Analysis and performance optimization of k - Nearest Neighbor approach for movie queries

Gaurav Singh
Harsh Shah
Manjari Akella

The Idea

- To implement a movie query engine that allows users to find a movie of interest by starting the query from a movie that they have already seen
- Example of a query –

"When Harry met Sally" – Romance + Thriller = "From dusk till dawn"

Data and Representation

- Custom collected data set crawled from
  - Freebase
  - MovieLens 10M
- Feature Vector
  - 3484 x 276 vector (row = movies, column = genres)
  - Each genre has a value 0 or 1
    - 0 if movie is categorized in that genre
    - 1 if it isn't

ML Algorithm

- kNN – star of the show
- Optimization of KNN –
  - Pre-clustering using k-means
  - kd-trees
    - Space partitioning data structure for organizing points in a k-dimensional space
  - Locality Sensitivity Hashing(LSH)
    - method of performing probabilistic dimension reduction of high-dimensional data

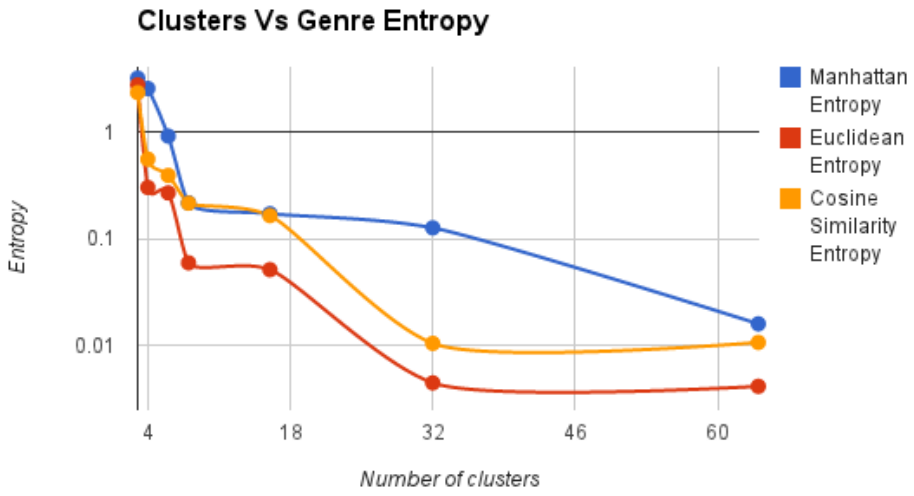
K-d Trees vs. LSH

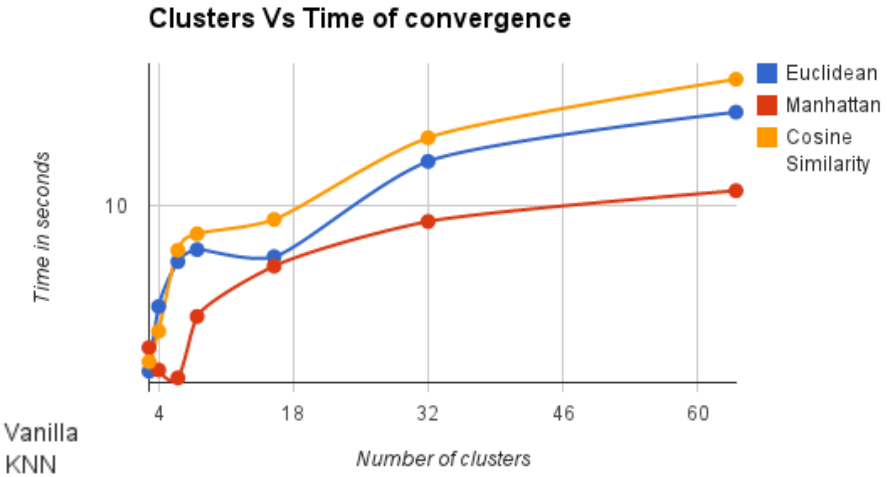
- If the dimensionality of the data is k, a k-d tree works well only if the number of data points is more than 2^k.
- Otherwise, we end up testing nearly all the nodes in the dataset making the complexity O(n) instead of the desired O(log(n)).
- LSH is expected to perform much better when the number of dimensions is large, like in our case.

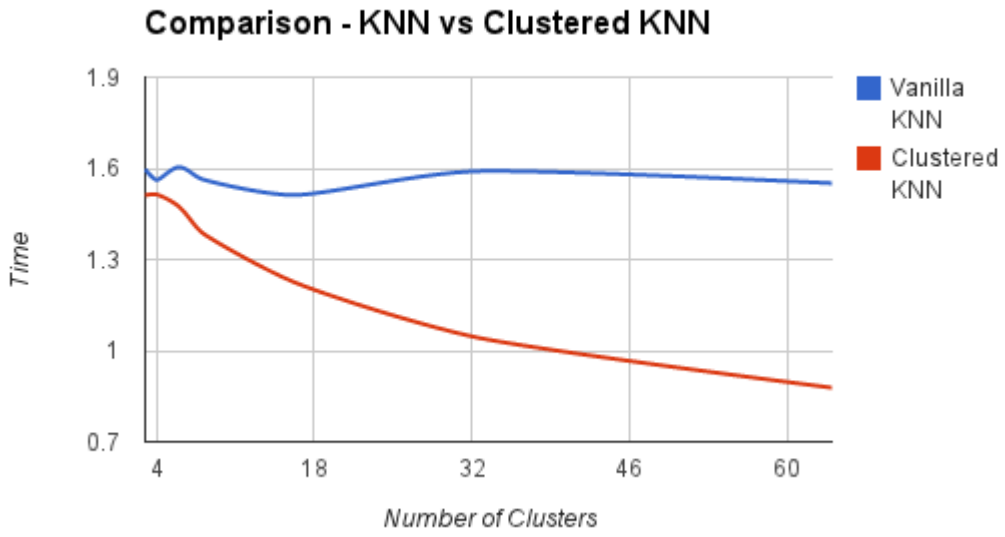
Results

- Tasks Completed:
  - kNN classification
  - K Means clustering of the data set using various distance metrics and values of k
  - kNN Classification on clustered data
- Tasks Remaining:
  - Comparison with k-d trees or LSH

Results


Clusters Vs Genre Entropy
Entropy
1
0.1
0.01
4
18
32
46
60
Number of clusters
Manhattan Entropy
Euclidean Entropy
Cosine Similarity Entropy


Clusters Vs Time of convergence
Time in seconds
10
4
18
32
46
60
Number of clusters
Euclidean
Manhattan
Cosine Similarity


Comparison - KNN vs Clustered KNN
Time
1.9
1.6
1.3
1
0.7
4
18
32
46
60
Number of Clusters
Vanilla KNN
Clustered KNN

Project Takeaways

- For us:
  - MQL (Metaweb Query Language)/Freebase – A new API and data repository
  - Ability to assess trade-offs and choose appropriate optimization methods based on dataset in hand (i.e. understanding of scenarios where certain algorithms fail)
  - An in-depth understanding of KNN
- For you:
  - A new movie query engine (Yes, we plan to put it online with an interactive interface)

References

- Locality-Sensitive Hashing for Finding Nearest Neighbors [Malcolm Slaney and Michael Casey]
- Multidimensional Binary Search Trees Used for Associative Searching [Jon Louis Bentley, Stanford University]
- Our best friend, www.google.com

Thank You

Questions ?