ISSN:0975-9646

Spatial Knowledge for Disaster Identification

R.Sanjeev Reddy[#], J.Kishore Kumar*, A.V.Sriharsha[#]

 # Research Scholar,
S.V.University,Tirupathi,
A.P.,India

** Lecturer in CS ,*
S.G.Govt Degree College,Piler,
A.P.,India

Abstract: Recent developments in information technology have enabled collection and processing of vast amounts of personal data, business data and spatial data. It has been widely recognized that spatial data analysis capabilities have not kept up with the need for analyzing the increasingly large volumes of geographic data of various themes that are currently being collected and archived. On one hand, such a wealth of data holds great opportunities for geographers, environmental scientists, public health researchers, and others to address urgent and sophisticated geographic problems, e.g., global change, epidemics etc,. Our study is carried out on the way to provide the mission-goal strategy (requirements) to predict the disaster. The co-location rules of spatial data mining are proved to be appropriate to design nuggets for disaster identification and a framework has been suggested. Principal Component Analysis is a statistical method for identifying patterns.

Keywords: spatial data mining, collocation rule mining, PCA

1. INTRODUCTION:

Geography is an integrative discipline and geographic data under analysis often span across multiple domains. The complexity of spatial data and geographic problems, together with intrinsic spatial relationships, constitute an enormous challenge to conventional data mining methods and call for both theoretical research and development of new techniques to assist in deriving information from large and heterogeneous spatial datasets. *(Han and Kamber 2001; Miller and Han 2001; Gahegan and Brodaric 2002).*

For a long time spatial analysis of health data was restricted to the mapping of individual cases or rates of particular diseases and other health relevant parameters. More of these 'health' maps have become available as the use of geographical information systems in health related contexts increased. Many literary research works has been taken place such as [1][4][11].

Although disease maps may provide clues of the etiology of diseases and may facilitate decisions concerning planning of health systems, sound analytical methods are needed to assess associations between health events and etiological factors that vary gradually over geographical regions. The last two decades saw therefore a vast and still ongoing development of statistical methods for the analysis of spatial data. *((Bailey & Gatrell 1995;Cressie 1991, Haining 1990).*

This paper describes a formula implemented as Hazard science to Risk Science, towards understanding the hazards and their consequences (risks), following a probabilistic approach using spatial data mining [1].

Due to larger heterogeneity of spatial data, the providers of geographic data specify different models for same spatial objects. Context specific semantics is one of the best approach suggested which deals with provision of feature space derivations. An Ontological analysis need to be done on the fundamentals of the domain space.

A feature space consists of all input data objects, each of which is typically described by many variables (some of which, in a spatial dataset, may represent geographic characteristics and relationships). Unknown and unexpected patterns, trends or relationships can hide deep in such a huge feature space and make it very hard for analytical methods or visual approaches to find. (Miller and Han 2000).

A hypothesis space is formed by all possible configurations of the tools used to detect patterns in a feature space. Characteristically, however, the hypothesis space for a large and high dimensional geographic dataset has an extreme degree of complexity. This is caused by several factors. First, each pattern may involve a different subset of variables from the original data, and the number of such subsets (hereafter subspaces), i.e., possible combinations of attributes, is huge. Second, inside a subspace, potential patterns can be of various forms (e.g., clusters can be various shapes). Third, for a specific pattern form (e.g., cluster of a specific shape), its parameter space is still huge, i.e., there are many ways to configure its parameters. Fourth, patterns can vary over geographic space, i.e., patterns can be different from region to region.

The richness of attributes (variables, or dimensions) in a data set can provide both opportunities and challenges for data analysis. On one hand, the availability of many attributes within the data enables the identification of complex (and preferably unexpected) patterns (e.g., multivariate relationships across domains). On the other hand, it is inevitable that irrelevant attributes exist in the data and the result can be misleading or useless if the analysis method is

unable to discriminate between relevant and irrelevant attributes.

2. APPLYING SPATIAL DATA MINING

Spatial data mining becomes more interesting and important as more spatial data have been accumulated in spatial databases [9]. Mining spatial co-location patterns is an important spatial data mining task with broad applications.

2.1. SPATIAL STATISTICS

Using spatial statistics measures, dedicated techniques such as cross *k*-functions with Monte Carlo simulations have been developed to test the collocation of two spatial features. Some other economic method for finding and evaluation of the collocation is to arbitrarily partition the space into a lattice, counting the number of instances for each spatial feature that are related to the each cell of lattice.

2.2. EXTRACTING PATTERNS BY CONCEPT GENERALIZATION

At the outset the studies include, the spatial data mining problem of how to extract a special type of proximity relationship – namely that of distinguishing two *clusters* of points based on the types of their neighboring *features* is another study[2][6][8]. Classes of features are organized into concept hierarchies. Furthermore, the issues of which discriminators are "better" than other by introducing the notion of maximal discriminators, and by using a ranking system to quantitatively weigh maximal discriminators from different concept hierarchies[3].

2.3. SUPPORT OF CLUSTERING TECHNIQUES

A reasonable and rather popular approach to spatial data mining is the use of clustering techniques to analyze the spatial distribution of data. While such techniques are effective and efficient in identifying spatial clusters, they do not support further analysis and discovery of the properties of the clusters.

2.4 PRINCIPAL COMPONENT ANALYSIS

Principal components analysis (PCA). Statistical way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of High dimension, where the luxury of graphical representation is not available, PCA is A powerful tool for analyzing data. This is a six step process.

Step 1:	Get data
Step 2:	Subtract the mean
Step 3:	Calculate the covariance matrix
Step 4:	Calculate the eigenvectors and eigenvalues of the covariance matrix
Step 5:	Choosing components and forming a feature vector
Step 6:	Deriving the new data set

2.5. MINING COLLOCATION PATTERNS

Mining collocation patterns gives the standard of observing the generic characteristics of a given spatial zone with more relevant boolean features with their *s%(support)* and *c(confidence)* [6]. The work of mining co-location patterns into spatial statistics approaches and combinatorial approaches [7]. The spatial co-location pattern mining framework presented in the erstwhile works has bias on popular events. It may miss some highly confident but "infrequent" co-location rules by using only "*support*"-based pruning.

In a spatial database S, let F = $\{f_1,...,f_k\}$ be a set of *boolean spatial features*. Let I = $\{i_1,...,i_n\}$ be a set of n instances in the spatial database S, where each instance is a vector consisting of [instance-id, location, spatial features]. ~ *Neighborhood relation R over pair wise locations in S exists* ~ is assumed. The object of this collocation rule mining is to find rules in the form of **A** □ **B**, where **A** and **B** are subsets of spatial features. **A** determines the set of spatial features that form the antecedent part of the rule and **B** defines the action and its consequential parts the support and the confidence. The rule indicates the coincidence of the spatial collocation rule absorbs the action of the rule in the "*nearby*" regions of the spatial objects that comply with the collocation rule. To capture the concept of the predicate *"nearby"*, the concept of neighbor-set L is a set of instances such that all pair-wise locations in L are neighbors. Neighborhood relation R may be defined based on Euclidean distance and neighboring instances are linked by edges. A *collocation pattern* C is a set of spatial features, i.e., C□□□F. A neighbor-set L is said to be a *row* instance of collocation pattern C if every feature in C appears in an instance of L, and there exists no proper subset of L does so. We denote all row instances of a collocation pattern C as *rowset*(C). In other words, *rowset*(C) is the set of neighbor-sets where spatial features in C collocate.

The conditional probability is the probability that a neighbor-set in *rowset*(A) is a part of a neighbor-set in *rowset*□(□□B). Intuitively, the conditional probability *p* indicates that, whenever we observe the occurrences of the spatial features in A, the probability to find the occurrence of B in a nearby region is *p*.

2.6. FINDING/ESTIMATING SYMPTOMS TO BUILD COLLOCATIONS

In the imaginary figure **Figure-2.** the landscape describes two important spatial marks, sea and lake. The epidemic spread is noted in the **Figure-3.** The water in the lake is afflicted by the lichens and mosses at the western zone of the lake as most of the water is stagnant and covered by marsh. The people utilizing the water resources at this zone may be affected by so many kinds of fecal contamination in water and food. The water in the sea is contaminated with the high salts and the crude oil and base products, as the people cannot take the water for the domestic purposes, the climatic changes are affected by the water contents in the shores of sea. People breading their lives at the shores will have the indirect

contamination of fecal material in the water and as well as in the form of moisture in the air.

As the lake water is supplied into the agriculture lands surrounding in the adjacent north-east zones, there may people who are affected indirectly by the virulent characteristics.

To find out the most probable symptoms or the causative agents in particular that affect people causing disease-deaths, a probabilistic study can be made on the collected demographic-health data.

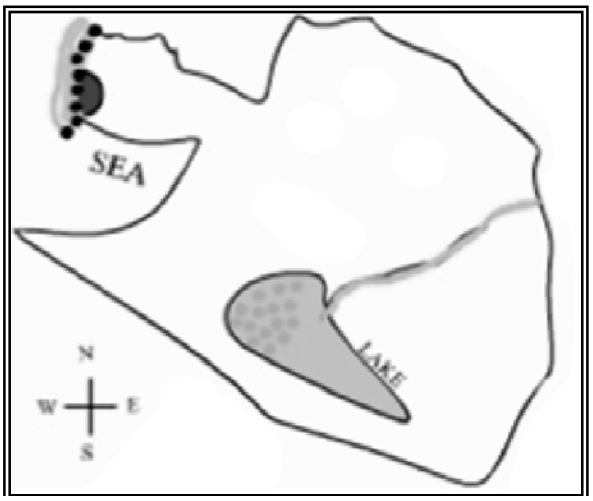


Figure-1.

As it is not easy to detect the spread of virulent features from the spatial data, the features are tested in the affected people who are native of the zone.

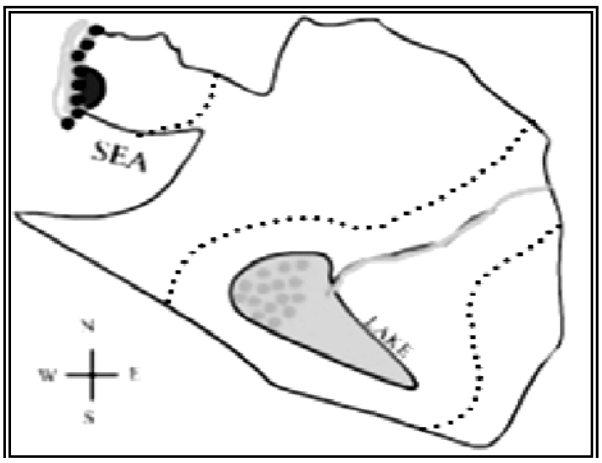


Figure-2.

3. RELATED WORK

A few epidemics that are spread due to common sources like contaminated water and contaminated food are shown in table.

(The table clearly explains about the causative agent, sources, reservoirs of the disease.)

Common Source Epidemic Diseases			
Disease	**Causative Agent**	**Infection Sources**	**Reservoirs**
Bacillary	*Shigella disenteriae (B)*	*Fecal contamination of food and water*	*Humans*
Cholera	*Vibrio cholerae (B)*	*Fecal contamination of food and water*	*Humans*
Giardiasis	*Giardia spp. (P)*	*Fecal contamination of water*	*Wild mammals*
Paratyphoid	*Salmonella partfyphi (B)*	*Fecal contamination of food and water*	*Humans*

Table-1.

The history of the epidemics has got their own influence in the world history which has taken lots of lives together. The epidemic is the result of an infection caused by any of the microbes. Every infection is a race between the microbes and the host. The microbe, following the indelible rules of evolution, strives to survive and reproduce, while the host's immune system mounts a warlike defense designed to find, destroy, and eliminate it. An agent that kills its host quickly cannot be expected to survive long enough to reproduce. Thus excessive virulence is not selected for in evolution. Gels, which can reproduce and be passed from one host to another, are favored.

Sometimes these epidemics would spread to a larger extent like a continent may turn into a pandemic. Here we are concentrating on the world's most deadly disease Cholera which was not only an epidemic but also a pandemic which shook the world with the fear of death. This disease has swept the world in seven major pandemics, including a major outbreak in South America, particularly Peru, as recently as 1991 (Avoid the cevyche!). Cases have also been reported along the Gulf coast of the U.S., usually the result of eating raw, infected shellfish. Cholera is endemic in India, Pakistan, Bangladesh, and the America. The details are shown below: The impact of the disease cholera on the world. The major cholera pandemics are generally listed as: First: 1817-1823, Second: 1829-1851, Third: 1852-1859, Fourth: 1863-1879, Fifth: 1881-1896, Sixth: 1899-1923: Seventh: 1961- 1970, and some would argue that we are in the Eighth: 1991 to the present. Each pandemic, save the last, was accompanied by many thousands of deaths. As recently as 1947, 20,500 of 30,000 people infected in Egypt died. Despite modern medicine, cholera remains an efficient killer.

3.1. COURSE OF THE DISEASE

Cholera (also called Asiatic cholera) is an infectious disease of the gastrointestinal tract caused by the Vibrio cholerae bacterium. These bacteria are typically ingested by drinking water contaminated by improper sanitation or by eating

improperly cooked fish, especially shellfish. Symptoms include diarrhea, abdominal cramps, nausea, vomiting, and dehydration. Death is generally due to the dehydration caused by the illness. When left untreated, cholera generally has a high mortality rate. Treatment is typically an aggressive rehydration regimen usually delivered intravenously, which continues until the diarrhea ceases. With treatment, mortality rates plummet. Cholera was first described in a scientific manner by the physician Garcia de Orta in the 16th century.

Can cholera be treated?

Cholera can be simply and successfully treated by immediate replacement of the fluid and salts lost through diarrhea. Patients can be treated with oral rehydration solution, a prepackaged mixture of sugar and salts to be mixed with water and drunk in large amounts. This solution is used throughout the world to treat diarrhea. Severe cases also require intravenous fluid replacement. With prompt rehydration, less than 1% of cholera patients die.

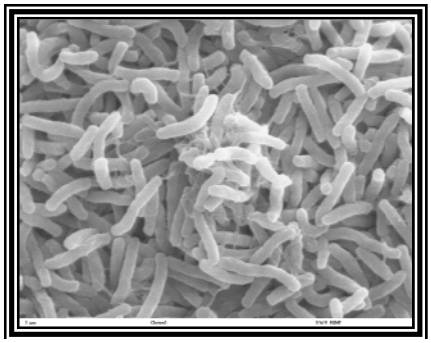
Antibiotics shorten the course and diminish the severity of the illness, but they are not as important as rehydration. Persons who develop severe diarrhea and vomiting in countries where cholera occurs should seek medical attention promptly.

The disease proceeds in possibly three stages:

(a) Invasion: at the end of the incubation period the symptoms are malaise, headache, severe diarrhea resulting in the so-called "rice water stool," (which derives its characteristic whitish color from intestinal tissue which is exfoliated (shed) and excreted along with innumerable *Vibrios*), anorexia, and a slight fever. This severe diarrhea can be as high as one liter per hour. The resulting loss of fluid and the accompanying electrolyte imbalance can lead to *hypovolemic* shock, renal failure, and cardiac failure

(b) Collapse: circulation is almost completely arrested, accelerated respiration, weak pulse, decreased systolic blood pressure, diminished or no urine output. This stage lasts from a few hours to one or two days. The mind remains clear until just before death, when coma occurs. Death follows shortly thereafter. Death can follow the onset of symptoms in little as six hours.

(c) Reaction: sometimes, even when the grim reaper is about to claim victory, vomiting ceases and diarrhea becomes less frequent and less watery, and convalescence follows.



Vibrio Cholerae
Figure-3.

4. THE LAW OF TOTAL PROBABILITY

Although there are many solutions to prevent diseases, finding the right area to apply the prevention measure with right inputs becomes the criterion. By applying the Law of Total Probability and Bayes' Theorem, the probabilities of the symptoms that are difficult to find alone are evaluated, by assuming the related symptoms occurrences, the inference is made whether the disease occurs or not occurs. The Bayes' theorem evaluates the reverse of conditionality of events; where the symptoms and the causative-agents are analyzed and found with a reciprocal equivalence. The Table-1 describes the most probable symptoms that cause the epidemics.

Let us assume, With a *priori probability*, a randomly chosen person has certain illness that will affect the society (leads to epidemic), given by P(I) = 0.001. Through the information that the person tested positive for the illness, and the reliability of the test, known to be P(Z/I) = 0.92 and P(Z/I')=0.04, according to the Bayes' theorem of *posterior probability* the person was sick with disease. The fact that the person had a positive reaction to the test may be considered as our data to build the collocation pattern.

The mining of collocation rule lay in two aspects. First, *how to identify and measure confident spatial collocation rules?* Secondly, *how to mine the patterns efficiently?*

The conditional probability of the collocation is the probability that a neighbor-set explaining the features of *existence of causative agent*, *infection sources*, is a part of the global neighbor-set in the *spatial domain* for this epidemic application Given a *spatial domain* in a database view ***S***, to measure the implication strength of a spatial feature in a collocation pattern, *a participation ratio Pr(C,f)* has to be defined. A feature *f* has a participation ratio *Pr(C,f)* in pattern *C* means whenever the feature *f* is observed, with probability *Pr(C,f)*, all other features in *C* are also observed in a neighbor-set.

In spatial application domain, as there are no natural transactions, for a continuous space, a *participation index* is proposed to measure the implication strength of a pattern from spatial features in the pattern.

For a collocation pattern ***C***, the participation index ***PI(C)*** = $\min_{f \in C} \{Pr(C, f)\}$. In other words, wherever a feature in C is observed, with a probability of at least ***PI(C)***, all other features in C can be observed in a neighbor-set. A high participation index value indicates that the spatial features in a collocation pattern likely show up together.

5. PROBLEM

Detection of the Epidemic

The collocation rules are very useful in detecting the affected areas by finding the symptoms of a disease. The collocation rule **C,** we use is

C: *cause of epidemic□□□ causative agent, infection sources; in the nearby region with high probability.*

This typical confident co-location rule involves with both frequent and rare events because although infection is quite

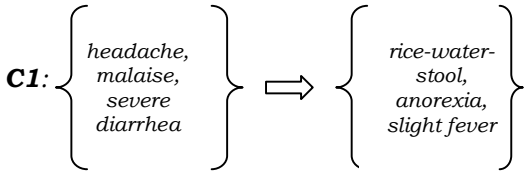
common and epidemic is rare the later factor implies the former one strongly.

As discussed in 2.4 and 2.5, by using sample identifiers, the collocation can be explained as follows:

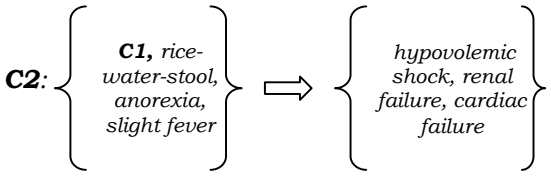
Assuming firstly, the '**b**' as the consequence of feature '**a**' is developed, forms a first level of collocation, which is identified by **a** → **b**, secondly, if the consequence '**c**' from the feature '**b**' is developed, forms a collocation, which is identified by **b** → **c**. As '**b**' already have an antecedent '**a**', the consolidated version of collocation, {**a, b**} → **c** can be formed. If '**c**' becomes another feature that can lead to the consequence of '**d**', then the notation wholly represents the cause of '**d**' as {**a, b, c**} → **d.** Also implies to {**a** ∪ **b** ∪ **c**} → **d** representation**.**

Similarly, considering the collocation pattern for the problem: ***C: {****cause of epidemic}□□□ {causative agent, infection sources}; in the* nearby *region with high probability.*

The collocation pattern is considered with practically proved parameters for cholera as follows …

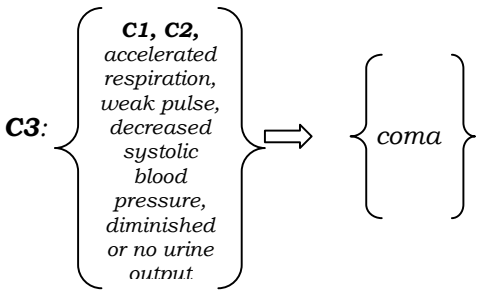


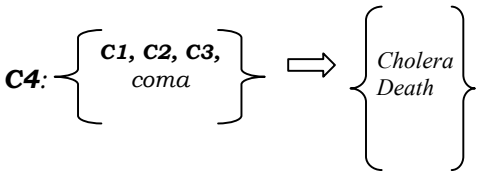
Probability <*excreted along with innumerable Vibrios with high probability*>



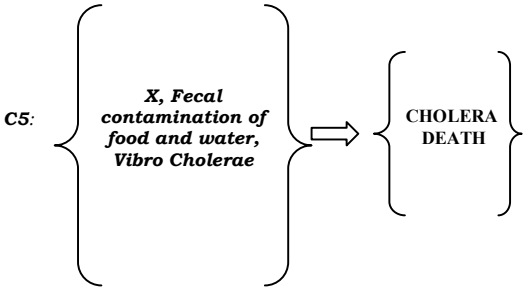
Probability <*loss of fluid, electrolyte imbalance with high probability*>

As the disease reaches "Collapse" stage, the circulation is almost completely arrested, accelerated respiration, weak pulse, decreased systolic blood pressure, diminished or no urine output.





Probability: <*the* loss *of fluid, electrolyte* imbalance *with high probability*>



Assuming ***X*** as defined representation of collocated sequence of patterns i.e., C_1, C_2, C_3, C_4 the resultant collocation C_5 is determined.

That *participation ratio* describes the intensities of the symptoms that play important role to form the collocation rule and builds the *reference feature*.

The general syntax for assessing the *reference feature* is *Pr(C, f)*.

If the probabilities of some features w.r.t C_1 are understood as having the maximum and minimum. If the lead feature of the collocation contains least probability then collocation is considered as feebly important. If the lead feature of the collocation contains higher probability then collocation is considered as highly important.

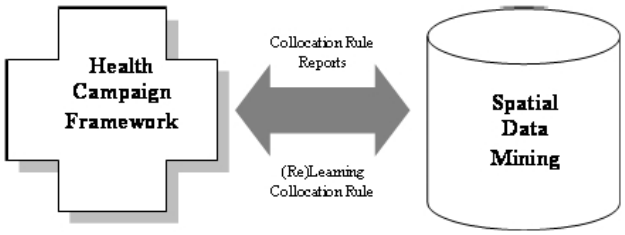
The probabilities mentioned in the problem are *<excreted along with innumerable Vibrios>*, *<loss of fluid, electrolyte imbalance>*. If one of them or some of them exhibit high probability, then there is a high significance of occurring the disease severely, for low exhibition of probability, the existing of the disease will be indicative. However, the features and the probabilities considered will prove the collocation to be appropriate for the causation of severity of cholera spectrum (simple cholera to cholera death).

Even though the participation index of the whole pattern could be low, there must be some spatial feature(s) with high participation ratio(s).

6. AN ABSTRACT FRAMEWORK

The built of framework explains the elements of the spatial knowledge support system in a *work flow strategy* and *component architecture strategy* [10][12]. The framework in work flow strategy is explained using conceptual process planning. The conceptual process planning, based on the

need of the mining of collocation patterns and help provided to the health campaigners, is briefed into *conceptual design* and *detail process design*. The following figure describes the conceptual design of the work flow strategy.

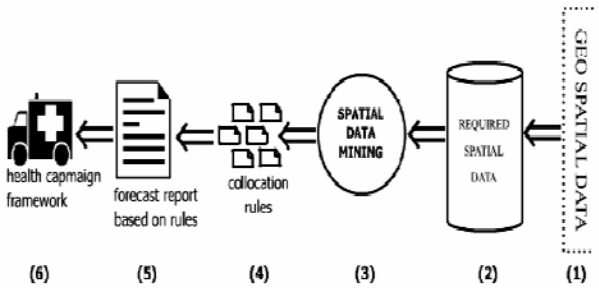


The conceptual design contains two important components, the spatial data mining infrastructure and the health campaign framework. The former defines all the necessary tools for gathering the spatial data from the spatial data warehouses. The later is defined into two classes, the software tools (developed), that defines all the necessary data base management tools that store and manage the spatial data and that which are required to transform the original geo-spatial data into the schema objects of the databases like tables, etc.

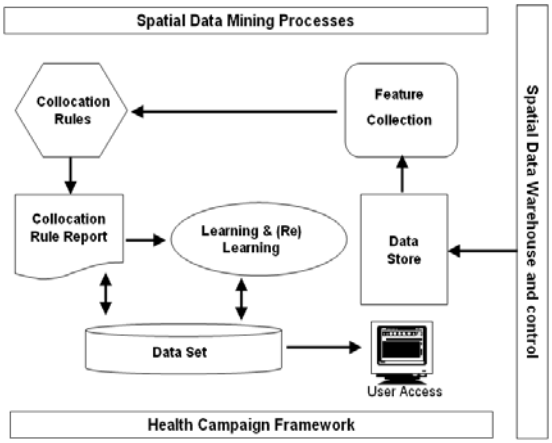
The elements of detail process design are related to the components of conceptual design. Acquiring the *spatial map* in a database representation with demographic data, a *learned database* of various sicknesses caused due to epidemics, *re-learning mechanism* to derive mappings of new patterns with learned patterns already in the database, an *antecedent-consequent based analysis*, spatial rule *generation* are associated to the first component of conceptual design framework.

The rule application and the health-campaign framework are associated to the second component of the conceptual design framework.

The following figure shows the detail process design framework of work flow strategy.



The collocation pattern formed by this sample region acts as a cautious measure or the forecast for the bio-medical researchers, analysts and other health-care-takers of the spatial zone which will be useful for them to take suitable remedial campaigns.



The boundaries of the framework are limited to design the semantic elements of the spatial knowledge support system. The detail process design explains the detailed functional decomposition of the components.

Further by analyzing the degree of relevance of the parameters given in the derivation of the Participation Index gives the basic idea of the virulent-bacterial-causative roots that develop the epidemic. Based on the parametric values, the quantitative analysis on participation index can be used to prove various alternatives of the input parameters and the seriousness of the virulent-bacterial-causative sources.

Socio-statistical methods related to health-science can be implemented to regulate the input variables that play a parametric role of collocation rule formation, in order to prevent the epidemic in the spatial zone, if not permanently, at least suitable preventive measures can be undertaken for the affect of such candidate epidemic in the interested spatial zone.

CONCLUSION

Epidemics, chronic diseases which are the major social disasters follow strategic-virulent disasters that affect the ecosystem of a spatial zone. So many parameters influence the spread of epidemic, difficult task is to find preventive measures. In this paper, a probabilistic study is made on the health demographic data, to find out the symptoms of the sick people. A collocation rule is defined as just a syntactic representation of the parameters in the form of antecedent and consequent. The participation index can boost or reduce the status of the collocation rule and there by its parameters that are used to form further. A collocation rule with weaker parameters may also exist in the spatial zone, but cannot play important role for the detection of the epidemic. Framework is described for the application of collocation rules i.e., spatial knowledge by the health campaign.

ACKNOWLEDGEMENTS

Our sincere thanks to A.V.Sreeharsha involving in the discussion throughout the experiment.

REFERENCES:

[1] Alan T.Murray, Ingrid McGuffog, John S.Western and Patrick Mullins, Exploratory Spatial Data Analysis for Examining Urban Crime, 2001

[2] Bavani Arunasalam, Sanjay Chawla, Pei Sun and Robert Munro, Mining Complex Relationships in the SDSS SkyServer Spatial Database, School of Information Technologies, University of Sydney. Proceedings of the 28th Annual International Computer Software and Applications Conference (COMPSAC'04) 2004, IEEE.

[3] Chawla, Shekkar, Spatial Databases: A Tour.

[4] Hardy Pundt, Evaluating the relevance of spatial data in time critical situations, University of Applied Sciences and Research, Faculty of Automatisation and Computer Science

[5] Huang, Shekkar, Xiong, Discovery Collocation Patterns from Spatial Data Sets: A General Approach, IEEECS, 2004.

[6] Knorr and Ng, Extraction of Spatial Proximity Patterns by Concept Generalization, (This research has been partially sponsored by NSERC Grants OGP0138055 and STR0134419. IRIS-2 Grants HMI-5 and IC-5 and a CITR Grant on "Distributed Continuous-Media File Systems – 1996)

[7] Koperski and Han, Discovery of Spatial Association Rules in GI Databases, (This research was supported in part by the research grant NSERC-OGP003723 from the Natural Sciences and Engineering Research Council of Canada and an NCE/IRIS research grant from the Networks of Centres of Excellence of Canada – 1995)

[8] Munro, Chawla, Complex Spatial Relationships, Proceedings of third IEEE International Conference on Data Mining (ICDM'03).

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Spatial Data Mining: A Database Approach, Institute for Computer Science, University of Munich, Proc' of 5th Int. Symposium on Large Spatial Databases (SSD'97).

[10] Masaharu Yoshioka, Yasuhiro Shamoto, Tetsuo Tomiyama, An Application of the Knowledge Intensive Engineering Framework to Architectural Design.

[11] Naresh Raheja, Ruby Ojha, Sunil R Mallik, Role of internet-based GIS in effective natural disaster management, R. M. Software India Pvt. Ltd. (RMSI)

[12] Shaw C. Feng, Y. Zhang : Conceptual Process Planning - A Definition and Functional Decomposition, Manufacturing Engineering Laboratory, National Institute of Standards and Technology, 1997.

ABOUT AUTHORS

R.Sanjeeva Reddy,. Research Scholar in PhD, Computer Science, SVUniversity, Tirupati.

J.Kishore Kumar, presently working as a lecturer in Dept.of CS,SGGDC, Piler,Chittoor Dist,A.P,INDIA

A.V.Sriharsha, Research Scholar in PhD, Computer Science and Engineering, SVUniversity, Tirupati. Completed his B.Tech in Computer Science from Andhra University, Completed M.Tech in Information Technology from Sathyabama University, Chennai.