Volume 4, Issue 11, November 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Online Image Retrieval Using Query Clustering Algorithm

K. Prabha
Research Scholar in Computer Science
Vivekanandha College for Women,
Unjanai, Tiruchengode, India

K. Rajeswari, M.Sc., M.Phil,
Assistant Professor in Computer Science
Vivekanandha College for Women,
Unjanai, Tiruchengode, India

Abstract - An online image retrieval system (similar to Google image search) where users search for images by submitting queries that are made of keywords. The queries formed by the users of a search engine are semantically refined, the keywords representing concise semantics. The aim is to improve user satisfaction by returning images that have a higher probability to be accepted (downloaded) by the user. The assumption is that the users search for images by issuing queries, each query being an ordered set of keywords. The system responds with a list of images. The user can download or ignore the returned images and issue a new query instead. As the user's issues queries and pick images the system annotates the images in an automatic manner and at the same time establishes relevance relations between the keywords. The new method is shown to posses the Query clustering algorithm using agglomerative method for discovering similar queries on a search engine to retrieve the similar images. Experimental result signifies the query clustering to show accuracy. This algorithm gives the precision and recall values, which are helpful in determine the efficiency of search engine queries.

Keywords - Query Clustering, Agglomerative Method, Data Mining, Image Retrieval System.

I. 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, cuts costs, or both. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Clustering is the process of making group of abstract objects into classes of similar objects. A cluster of data objects can be treated as a one group. While doing the cluster analysis, we first partition the set of data into groups based on data similarity and then assign the label to the groups.

Agglomerative method clusters have sub-clusters, which in turn have sub-clusters, etc. Agglomerative clustering starts with every single object in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.

Image retrieval system is an active area to propose a new approach to retrieve images from the large image database. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning', keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Image Retrieval (IR) is one of the most exciting and fastest growing research areas in the field of multimedia technology.

II. ANALYSIS OF RELATED WORK

K. Stevenson and C. Leung (July 2005) proposed text-oriented document searching are relatively mature on the Internet, image searching, which requires much more than text matching, significantly lags behind. We find that current technology is only able to deliver an average precision of around 42% and an average recall of around 12%, while the best performers are capable of producing over 70% for precision and around 27% for recall.

A. Bhattacharya and A.K. Singh (Nov 2005) given a large collection of medical images of several conditions and treatments. We propose to automatically develop a visual vocabulary by breaking images into n × n tiles and deriving key tiles ("ViVos") for each image and condition. We experiment with numerous domain-independent ways of extracting features from tiles (color histograms, textures, etc.), and several ways of choosing characteristic tiles (PCA, ICA).

J. Li and J. Wang (2006) developing effective methods for automated annotation of digital pictures continues to challenge computer scientists. These new techniques serve as the basis for the automatic linguistic indexing of pictures - real time (ALIPR) system of fully automatic and high-speed annotation for online pictures. In particular, the D2-clustering method, in the same spirit as K-Means for vectors, is developed to group objects represented by bags of weighted vectors.

D.M. Blei and A.Y. Ng, and M.I. Jordan (2003) understanding how topics within a document evolve over its structure is an interesting and important problem. In this paper, we address this problem by presenting a novel variant of Latent Dirichlet Allocation (LDA): Sequential LDA (SeqLDA). This variant directly considers the underlying sequential structure, i.e., a document consists of multiple segments (e.g., chapters, paragraphs), each of which is correlated to its previous and subsequent segments.

Z. Guo, S. Zhu, Y. Chi, Z. Zhang, and Y. Gong (2009) proposed document similarity measures are required for a variety of data organization and retrieval tasks including document clustering, document link detection, and query-by-example document retrieval. In this paper we examine existing and novel document similarity measures for use with spoken document collections processed with automatic speech recognition (ASR) technology.

Konstantinos A. Raftopoulos (Feb 2013) proposed Markovian Semantic Indexing (MSI), is presented in the context of an online image retrieval system. Assuming such a system, the users' queries are used to construct an Aggregate Markov Chain (AMC) through which the relevance between the keywords seen by the system is defined. The users' queries are also used to automatically annotate the images. A stochastic distance between images, based on their annotation and the keyword relevance captured in the AMC is then introduced.

III. QUERY CLUSTERING

Query clustering is a technique for discovering similar queries on a search engine. Also it is a class of techniques aiming at grouping users' semantically related, not syntactically related, and queries in a query repository, and accumulated with the interactions between users and the system.

Query clustering algorithm choosing an appropriate clustering algorithm is also very critical to the effectiveness and efficiency of the query clustering process. While choosing the clustering algorithm, the following things must be kept in mind:

- The algorithm should be capable of handling a large data set within reasonable time and space constrained.
- The algorithm should be easily extended to cluster new queries incrementally.
- The algorithm should not require manual setting of the resulting form of the clusters.

3.1 Agglomerative Method

Agglomerative method works by grouping the data one by one on the basis of the nearest distance measure of all the pair wise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed. Single linkage, complete linkage, average linkage and centroid distance between two points, grouping the data until one cluster is remaining.

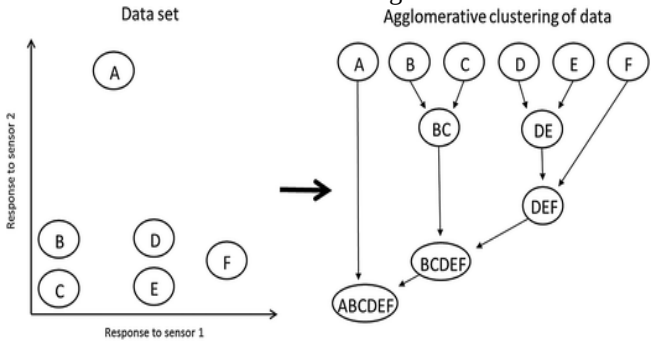

Figure 1 Agglomerative Clustering

Algorithm: Agglomerative clustering

Agglomerative clustering is starting out with *n* cluster for *n* data points, that is, each cluster consisting of a single data points.

Input: Number of cluster.

Output: One line per cluster which contains the points belonging to that cluster.

Method:

Step-1 Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

Step-2 Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

Step-3 Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].

Step-4 Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r, s) and old cluster (k) is defined in this way: d [(k), (r, s)] = min (d [(k), (r)], d [(k), (s)]).

Step-5 If all the data points are in one cluster then stop, else repeat from step 2.

Advantages:

- It can produce an ordering of the objects, which may be informative for data display.
- Smaller clusters are generated, which may be helpful for discovery.

IV. EXPERIMENTS AND RESULTS

Query Clustering Process

Query clustering is a technique for discovering similar queries on a search engine. In this work provides an overview of algorithms which are helpful in search engine optimization.

To search for images submit the queries that are made up of keywords. Generate possible Meaning of keywords using WordNet Dictionary. Generate Multiple Synonyms of the specified query using WordNet Dictionary. Retrieve tags according to synonyms applied on queries. Show in table 1.

Table 1 Flower Images Data's

Image name	Keyword	Tag id
451.jpg	Endosperm	Tag 1
452.jpg	Bartonia	Tag 2
488.jpg	Calla	Tag 3
964.jpg	Style	Tag 4
496.jpg	Spike	Tag 5
479.jpg	Wind flower	Tag 7

4.1 Distance Calculation:

Consider the Figure 2 to construct the distance value. Distance between a point X (X1, X2, etc.) and a point Y (Y1, Y2, etc.)**.**

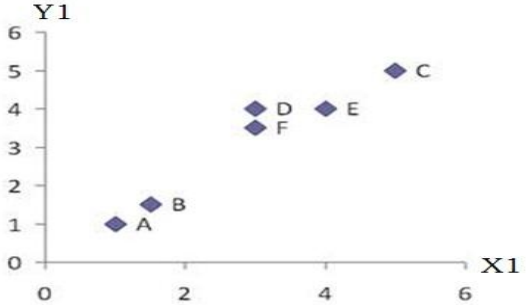

Figure 2 Graphs for Distance

Data value

	A	B	C	D	E	F
X1	1	1.5	5	3	4	3
X2	1	1.5	5	4	4	3.5

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Using this formula to find the distance calculation in table 2

$d_{AB} = ((1-1.5)^2 + (1-1.5)^2)^{1/2} = 0.5^{1/2} = 0.7071$

$d_{DF} = ((3-3)^2 + (4-3.5)^2)^{1/2} = 0.5$

Table 2 Distance Calculation

Dist	A	B	C	D	E	F
A	0	0.7	5.6	3.6	4.2	3.2
B	0.7	0	4.9	2.9	3.5	2.5
C	5.6	4.9	0	2.2	1.4	2.5
D	3.6	2.9	2.2	0	1	0.5
E	4.2	3.5	1.4	1	0	1.1
F	3.2	2.5	2.5	0.5	1.1	0

4.2 Matrix Calculation:

Construct the distance matrix using distance value from the Euclidean distance is shown in table 3.

Table 3 Distance matrix

Dist	A	B	C	D	E	F
A	0	0.7	5.6	3.6	4.2	3.2

B	0.7	0	4.9	2.9	3.5	2.5
C	5.6	4.9	0	2.2	1.4	2.5
D	3.6	2.9	2.2	0	1	0.5
E	4.2	3.5	1.4	1	0	1.1
F	3.2	2.5	2.5	0.5	1.1	0

From the table 3 (0.5) is minimum value. Merge two closest clusters is shown in table 3.1

Table 3.1 Reduced the Matrix 1

Dist	**A**	**B**	**C**	**D,F**	**E**
A	0	0.7	5.6	-	4.2
B	0.7	0	4.9	-	3.5
C	5.6	4.9	0	-	1.4
D,F	-	-	-	0	-
E	4.2	3.5	1.4	-	0

From table 3.1 update distance matrix. To find the minimum distance of $d_{(D, F)}$ is show in table 3.2

$d_{(D,F) \rightarrow A}=\min(d_{DA},d_{FA})=\min(3.6,3.2)=3.2$

$d_{(D,F) \rightarrow B}=\min(d_{DB},d_{FB})=\min(2.9,2.5)=2.5$

$d_{(D,F) \rightarrow C}=\min(d_{DC},d_{FC})=\min(2.2,2.5)=2.2$

$d_{(D,F) \rightarrow E}=\min(d_{DE},d_{FE})=\min(1,1.12)=1$

Table 3.2 Find Minimum Value

Dist	**A**	**B**	**C**	**D,F**	**E**
A	0	0.7	5.6	3.2	4.2
B	0.7	0	4.9	2.5	3.5
C	5.6	4.9	0	2.2	1.4
D,F	3.2	2.5	2.2	0	1
E	4.2	3.5	1.4	1	0

From the table 3.2 (0.7) is minimum value. Merge two closest clusters is shown in table 3.3

Table 3.3 Reduced the Matrix 2

Dist	**A,B**	**C**	**D,F**	**E**
A,B	0	-	-	-
C	-	0	2.2	1.4
D,F	-	2.2	0	1
E	-	1.4	1	0

From table 3.3 update distance matrix. To find the minimum distance of $d_{(A,B)}$ is show in table 3.4

$d_{C \rightarrow (A,B)}=\min(d_{CA},d_{CB})=\min(5.6,4.9)=4.9$

$d_{(D,F) \rightarrow (A,B)}=\min(d_{DA}, d_{DB,} d_{FA,} d_{FB})=\min(3.6,2.9,3.2,2.5)=2.5$

$d_{E \rightarrow (A,B)}=\min(d_{EA},d_{EB})=\min(4.2,3.5)=3.5$

Table 3.4 Find Minimum Value

Dist	**A,B**	**C**	**D,F**	**E**
A,B	0	4.9	2.5	3.5
C	4.9	0	2.2	1.4
D,F	2.5	2.2	0	1
E	3.5	1.4	1	0

From the table 3.4 (1) is minimum value. Merge two closest clusters is shown in table 3.5

Table 3.5 Reduced the Matrix 3

Dist	**A,B**	**C**	**(D,F),E**
A,B	0	4.9	2.5
C	4.9	0	1.4
(D,F),E	2.5	1.4	0

From the table 3.5 (1.4) is minimum value. Merge two closest clusters is shown in table 3.6

Table 3.6 Reduced the Matrix 4

Dist	**A,B**	**((D,F),E),C**
A,B	0	2.5
((D,F),E),C	2.5	0

Final Result

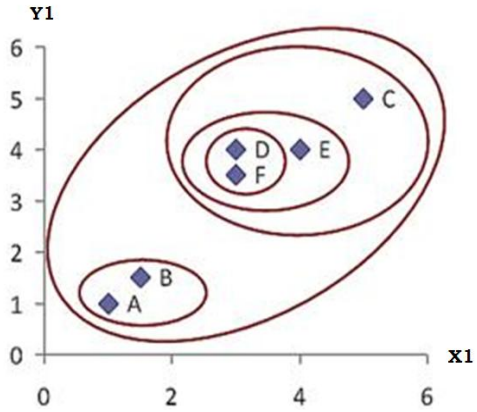

Figure 3 Graphs for Final Cluster

4.3 Probability Calculation

Using below formula to find the probability value in table 4. Calculate the probability of images to be retrieved using the query and matching with the generated synonyms. It is calculated from the distance of query with keywords.

$$\text{Probability of an event happening} = \frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}}$$

Example

0.2, 0.4, 0.3, 0.5, 0.6, 0.2

Probability of 0.2 = 2 /6 = 0.333

Table 4 Calculate the Probability

Image Name	**Keyword**	**Probability Value**
451.jpg	Endosperm	0.3
452.jpg	Bartonia	0.1
488.jpg	Calla	0.1
964.jpg	Style	0.1
496.jpg	Spike	0.2
444.jpg	Blossome	0.2

4.4 Search Result

Download the list of Images retrieved. Show in table 5.

Table 5 Search Result

Precision

Precision is defined as the ratio of the number of words that correctly retrieved to the total number of words retrieved in every image search.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}}$$

Recall

Recall is the ratio of the number of words that retrieved correctly to the number of words.

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total relevant images in collection}}$$

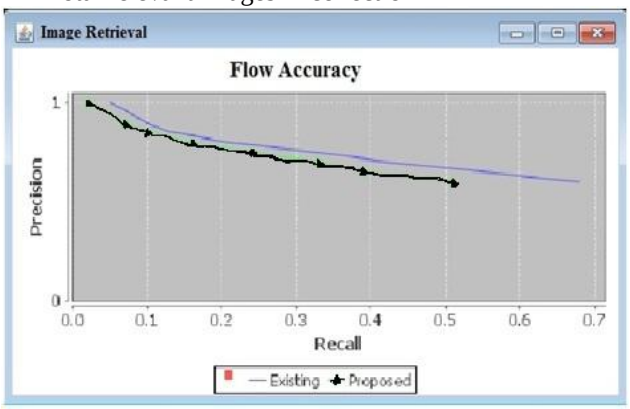

Figure 4 Comparison Graph of Existing and Proposed System

V. CONCLUSION

The Query clustering algorithm using agglomerative method helps to retrieve the images in large database. The effectiveness of the proposed framework compared with other presented retrieval algorithms. This algorithm gives more accurate results than the query mining algorithm. Experimental results show that user profiles which capture both the user's positive and negative preferences perform the best among all of the profiling strategies studied. This algorithm gives the better precision and recall values, which are helpful in determine the efficiency of search engine queries.

VI. FUTURE WORK

In future work it can be done for voice recording and voice searching. And also in future, using ranking based image retrieval (RBIR) method can be done to provide ranking for each user query. Many times to give same query for image retrieval, the query will arrange priority based and the query easily retrieve the images.

REFERENCES

[1] User Interfaces in C#: Windows Forms and Custom Controls by Matthew MacDonald.
[2] Applied Microsoft® .NET Framework Programming (Pro-Developer) by Jeffrey Richter.
[3] Practical .Net2 and C#2: Harness the Platform, the Language, and the Framework by Patrick Smacchia.
[4] Data Communications and Networking, by Behrouz A Forouzan.
[5] Computer Networking: A Top-Down Approach, by James F. Kurose.
[6] Gabriel R. Bitran and Rene Caldentey. An overview of pricing models for revenue management.
[7] N. Bruno and S. Chaudhuri. An online approach to physical design tuning.
[8] Xi-Ren Cao, Hong-Xia Shen, R. Milito, and P. Wirth. Internet pricing with a game theoretical approach: concepts and examples.

[9] Ch Chen, Muthucumaru Maheswaran, and Michel Toulouse. Supporting co-allocation in an auctioning-based resource allocator for grid systems.
[10] S. Choenni, H. M. Blanken, and T. Chang. On the selection of secondary indices in relational databases.
[11] D. Dash, Y. Alagiannis, C. Maier, and A. Ailamaki. Caching all plans with one call to the optimizer.
[12] Debabrata Dash, Verena Kantere, and Anastasia Ailamaki. An economic model for self-tuned cloud caching.
[13] Carsten Ernemann, Volker Hamscher, and Ramin Yahyapour. Economic scheduling in grid computing.
[14] Guillermo Gallego and Garrett van Ryzin. Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons.
[15] A. Ghose, V. Choudhary, T. Mukhopadhyay, and U. Rajan. Dynamic pricing: A strategic advantage for electronic retailers.

