A CLUSTER BASED APPROACH FOR GEOGRAPHIC MAPPING

Piyali Sarkar
M.Tech C.S.E. Dept. of Computer Engineering
St.Vincent Pallotti College of Engg & Tech. Nagpur, India

Abstract- Today we can't imagine our lives without cell phones which have become part and parcel of our life. Technology today had made life easier and better. Often a country's technological advancement is measured on the basis of how people can easily communicate with each other. Now-a-days many application software are being introduced in markets for various purposes to make life easier. The apps which are widely used are the GPS system or Google maps which are used for navigation purposes. GPS systems can be used to calculate latitudes, longitudes and altitudes of current location irrespective of their location. The work which has been done is addition and only a small contribution to the work done by researchers all around the world in the field of geo-tagging. Here a dummy database is created which includes landmarks, city and state and various statistical processes are used to cluster data according to the distance measured. All the three methods are compared and analyzed on the basis of formation of fair clusters. It is represented on map according to the database of longitudes and latitudes and distance measures. Then the clusters are decided whether the landmarks are present in the same city or not according to a predefined value of radius. For instance, if the data is in format India gate, Delhi; Lotus Temple, Delhi; Taj-Mahal, Agra, Uttar-Pradesh; then India Gate and Lotus temple should be present in the same city of Delhi. Taj mahal, present in Agra in near to Delhi. But it should not be shown in the cluster. There might be a possibility that within 100 m radius while representing clusters it may be included. We have tried to solve this problem and also tried to give accurate results as much as possible. Also the data which is used is taken in sequential format rather than random points.

Keywords—Geo-tagging, geo-coding, clustering, feature extraction, statistical methods.

I. INTRODUCTION

Mobiles are no longer restricted to calls. They have been used for various purposes like music, video, text,To-do lists and many more. Moreover, an added advantage to all recent technologies, they can be controlled and used by the end users.

We can state that mobile has become a tool which is widely used to access the internet. Also they are used for playing games, for television and GPS and also video calling and barcode reading. In past few years technology has witnessed many researchers proposing high number of navigation systems. Some of these systems were built successfully and have been in use since for outdoor applications namely GPS system or GLONASS (Global Navigation Satellite System) which are widely used in transport applications to find the shortest route to the destination and also for pedestrian navigation.

Geo-tagging refers to the process of adding tags or information, basically geographical information to any media posted on social sites. Geo-tagging can be further classified as Geo-Coding, Geo-Blogging and Geo-Microformat. Geo-Coding refers to the process of taking non-coordinate based geographical identifiers such as street address and searching associated geographic coordinates. Geo-Blogging refers to the process of tagging blog data to attract search engine users searching for a particular blog which contains any geographical information. Geo-microformat is a process where it represents a physical location on the Earth using latitudes and longitudes. Here the main focus is on Geo-coding and the mapping is done on the principles of geo-microformat.

Geo-Coding or in other words forward geo-coding refers to give meaningful description to a particular location such as street address or postal address or place name by using geographic co-ordinates such as latitudes, longitudes and elevation for projection on map. Reverse Geo-coding refers to give meaningful description to a geographic co-ordinates for location or place name. Further Geo-coding can be classified as interpolation and point-level geo-coding. The address interpolation technique can be well described with the help of a suitable example. Consider a street address 166 Pandey layout. The entire locality ranges from 110-210. Let us consider that even numbers fall on the west side and odd numbers fall on the east side of the street. The work of geo-coding is to match an address with a street block as every block contains a different street range, and would interpolate or calculate the position. In this example, 166 is almost half way from 110 towards 210. This point would be on the west side of the street. A point will be mapped at this location. But this process has no provisions to solve the problem of ambiguity. The problem arises when the street address would be 166 pandey layout and 166 E pandey layout. However this problem of ambiguity can be solved by defining proper address sets including city, state and country. Another problem that interpolation faces is to map addresses which have not been added to the database. Point level geo-coding gives more accurate mapping by using centroid of a particular building or land parcel and then maps the location. This is more effective for flood determination.

Geocoding is the process of assigning a location, usually in the form of coordinate values, to an address by comparing the descriptive location elements in the address to those present in the reference material.[1][2]

Geocoding, also called address matching, can also be defined as the process of plotting geographic locations (identifiers) for data that is stored in tabular format. One can use geocoding to plot the geographic positions of a set of addresses in a customer database or one can plot Latitude and Longitude points recorded from a GPS unit. Thus, Geocoding requires a database of properly formed address, a database of streets and a set of rules for matching the addresses to the streets.

Addresses come in many forms, ranging from the common address format of house number followed by the street name and succeeding information to other location descriptions such as postal zone or census tract. In essence, an address includes any type of information that distinguishes a place.

W Olive Ave
S 4th St
E Fern Ave
Grant St
205 W Fern Ave,
Redlands, CA 92373
S Buena Vis St
Grant St
Cypress Ave
orth Pl
0m
304.43m
608.86m
1:14204

When you type an address or a place name in the search box and in return the map shows a marker at the place. The process of associating an address or a place name with coordinates on the map is called **Geocoding**. In a spatial database this is done as a point layer with name of the place as an attribute to the point location. This is one way of geocoding. For addresses, the associated coordinates are not saved in a database directly, but computed using a method called linear referencing. The start and end addresses along a line segment are saved and intermediate addresses are interpolated and the coordinates are calculated.

In some online mapping service, you may have seen satellite imagery. When these images are captured from a satellite or an airplane, they are just plain images, like photographs. But to display these images on a map, they need to be associated with map coordinates. This process is called **Geo-referencing**. Once the image is associated with the map coordinates it can be overlaid on top of street maps. For georeferencing, you can use a GIS software such as ArcGIS or QGIS to georeference an otherwise un-referenced image or scanned maps, and load them into Oracle Spatial.[2]

The widely used technique is the GPS system. GPS systems were first developed for military applications. But today any handheld devices have GPS unit which can receive radio signals that satellites transmit. Irrespective of our location atleast four satellites are visible. GPS is used to provide location and time information. GPS systems are accurate theoretically up to 14 nanoseconds but are not authenticated and also don't provide any provisions to avoid false data shown on maps. For example, if someone feeds information about Parvati complex near Sadar, inspite of Padole Square which is the actual address of the mentioned place then it would show the same on the map.

Many companies have found geo-location or geo-targeting technology to be of value in Internet advertising. Pay-per-click search engines like Google and Yahoo offer the ability for advertisers to deliver targeted advertising banners based on the location of the website visitor's IP address. For marketers, geocoding is critical in targeting specific demographics. Appending demographic census track data to latitude and longitude coordinates helps marketers target the right demographics - those who would be most likely to respond to their offer or marketing message. Insurance companies are relying increasingly on geocoding techniques to help set premiums and make underwriting decisions based on the physical location of the insured property. Take Hurricane Katrina and storm-surge damage, for example. Most insurance carriers have their own set of rules and criteria when it comes to underwriting, such as property elevation and determining the distance of the property from/to the coast. Such an imprecise standard may leave carriers insuring properties that may not be situated in a flood zone, but are actually in a storm-surge zone - where the flood exclusions in their policies would not apply. [3]

One of the most widespread uses of geocoding technology is in store/dealer locators. Businesses use geo-coded data to ascertain proximity to potential customers, distance to suppliers and competitors, service areas and delivery routes. You've probably experienced a locator lookup yourself - maybe to find a restaurant, pet shop or the Sprint Nextel store nearest your home or business. However, to adequately serve its 52 million customers, Dominos often has multiple stores located within the same ZIP Code. So which neighborhood store is closest to a particular customer's home? Using a geocoding solution to power the Store Locator on its website allows Dominos to turn the street addresses of its stores into usable location information - so customers can actually determine which store is closest in relation to their home address. But what if the address is wrong? Without accurate addresses, it would be difficult to obtain accurate geocoding. For instance, a geocoding application might not recognize the difference between 123 Pandey Layout and 123 S. Pandey Layout, which could be two totally different addresses located miles apart. A bad address diminishes the accuracy of a store locator - it's the biggest reason why some store locators get it wrong. That's why businesses are more proactive about integrating routines for address verification with their geocoding initiatives.

II. LITERATURE SURVEY

The most widely used technique in the field of geo-tagging is Chi square method. Chi square is used to define a relation between two points if there is no relation between them. The work of chi square can be defined as guessing. In context of statistics, chi square is defined as the comparison or difference between the observed and expected value. In context of text mining chi square is used by considering a term 't' in a document and looking up to the probability of occurrence of 't' in the given document and probability that the term 't' is present in other document i.e. other than the preferred document. In geo-tagging it is used as every city is counted and compared that how many times it has occurred.[1][2]

The best way to describe this method let us take an example of some cities from state Maharashtra and Madhya Pradesh. The cities and their corresponding states are listed in the following table 1:

City	State
Nagpur	Maharashtra
Bhopal	Madhya Pradesh
Mumbai	Maharashtra
Nasik	Maharashtra
Aurangabad	Maharashtra
Bilaspur	Madhya Pradesh
Raipur	Madhya Pradesh

Table 1. List of city and their respective states

The above table is a list of cities with their respective states. As we observe, every city is a unique city. No city is repeated. Therefore the observed value will be 1 for all cities as they occur only once in a particular state and also state other than the one in which they are present. Next is the expected value. First the observed value is summed up. Here in this example it is 7. Now there are two states shown in the above table. So the expected value will be 7/9 for each cell. The expected value will be 0.7777. This value will be same for all the cities as they have the same observed value. This is shown in the table below :

City	State	Observed Value	Expected Value
Nagpur	Maharashtra	1	0.7777
Bhopal	Madhya Pradesh	1	0.7777
Mumbai	Maharashtra	1	0.7777
Nasik	Maharashtra	1	0.7777
Aurangabad	Maharashtra	1	0.7777
Bilaspur	Madhya Pradesh	1	0.7777
Raipur	Madhya Pradesh	1	0.7777

Table 2. Expected and Observed Values

Now Chi Square can be calculated by the formula given below:

$$\chi^2 = \sum \frac{(O-E)^\wedge 2}{E}$$

Where 'O' is the observed value and 'E' is the expected value. The difference is squared and divided by the expected value.

The pro of this method is that it is very easy to implement. But the cons of this procedure are it lacks in accuracy. Accurate clustering requires precise definition of closeness between a pair of objects. Moreover chi square never works for continuous data and is designed to analyze categorical data. Since the work of chi square is guessing, accurate relations are not defined.

To overcome these drawbacks of chi square, KDE was introduced. KDE estimates the density directly from the data without assuming a particular form for the underlying distribution.

Mathematically, a kernel is a positive function. This is given by K(x; h) where h is the bandwidth parameter and this function is controlled by 'h'. [2] If the kernel form is 'Gaussian' the density estimate at a point 'y' within a group of points x_i where i=1,2,3,……,N is given by :

$$P_K(y) = \sum_{i=1}^{N} K((y - xi)/h)$$

The bandwidth act as a smoothing parameter controlling the trade off between bias and variance in the result.

In context of geo-coding, KDE can be used with a valid distance metrics though the results are only normalized by Euclidean distance. For navigation purpose haversine distance is used which gives much accurate circles on the sphere between two points from their latitudes and longitudes. The haversine distance can be calculated by the following equation:

$$\text{Haversin } (d/r) = \text{haversine}(\varphi_2-\varphi_1) + \cos(\varphi_1)\cos(\varphi_2)$$
$$\text{haversine } (\lambda_2-\lambda_1)$$

Where λ_1 and λ_2 are longitudes at point 1 and 2 respectively and φ_1 and φ_2 are latitudes at points 1 and 2 respectively. Further the haversine function is given as $\sin^2(\theta/2)$. Also, haversine (d/r)=h which is calculated as inverse sine function given by r $\text{haversine}^{-1}(h) = 2r \arcsin(\sqrt{h})$ where arcsine is nothing but inverse sine. 'r' is the radius of the sphere. Here the image given below shows a geo-spatial data representing two different species in South American continent.

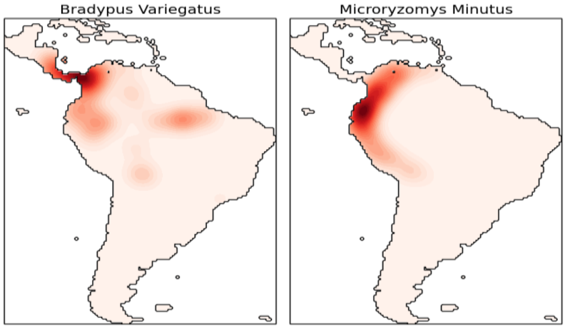


Figure 2. KDE with haversine distance

KDE lacks because it yields density estimates that have discontinuities and weights equally all points regardless of their distance to the estimation point. The basic principle of kernel density estimation is suppose there is a point say 'A' and it is to be estimated that whether it is possible to club within the cluster. But the entire densities of the cluster or nearer points are clubbed together. Even if point A should be within the cluster or even if related then it is not included within the cluster as it is not nearer to the entire cluster. Thus point wise consideration is not considered. Also in KDE if the bandwidth is smaller it leads to unsmooth density distribution. Performance degrades in higher dimensions.

III. PROPOSED METHODOLOGY

Thus to overcome these drawbacks of Chi square and KDE cosine similarity method has been used. In context of text mining cosine similarity measure is used in the following manner: consider a document such as

T= {The sky is blue

The sun is bright

The sun in the sky is bright.}

In cosine similarity, the first statement is compared with itself and with other two statements. In the first statement when it is compared with itself all the words are same i.e. the statement will be same and its value will be 1. When the statement will be compared with the second statement the value will be much less. When it will be compared with the third statement there are many similar words i.e. 'The' , 'sky', 'is' therefore the value will be nearer to 1.

The documents will be ranked accordingly. If the problem of synonymy occurs, cosine inspite of allowing partial matching this problem can be overcome by using linear algebra or word text or any geographical gazette/dictionary is used.

In context of geo-coding,

$$\cos\theta = \overrightarrow{a}\,.\,\overrightarrow{b} \;/\; ||\overrightarrow{a}||\;\;||\overrightarrow{b}||$$

The cosine similarity calculates the cosine of angle between two vectors. By this formula it can be stated that how can two points be related by seeing the angle and not the magnitude.

Cosine similarity is mostly used in high dimensions. For instance consider a database which consists of cities and states. Every city is a unique city within that particular state and other than that state. Therefore a different dimension is assigned to each city say in terms of latitudes and longitudes. Also, the cities will be characterised by a vector where the value of each dimension may depend on its occurrence in the entire database. In this situation cosine gives how two points can be likely to each other. Cosine similarity gains its popularity due to measurement of cohesion in clusters in data mining. Also, cosine similarity is a good support for positive spaces. But it does have a name in good metrics. The reason behind this is that it doesn't have triangle inequality property. The triangle inequality property states that any side of a triangle is less than the sum of other two sides of a triangle. To get through this, angular similarity is used i.e. converted to angular similarity.

IV. EXPERIMENTAL RESULTS

All the three methods are compared on the basis of accuracy and elapsed time. Accuracy in terms of cluster formation is measured for all the three statistical methods. The following table shows the comparison on basis of elapsed time by all the three methods.

Methods	Elapsed Time(Sec)
Chi Square	6.6196
Cosine Similarity	2.4329
KDE	4.9948

Table 3. Comparison on the basis of Elapsed time

The following graph is the pictorial representation of the above table showing the most effective method in terms of elapsed time.

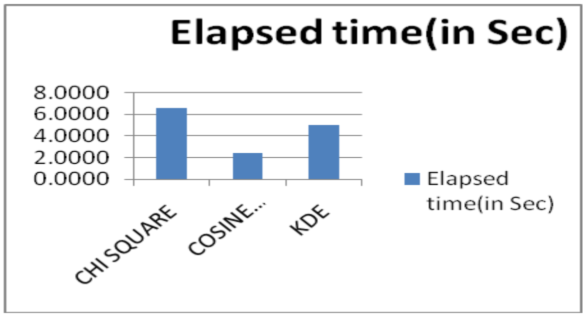


Figure 3. Graph representing the Elapsed time of all the three methods

The values of first 16 records are as follows:

CITY	CHI VALUES	COSINE VALUES	KDE VALUES
Achalpur	0.9208	0.0045	0
Achhnera	0.9208	0.0625	0.42
Adalaj	0.9208	0.0758	1.8
Adilabad	0.9208	0.0803	3.12
Adityapur	0.9208	0.0116	4.38
Adoni	0.9208	0.1205	6.3
Adoor	0.9208	0.1294	7.08
Adra	0.9208	0.1651	8.64
Adyar	0.9208	0.2007	9.54

Table 4. Calculated Values of all the three methods

There are total 1400 records of nearly all cities in different states. Here the cities are unique cities and are considered in a matrix formation while calculation of cosine similarity and KDE representing number of rows as i and columns as j. The graph shown below is for 50 records :

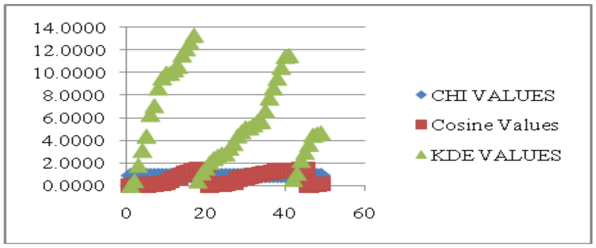


Figure 4. Graph showing the values for the methods

For clustering, a database with latitudes and longitudes will be used. The points or the cities which will be plotted using latitudes and longitudes will be shown on the map. They are clustered on the basis of states right from the pre-processing step.

Accuracy is compared by comparing the values calculated by each of the three methods out of the total. The analysis is as follows :

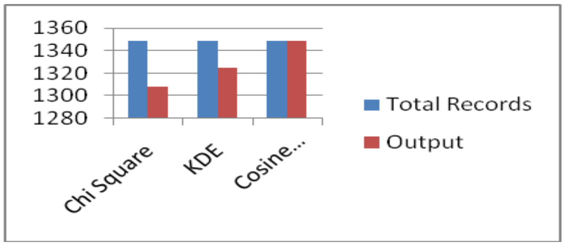
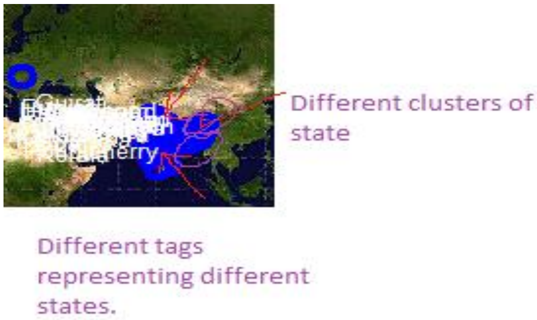


Figure 5. Graph showing the accuracy of each method out of the total cities considered

The clusters thus formed are shown in the figure below :

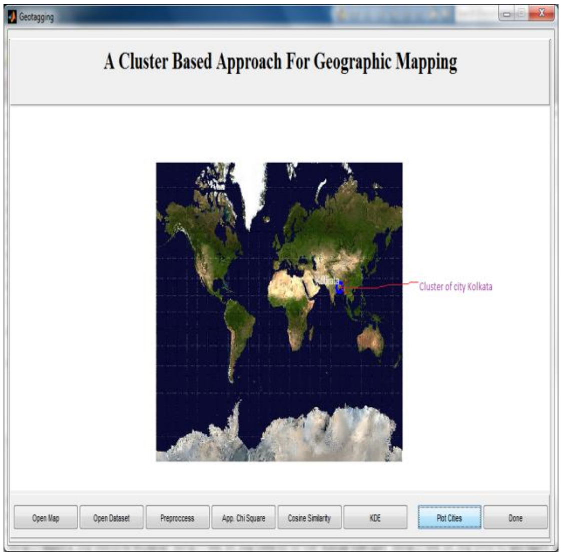


Different clusters of state

Different tags representing different states.

Figure 6. Clusters of all states

For a single city, the cluster is shown below :



A Cluster Based Approach For Geographic Mapping

Figure 7. Cluster of the city Kolkata which includes some of its landmarks.

CONCLUSION

Cosine is known as one of the best metric as compared to Chi square in terms of accuracy and time. It is more accurate because Chi square guesses the relation but not cosine similarity. Some of the problems which occur with cosine are also considered and solved regarding synonymy and triangle inequality by using appropriate measures. As compared to KDE, as it faces the problem of considering individual points, cosine gives more accurate results in such case. Both work in higher dimensions where KDE performs more accurately but not with smaller dimensions where cosine works far better. Also cosine has least elapse time and also more accurate than other two methods.

REFERENCES

[1] S. Dinesh, K.S. Kannan, "Kernel Based Tagging Method Using Spatial Paradigm," *International Journal of Computer Science and Mobile Computing, IJCSMC, Vol.3, Issue. 11, November 2014, Pg. 73-80.*

[2] Olivier Van Laere, Jonathan Quinn, Steven Schockaert , Bark Dhoedt, Member, IEEE, "Spatially Aware Term Selection for Geo-tagging," *IEEE transactions on Knowledge and Data Engineering, Vol. 26, No.1, January 2014*.

[3]Y.Yang and J.O. Pedersen, "A Comparative Study On Feature Selection in Text Categorization," *Proc. 14th Int'l Conf. Machine Learning, pp. 412-420 , 1997.*

[4] C. Hauff and G.J. Houben, "WISTUD at MediaEval 2011: Placing Task," *Proc. Working Notes Of the MediaEval Workshop, 2011.*

[5] Martha Larson, Mohammad Soleymani, Pavel Serdyukov,"Automatic Tagging and Geotagging in video collections and communities" ICMR,2011.

[6] Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, Christopher B. Jones, " Geo-referencing Wikipedia Documents Using Data From Social Media Sources," *ACM Transactions on Information Systems*

[7] S. Ahern, M. Naaman, R. Nair, and J. Yang, " World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections," *Proceedings of the SeventhACM/IEEE-CS Joint Conference on Digital Libraries, May 2007*