon approximate nearest-neighbor algorithm with near-linear storage and query time improves asymptotically linear search in all dimensions.

Aristides Gioniset. al[11], examine a novel scheme for approximate similarity search based on hashing. The basic idea is to hash the points from the database so as to ensure that the probability of collision is much higher for objects that are close to each other than for those that are far apart. The method gives significant improvement in running time over other methods for searching in high dimensional spaces based on hierarchical tree de-composition. This scheme scales well even for relatively large number of dimensions(more than 50). Previous technique[6] solve this problem efficiently only for the approximate case Accurate and efficient Near neighbor Search in High Dimensional Spaces. In this are designs to solve r-near neighbor queries for a fixed query range or for set of query ranges with probabilistic guarantees .and then extend for nearest neighbor queries.

III. PROPOSED METHODOLOGY

Spatial queries with keywords have not been extensively explored. In the past years, the community has sparked enthusiasm in studying keyword search in relational databases. It is until recently that attention was diverted to multidimensional data. The best method to date for nearest neighbor search with keywords is due to Felipe et al. They nicely integrate two well-known concepts: R-tree, a popular spatial index, and signature file, an effective method for keyword-based document retrieval. By doing so they develop a structure called the IR2 -tree, which has the strengths of both R-trees and signature files. Like R-trees, the IR2 - tree preserves objects' spatial proximity, which is the key to solving spatial queries efficiently. On the other hand, like signature files, the IR2 -tree is able to filter a considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined.

In this work, here design a variant of inverted index that is optimized for multidimensional points, and is thus named the spatial inverted index (SI-index). This access method successfully incorporates point coordinates into a conventional inverted index with small extra space, owing to a delicate compact storage scheme. Meanwhile, an SI-index preserves the spatial locality of data points, and comes with an R-tree built on every inverted list at little space overhead. As a result, it offers two competing ways for query processing. We can (sequentially) merge multiple lists very much like merging traditional inverted lists by ids. Alternatively, we can also leverage the R-trees to browse the points of all relevant lists in ascending order of their distances to the query point. As demonstrated by experiments, the SI-index significantly outperforms the IR2 -tree in query efficiency, often by a factor of orders of magnitude.

This work addresses a novel clustering algorithm to discover the latent semantics in a text corpus from a fuzzy linguistic perspective. Besides the applicability in text domains, it can be extended to the applications, such as data mining, bioinformatics, content-based or collaborative information filtering, and so forth.Web documents can constitute several latent semantic topics equipped with numerical coefficients (fuzzy linguistic coefficients) that indicate the significance levels of these inherent situations. A collection of documents and its corresponding fuzzy linguistic topological space L are two finite and discrete topological spaces, where L = {C1, C2, · · · ,Cn} and Ci denotes a semantic topological category. A discrete topological category is composed of all discrete features, that is, attribute-value pairs. The features in a document are extracted by using semi-supervised learning schemes called named entities. Named entity recognition (NER) can identify one item from a set of features that have similar attributes, i.e., named categories.Examples of named categories are person, affiliations, location, and so on. Consider the polyse my like the term "jaguar" can be classified as "animal," "vehicle," and so forth. If the term "jaguar" is associated with the items, such as "cat,""tiger," and "feline," the term "jaguar" is more possible to be classified into the named category "animal."

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2017

Fuzzy linguistic coefficient is given to measure the possibilities of a term belonging to every category where the term is associated with other co-occurring terms.

The general framework of our clustering method consists of two phases. The first phase, feature extraction, is to extract key named entities from a collection of "indexed" documents; the second phrase, fuzzy clustering, is to determine relations between features and identify their linguistic categories.The kernel of the first phrase is to identify the key features and their named categories. In order to identify features in documents, we deployed the named entity recognition method.From a given sentence, NER method first finds out the segmented entities composed of a sequence of words, and then classifies the entities by a type or named category, such as person, organization, location, and so on. This work considers only noun entities, especially some representative entities. Therefore, discriminative linear chain conditional random field (CRF) was used to choose the particular features in the corpus. A CRF is a simple framework for labeling and segmenting data that models a conditional distribution P(z|x) by selecting the label sequences to label a novel observation sequence x with an associated undirected graph structure that obeys the Markov property. When conditioned on the observations that are given in a particular observation sequence, the CRF defines a single log-linear distribution over the labeled sequence. CRF model does not need to explicitly present the dependencies of input variables x affording the use of rich and global features of the input, thus allows relaxation of the strong independent assumptions made by HMMs.

IV. RESULTS AND DISCUSSION

The experiment was carried out using two clustering models, namely; K-means and Fuzzy C-Means clustering algorithms. This is in view to finding out which of the cluster best suits the document URLs in terms of classifying the pre-processed data, trained data, testing, and making prediction using the model obtained from the training process. The detailed procedure of the experimentation is as follows:

Datasets information
The web document URLs datasets from the UCI Machine Learning Repository is used, Log of anonymous users of www.microsoft.com; predict areas of the web site a user visited based on data on other areas the user visited. The dataset available at https://archive.ics.uci.edu/ml/machine-learning-databases/anonymous/anonymous-msweb.data

Dataset Title: "Anonymous web data from www.microsoft.com".

Dataset URL: https://archive.ics.uci.edu/ml/machine-learning-databases/ anonymous/ anonymous - msweb.data

The data records the use of www.microsoft.com by 38000 anonymous, randomly-selected users. For each user, the data lists all the areas of the web site (Vroots) that user visited in a one week timeframe.

Users are identified only by a sequential number, for example, User #14988, User #14989, etc. The file contains no personally identifiable information. The 294 Vroots are identified by their title (e.g. "NetShow for PowerPoint") and URL (e.g. "/stream"). The data comes from one week in February, 1998.

Dataset format:

The data is in an ASCII-based sparse-data format called "DST". Each line of the data file starts with a letter which tells the line's type. The three line types of interest are:

-- Attribute lines:

For example, 'A,1277,1,"NetShow for PowerPoint","/stream"'

Where:

'A' marks this as an attribute line,
'1277' is the attribute ID number for an area of the website
(called a Vroot),
'1' may be ignored,
'"NetShow for PowerPoint"' is the title of the Vroot,
'"/stream"' is the URL relative to "http://www.microsoft.com"

-- Case and Vote Lines:

For each user, there is a case line followed by zero or more vote lines.
For example:

C,"10164",10164
V,1123,1
V,1009,1
V,1052,1

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2017

Where:
'C' marks this as a case line,
'10164' is the case ID number of a user,
'V' marks the vote lines for this case,
'1123', 1009', 1052' are the attributes ID's of Vroots that a user visited.
'1' may be ignored.

Number of Instances:
-- Training: 32711
-- Testing: 5000
Each instance represents an anonymous, randomly selected user of the web site.

Number of Attributes: 294

Attribute Information:
Each attribute is an area ("vroot") of the www.microsoft.com web site.
The datasets record which Vroots each user visited in a one-week timeframe inFeburary 1998.

PROPOSED WORK

The proposed experimental work on K-means and Fuzzy C-means obtained interesting results. While some past performance evaluation experiments have been performed on performance evaluation of clustering algorithms, the main aim of these experiments have been to compare performance of flat and hierarchical clustering algorithms. This experimental work in contrast tried to evaluate performance of different flat clustering algorithms with use of different representation and feature selection schemes. The work attempted do an extensive evaluation of three important flat clustering algorithms (K-means and Fuzzy C-Means) with a number of variations of representation schemes and feature selections (with or without stop words and with or without stemming). The preliminary results obtained present some commonly speculated and few relatively less known findings. Both the representation scheme used and the feature vector selected have affected the quality of the clustering result. Moreover, the magnitude of their effect varies with different algorithms. .The data is very sparse, so vroot visits are explicit; nonvisits are implicit (missing).

Training set i.e. set of URLs is given input to this problem. The URLs are partitioned into group of similar pages (called as dup-cluster) from one or more domain. URL-based de-duping methods, strategy is to learn, by mining these dup-clusters, rules that transform duplicate URLs to the same canonical form.Example of URLs to be de-duplicated and possible canonical forms. URLs of a same dub-cluster point to the same or similar content. URLs from different dub-clusters likely correspond to different content. Thu, in this example, whereas contents of u0 and u2 are the same, contents of u0 and u5 are different.

Table 4.1 Dub-cluster

Dub-cluster	URL
C_1	u_0=http://britney.com.br/?id=5 u_1=http://britney.com.br/index.php?id=5 u_2=http://Britney.com.br/?id=5 u_3= www.britney.com.br/?id=5 n_1=www.britney.com.br/index.php?id=5
C_2	u_4=http://britney.com.br/?id=7 u_5=http://Britney.com.br/index.php?id=7 n_1=www.britney.com.br/index.php?id=5

In above table, U = {u1, u2, u3, u4, u5} is partitioned in dup-clusters C1 and C2. The canonical form of the URLs in C_1 and C_2 are given by n_1 and n_2, respectively. This process, called as URL normalization, identifies, at crawling time, whether two or more URLs without fetching their contents. As crawlers have resources constraints, the best methods are those that achieve larger reductions with smaller false positive rates using the minimum number of normalization rules. A normalization rule is a description of the conditions and operations necessary to transform a URL into a canonical form.

UGC Approved Journal
IJIREEICE
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526
International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2017

The multiple sequences alignment of sequences can be the taken as a natural generalization of the above method .This method is taken into consideration when sequences greater than two number of sequences are present in the training dataset, In this gaps are inserted at arbitrary position in any number of sequences to be aligned, so the result to get have the same size l. the sequences are arrange in k lines and l column. So that spaces of sequences can be occurred in single column. The Multiple Sequences alignment method is known as NP-hard problem there are different solution have been invented to find a heuristic solution and to overcome this problem. So here use the method called as Progressive Alignment to align cluster of duplicate URL. In dupcluster there are multiple URL are present .in this method perform the alignment from already selected sequences and then after that new sequences is aligned with pervious alignment, from that to get the final multiple sequences alignment. This process is repeated until get all sequences been aligned.
The progressive alignment uses the method which is known as greedy policy, in that once gaps is given, which is not removed for a sequences alignment. So gaps are preserved until the final solution .the most similar sequences are selected due to error rate in alignment at every stage seen to be decrease and increase if most similar sequences are selected. It determine the best order of sequences for the alignment .Mostly similar Sequences are alignment first and at end most different Sequences due to this reduces the error which introduced by this heuristic solution.

The step of k-means clustering is the beginning to determine number of cluster and assume the centroid or center of these clusters. Here can take any random number of urls in dataset as the initial centroids or the first objects in sequence can also serve as the initial centroids.

Web document (Multidimensional dataset) URLs clustring using K-means and Fuzzy C-means

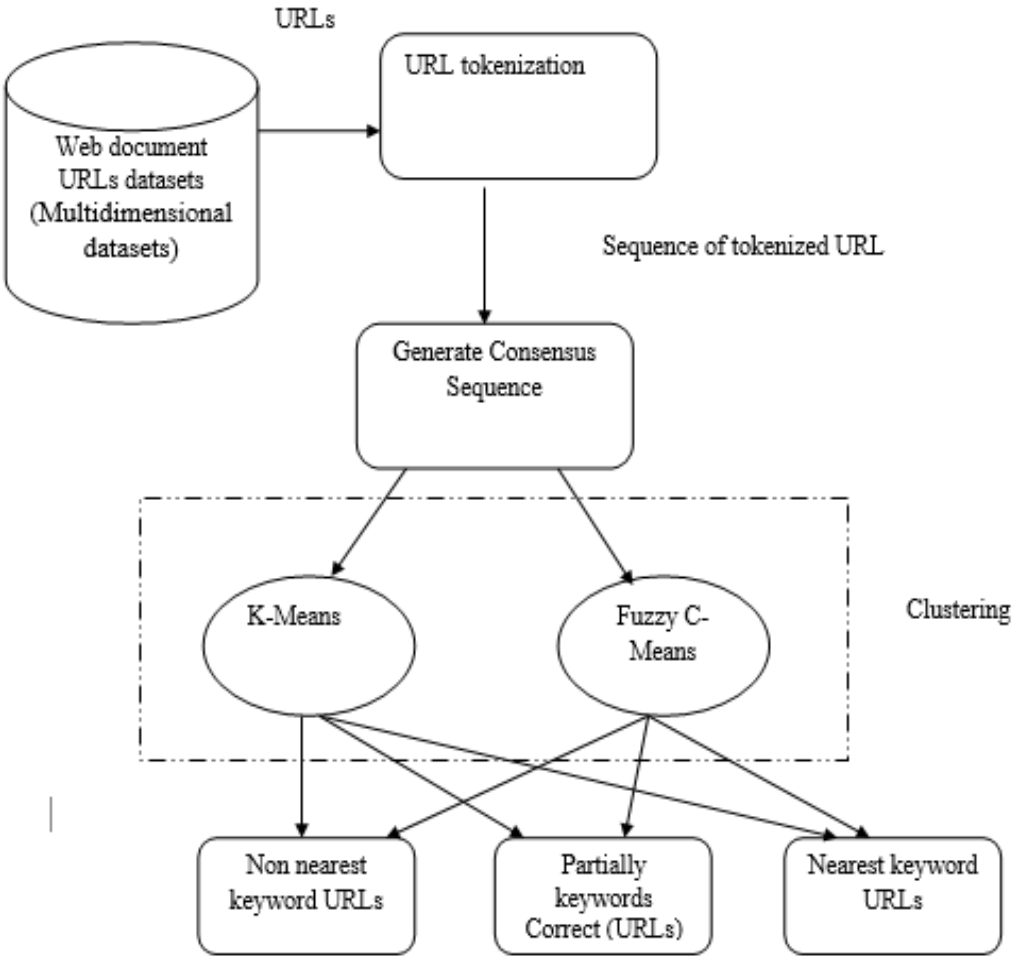

Figure 4.2 Framework of the Proposed Methodology

Time Comparision for K-Means and Fuzzy C-Means Algorithm
Fig 4.3 represents the Time chart of the resources with respect to expected completion time of tasks.

UGC Approved Journal

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2017

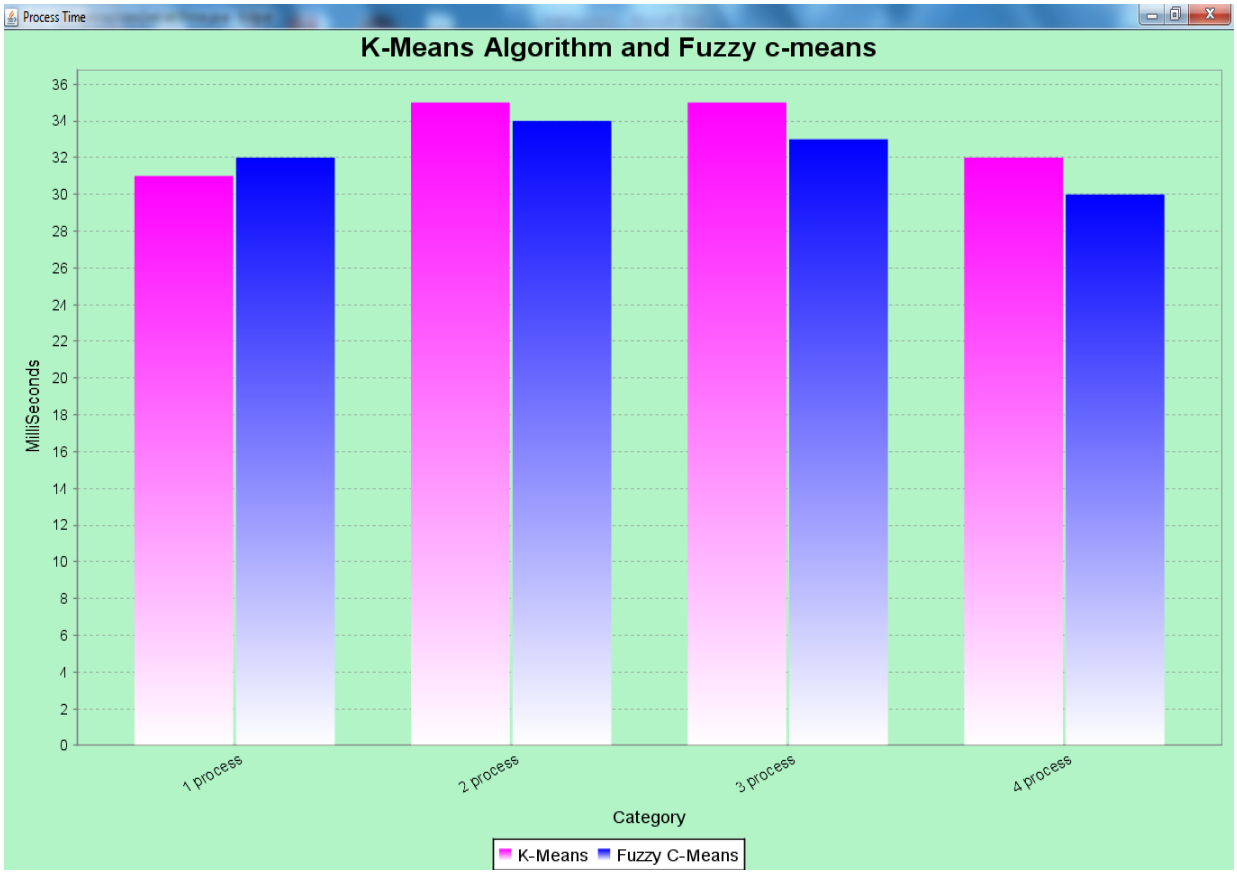

Fig 4.3 The graph for Comparison of K-Means and Fuzzy C-Means

Table 4.1 The table for Comparison of K-Means and Fuzzy C-Means

Illustration No	No of Process	Time in Milli Seconds	
		K-means	Fuzzy c-means
1	1	31	32
2	2	34	33
3	3	34	33
4	4	32	31

Table 4.3 shows that the minimum completion time of Fuzzy C-Means (FCM) are less when compared to K-Means. FCM produces close results to K-Means clustering though it evolves more fuzzy measures calculations in the algorithm. In fact, FCM clustering which constitute the oldest component of software computing are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster. They have been mainly used for discovering association rules and functional dependencies as well as image retrieval. So, overall work is that Fuzzy C-Means algorithm seems to be superior to K-Means algorithm.

V. CONCLUSION

The proposed approach used K-means and FCM algorithms for clustering the web document URLs in Multidimensional datasets. This approach keeps the related nearest keyword in URls in the same cluster. A single term is not able to identify a latent concept in a Multidimensional datasets, for instance, the term “Network” associated with the term “Computer,”“Traffic,” or “Neural” denotes different concepts. A group of solid co-occurring named entities can clearly define a CONCEPT. The semantic hierarchy generated from frequently co-occurring named entities of a given collection of Multidimensional datasets, form a simplifies complex. The complex can be decomposed into connected components at various levels (in various levels of skeletons). Here believe each such connected component properly identify a concept in a collection of Multidimensional datasets.

UGC Approved Journal

IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2017

To identify and discriminate the correct topics in a collection of Multidimensional datasets, the combinations of features and their co-occurring relationships are the clue, and the possibilities display how significant they will be. All features in Multidimensional datasets compose a topologically probabilistic space, more specifically simplical complex associated with probabilistic measures to denote the underlying structure. The complex can be geographically decomposed into inseparable components at various levels (in various levels of skeletons) that each component properly corresponds to topics in a collection of documents. Of course, the topics that a component induced are either topologically distinguishable, or perfectly included in other induced topics.

Future Direction

Here can effectively discover such a maximal fuzzy simplexes and use them to cluster the collection of Multidimensional datasets URLs. Based on the web site and the experiment, here find that dub-cluster is a very good way to organize the unstructured and semi structured data into several semantic topics. It also illustrates that geometric complexes are an effective model for automatic web documents clustering. Future work would focus on improving the cluster sets by semantic based clustering and ranking the documents in each cluster using topic based modeling.

REFERENCES

[1] Li, Zhisheng, Ken CK Lee, BaihuaZheng, Wang-Chien Lee, Dik Lee, and Xufa Wang. "Ir-tree: An efficient index for geographic document search." IEEE Transactions on Knowledge and Data Engineering 23, no. 4 (2011): 585-599.
[2] Cao, Xin, Gao Cong, and Christian S. Jensen. "Retrieving top-k prestige-based relevant spatial web objects." Proceedings of the VLDB Endowment 3, no. 1-2 (2010): 373-384.
[3] Cong, Gao, Christian S. Jensen, and Dingming Wu. "Efficient retrieval of the top-k most relevant spatial web objects." Proceedings of the VLDB Endowment 2, no. 1 (2009): 337-348.
[4] Basu Roy, Senjuti, and KaushikChakrabarti. "Location-aware type ahead search on spatial databases: semantics and efficiency." In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pp. 361-372. ACM, 2011.
[5] Zhang, Dongxiang, Beng Chin Ooi, and Anthony KH Tung. "Locating mapped resources in web 2.0." In Data Engineering (ICDE), 2010 IEEE 26th International Conference on, pp. 521-532. IEEE, 2010.
[6] Ashlesh S. Patole; ShripadraoBiradar " A Survey on Best Keyword Cover Search " IJIRCCE Vol. 3, Issue 11, November 2015 ISSN(Online): 2320-9801 ISSN (Print): 2320-9798
[7] Ke Deng; Xin Li; Jiaheng Lu; Xiaofang Zhou," Best Keyword Cover Search" Knowledge and Data Engineering, IEEE Transactions on Year: 2015.
[8] W. Li and C. X. Chen, "Efficient data modeling and querying system for multi-dimensional spatial data," in GIS, 2008, pp. 58:1–58:4.
[9] D. Zhang, B. C. Ooi, and A. K. H. Tung, "Locating mapped resources in web 2.0," in ICDE, 2010, pp. 521–532.
[10] V. Singh, S. Venkatesha, and A. K. Singh, "Geo-clustering of images with missing geotags," in GRC, 2010, pp. 420–425.
[11] V. Singh, A. Bhattacharya, and A. K. Singh, "Querying spatial patterns," in EDBT, 2010, pp. 418–429.
[12] J. Bourgain, "On lipschitz embedding of finite metric spaces in hilbert space," Israel J. Math., vol. 52, pp. 46–52, 1985.
[13] H. He and A. K. Singh, "Graphrank: Statistical modeling and mining of significant subgraphs in the feature space," in ICDM, 2006, pp. 885–890.
[14] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, "Collective spatial keyword querying," in SIGMOD, 2011.
[15] C. Long, R. C.-W. Wong, K. Wang, and A. W.-C. Fu, "Collective spatial keyword queries: a distance owner-driven approach," in SIGMOD, 2013.
[16] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsuregawa, "Keyword search in spatial databases: Towards searching by document," in ICDE, 2009, pp. 688–699.
[17] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in SCG, 2004.
[18] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, "Hybrid index structures for location-based web search," in CIKM, 2005.
[19] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, "Processing spatialkeyword (SK) queries in geographic information retrieval (GIR) systems," in SSDBM, 2007.
[20] S. Vaid, C. B. Jones, H. Joho, and M. Sanderson, "Spatio-textual indexing for geographical search on the web," in SSTD, 2005.