IMPLEMENTATION OF K-MEANS AND K-MEDOIDS CLUSTERING ALGORITHMS

B.Tech. Final Year Project Report

BY

ANISHA PAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NEOTIA INSTITUTE OF TECHNOLOGY, MANAGEMENT & SCIENCE
JHINGA, AMIRA, D.H.ROAD, 24-PARAGANAS (S) - 743368, WB (INDIA)

November, 2016

IMPLEMENTATION OF K-MEANS AND K-MEDOIDS CLUSTERING ALGORITHMS

A Major Project Report

Submitted in partial fulfillment of the requirements for the award of the degree

Of

Bachelor of Technology

In

COMPUTER SCIENCE AND ENGINEERING

BY

ANISHA PAL	**14400113005**	**131440110005**

Under the Guidance of
Prof. Subrata Bose

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NEOTIA INSTITUTE OF TECHNOLOGY, MANAGEMENT & SCIENCE
JHINGA, AMIRA, D.H.ROAD, 24-PARAGANAS (S) - 743368, WB (INDIA)

November, 2016

CERTIFICATE OF ORIGINALITY

I hereby certify that the work which is being presented in the B.Tech. Final Year Project Report entitled **"IMPLEMENTATION OF K-MEANS AND K-MEDOIDS CLUSTERING ALGORITHMS",** in partial fulfilment of the requirements for the award of the **Bachelor of Technology in Computer Science & Engineering** and submitted to the Department of Computer Science & Engineering of Neotia Institute of Technology, Management & Science, West Bengal is an authentic record of my own work carried out from July, 2016 to November, 2016 under the supervision of **Prof. Subrata Bose**.

The matter presented in this thesis has not been submitted by me for the award of any other degree elsewhere.

Signature of Candidate

Anisha Pal

14400113005

131440110005

CERTIFICATE OF RECOMMENDATION

This is to certify that the Project entitled “IMPLEMENTATION OF K-MEANS AND K-MEDOIDS CLUSTERING ALGORITHMS” has been submitted by **MS. ANISHA PAL** under my guidance in partial fulfilment of the degree of Bachelor of Technology in Computer Science & Engineering of Neotia Institute of Technology, Management & Science, Jhinga, WB during the academic year 2016-2017.

Signature of Supervisor
Prof. Subrata Bose

Head of the Department
Department of Computer Science & Engineering
Neotia Institute of Technology, Management & Science, Jhinga, West Bengal

Date:
Place:

ACKNOWLEDGEMENT

This Project has been carried out to meet the academic requirements of Maulana Abul Kalam Azad University of Technology, West Bengal. I would like to put on record, my appreciation and gratitude to all who have rendered their support and guidance. Without them, it would not have been possible for me to shape this project.

I have received immense guidance from my project supervisor Mr. Subrata Bose, Professor and Head of the Computer Science Engineering Department, NITMAS. I would therefore like to convey my sincere gratitude to him.

Anisha Pal

INDEX

Sl No	Topic	Page
1	Abstract	7
2	Introduction	8
3	Cluster Analysis	10
	3.1 Application of Cluster Analysis	10
	3.2 Application examples of Cluster Analysis	11
4	Basic steps to develop a Clustering Task	12
5	Partitioning Algorithms: Basic Concept	14
	5.1 Points to remember	15
6	The K-means Clustering method	16
	6.1 K-means algorithm to be followed	16
	6.2 An example of K-means clustering	17
	6.3 Problems of the K-means method	18
7	Elbow method	18
8	K-medoids	20
	8.1 What makes the distance measure in K-medoid better?	21
9	Work done so far	22
10	Plan of work for next semester	22
11	References	23

LIST OF FIGURES

Fig	Topic	Page
1	Partitioning Algorithm	13
2	Hierarchical based Algorithm	14
3	Density based Algorithm	14
4	Grid based Algorithm	15
5	Steps of K-means Clustering Algorithm	18
6	Elbow method graph	20

ABSTRACT

Data clustering is an unsupervised data analysis and data mining technique, which offers refined and more abstract views to the inherent structure of a data set by partitioning it into a number of disjoint or overlapping (fuzzy) clusters. Hundreds of clustering algorithms have been developed by researchers from a number of different scientific disciplines. The intention of this report is to present a special class of clustering algorithms, namely K-means and K-medoids (under partitioning based), where cluster validation is an important step and it is evaluated based on the similarity between two clusters. Cluster analysis is used everywhere be it during public census, the classification of species of animals and plants or similar diagnostic cases.

INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases

Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- Non-operational data, such as industry sales, forecast data, and macro-economic data
- Metadata - data about the data itself, such as logical database design or data dictionary definitions

Information

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products is selling and when.

Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data Warehouses

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

What can data mining do?

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

CLUSTER ANALYSIS

Cluster analysis (or clustering, data segmentation) is finding the similarities between data according to the characteristics found in the data and grouping similar data objects into clusters

It is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Applications of Cluster Analysis

- Data reduction
    - Summarization: Preprocessing for regression, PCA, classification, and association analysis
    - Compression: Image processing: vector quantization
- Hypothesis generation and testing
- Prediction based on groups
    - Cluster & find characteristics/patterns for each group
- Finding K-nearest Neighbors
    - Localizing search to one or a small number of clusters

- Outlier detection: Outliers are often viewed as those “far away” from any cluster

Application examples of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.

- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.

- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.

- Clustering also helps in classifying documents on the web for information discovery.

- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

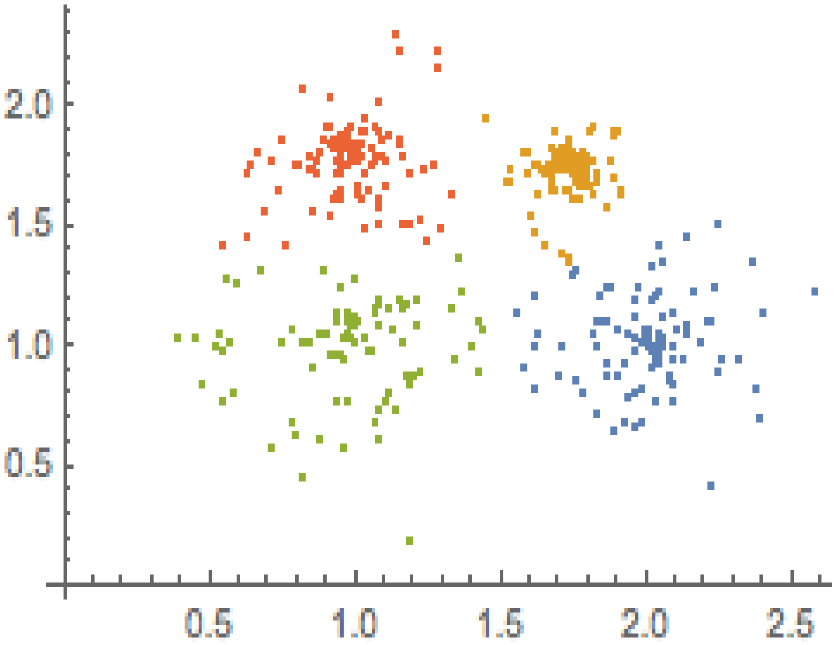
- Climate: understanding earth’s climate, finding patterns of atmospheric and ocean

- Economic Science: market research.

BASIC STEPS TO DEVELOP A CLUSTERING TASK

There are different methods of clustering, known as:

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARA, CLARAN



Fig, 1 Partitioning Algorithm

- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, CAMELEON

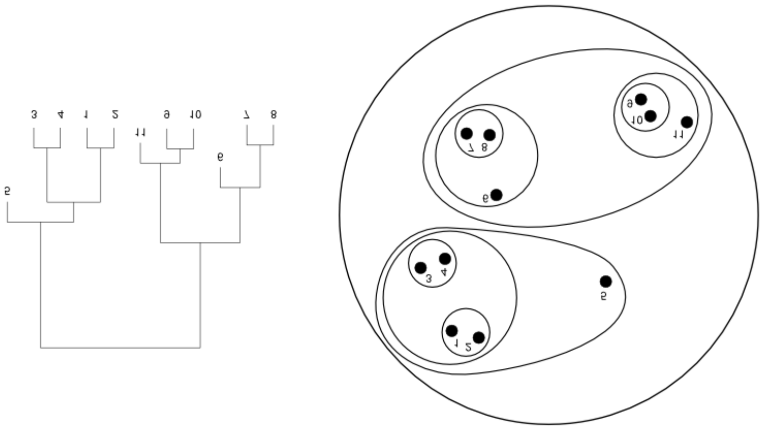


Fig.2 Hierarchical based Algorithm

- Density-based approach:
    - Based on connectivity and density functions
    - Typical methods: DBSACN, OPTICS, DenClue

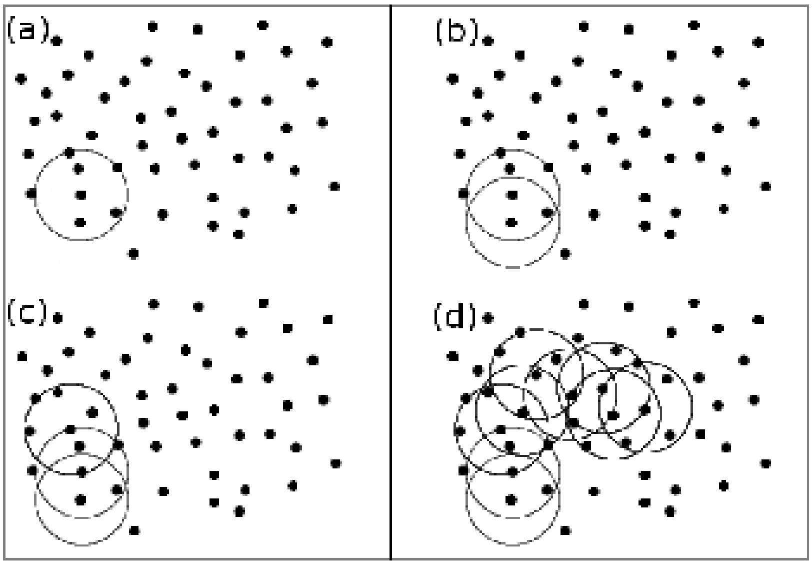


Fig.3 Density-based Algorithm

- Grid-based approach:
    - Based on a multiple-level granularity structure
    - Typical methods: STING, WaveCluster, CLIQUE

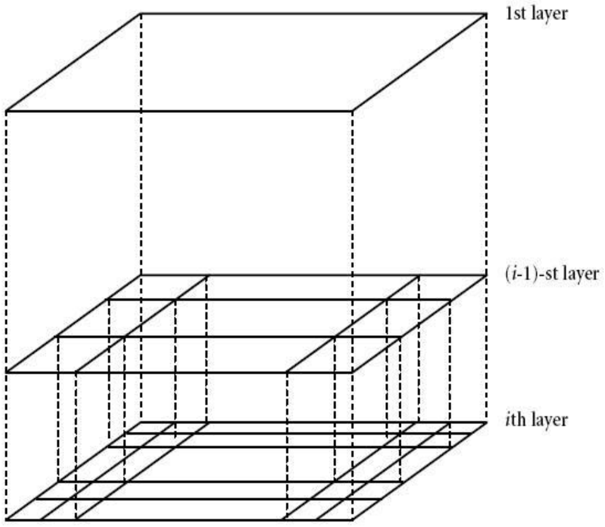


Fig.4 Grid-based Algorithm

PARTITIONING ALGORITHMS: BASIC CONCEPT

Suppose we are given a database‘d’ of ‘n’ objects and the partitioning method constructs ‘k’ partition of data(clusters). Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.

- Each object must belong to exactly one group.

- The sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (d(p, c_i))^2$$

Points to remember

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.

- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

  - Given *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion

    - Global optimal: exhaustively enumerate all partitions

    - Heuristic methods: *k-means* and *k-medoids* algorithms

    - *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster

    - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster.

THE *K-MEANS* CLUSTERING METHOD

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function.

The objective function

$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$,Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j , is an indicator of the distance of the *n* data points from their respective cluster centers.

K-means Algorithm to be followed

Given *k*, the *k-means* algorithm is implemented in four steps:

- Partition objects into *k* nonempty subsets
- Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
- Assign each object to the cluster with the nearest seed point
- Go back to Step 2, stop when the assignment does not change

An Example of *K-Means* Clustering

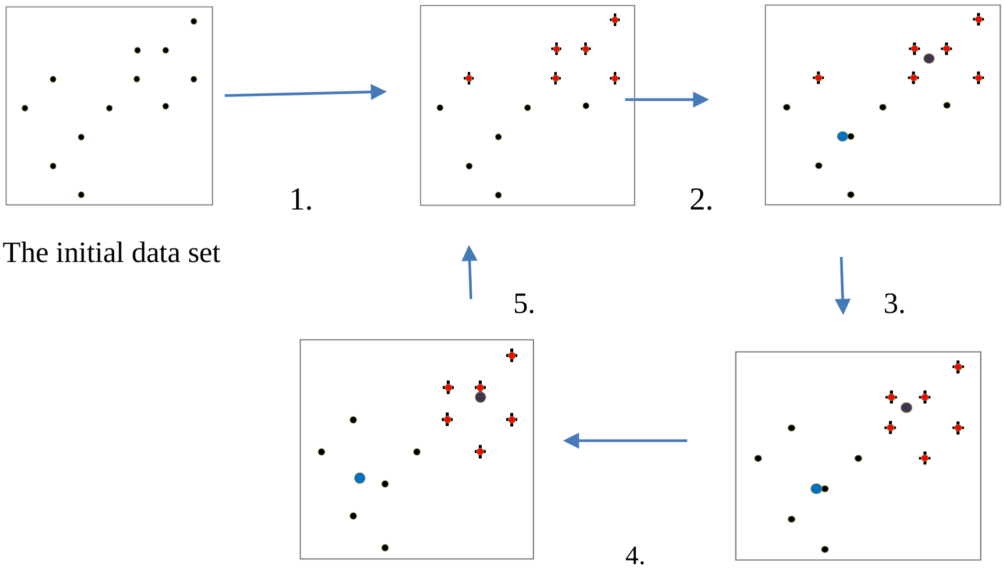


Fig.5 Steps of K-means Clustering Algorithm

Arbitrarily partition objects into k groups

- Update the cluster centroids

- Reassign objects
- Update the cluster centroids
- Loop if needed

Problem of the K-Means Method

- **Sensitive to scale**

  Rescaling your datasets will completely change the results. While this it is not bad, not realizing that *you have to spend extra attention to scaling your data* is bad.

- **Even on perfect data sets, it can get stuck in a local minimum**

- **Means are continuous**

- **Too easy to use badly**
  All in all, it's too easy to throw k-means on the data, and nevertheless get a result out (that is pretty much random, but it is not noticeable). It would be better to have a method which can fail if you haven't understood the data.

ELBOW METHOD

The **Elbow Method** looks at the percentage of variance explained as a function of the number of clusters. One should choose a number of clusters so that adding another cluster doesn't give much better modelling of the data. If one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This elbow cannot always be unambiguously identified. Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test. A slight variation of this method plots the curvature of the within group variance.

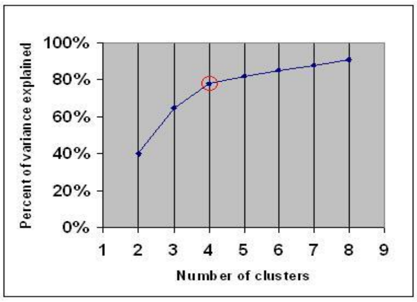


Fig.6 Elbow method graph

K-MEDOIDS

The k-means algorithm is sensitive to outliers. Since an object with an extremely large value will substantially distort the distribution of the data

Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster

The ***k*-medoids algorithm** is a clustering algorithm related to the *k*-means algorithm and the medoid shift algorithm. Both the *k*-means and *k*-medoids algorithms are partitioned (breaking the dataset up into groups) and both attempt to minimize the distance between points labelled to be in a cluster and a point designated as the center of that cluster. In contrast to the *k*-means algorithm, *k*-medoids chooses data points as centers (medoids or exemplars) and works with an arbitrary metric of distances between data points.
K-medoid is a classical partitioning technique of clustering that clusters the data set of *n* objects into *k* clusters known *a* priori.

The most common realization of *k*-medoid clustering is the **Partitioning around Medoids (PAM)** algorithm

- *K-Medoids* Clustering: Find *representative* objects (medoids) in clusters

- *PAM* (Partitioning Around Medoids, Kaufmann & Rousseeuw 1987)

    - Starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
    - *PAM* works effectively for small data sets, but does not scale well for large data sets (due to the computational complexity)

- Efficiency improvement on PAM

- *CLARA* (Kaufmann & Rousseeuw, 1990): PAM on samples
- *CLARANS* (Ng & Han, 1994): Randomized re-sampling

What makes the distance measure in k-medoid better?

- **K-medoid is more flexible**

  First of all, k-medoids can be used with any similarity measure. K-means however, may fail to converge - it really must only be used with distances that are consistent with the mean. So e.g. Absolute Pearson Correlation must not be used with k-means, but it works well with k-medoids.

- **Robustness of medoid**

  Secondly, the medoid as used by k-medoids is roughly comparable to the median (in fact, there also is k-medians, which is like K-means but for Manhattan distance). If you look up literature on the median, you will see plenty of explanations and examples why the median is more robust to outliers than the arithmetic mean. Essentially, these explanations and examples will also hold for the medoid. It is a more robust estimate of a representative point than the mean as used in k-means.

WORK DONE SO FAR

In the current semester, I have implemented the K-means clustering algorithm taking input. The inputs are obtained by two ways:

- Taking input numbers from the user and also asking for the number of clusters to be formed with the given number.
- Taking input numbers from a file that contains a set of single data values.
- Implementation of k-means algorithm on data values which are in pairs such (height, weight) of persons
- Read input data from file
- Application of Elbow Method to find the elbow point

PLAN OF WORK FOR NEXT SEMESTER

Following the work done during the 7th semester, the tasks that are ahead to be done are

- Pictorial representation of clusters in Pie/Bar Chart
- Application of K-medoids algorithm to remove the outliers while cluster formation.

REFERENCES

- https://en.wikipedia.org/wiki/Cluster_analysis
- Data Mining: Concepts and Techniques, Jiawei Han.
- https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
- https://en.wikipedia.org/wiki/K-means_clustering
- https://en.wikipedia.org/wiki/K-medoids
- https://www.researchgate.net/publication/220215167_A_simple_and_fast_algorithm_for_K-medoids_clustering
- www.google.com/images
- www.bing.com/images