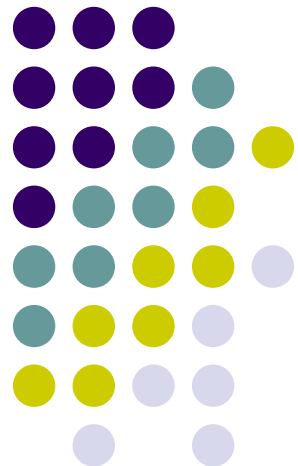
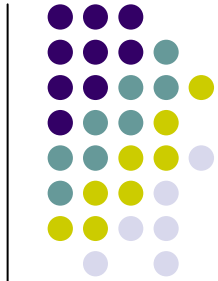
Similarity Search: A Matching Based Approach

Anthony Kum Hoe Tung
National University of Singapore
atung@comp.nus.edu.sg



Joint Work with
Rui Zhang — University of Melbourne
Nick Koudas — University of Toronto
Beng Chin Ooi — National University of Singapore

About my research interest



Techniques

- Association rules discovery
- Sequential Pattern Discovery
- Cluster analysis
- Outlier Detection
- Classifier Building
- Data Cube/Data Warehouse Construction
- Visualization ...

Applications

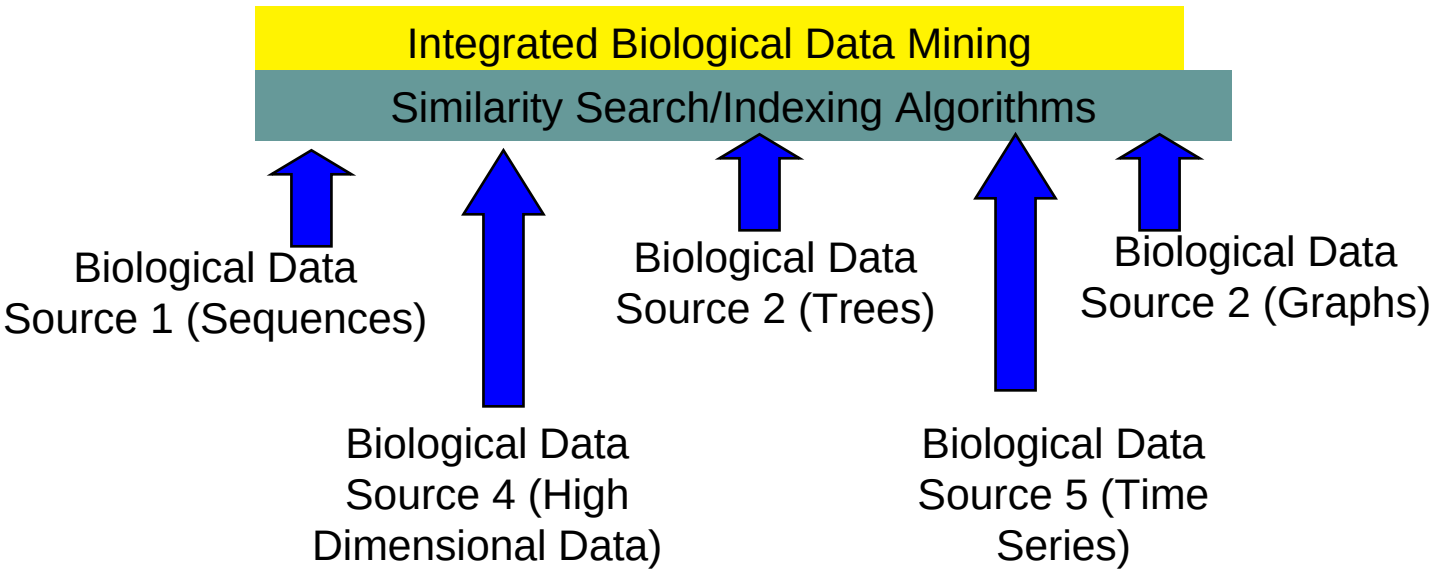
- Spatial Data Mining
- Biological Data Mining
- Personal Information Management

Principles/ Foundation

- **Database Technology:**
  - **Indexing, Compression, Data Structure**
- **AI/ Machine Learning**
- **Statistics**
- **Information Theory**
- **Theoretical CS :**
  - **Approximate, Random, Online Algorithms**
- **Mathematical Programming**
- **Computational Geometry …**

About My Research (II)

- Why similarity search?
  - Distance function is a core issue in data mining. Without a good way to judge what is similar/dissimilar, no effective mining can be done
- As such, my research focus for the past few years have been focus on similarity search for high dimensional data, sequences, trees and graphs
- Eventually, mining algorithms can be developed on top of them for application such as integrated mining of biological data, personal information management etc



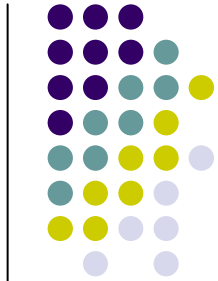
About My Research (III)

- Sequences
  - "Indexing DNA Sequences Using q-grams". **Best Paper Award**. DASFAA 2005
  - Indexing Mixed Types for Approximate Retrieval, VLDB'06
- Trees
  - "Similarity Evaluation on Tree-structured Data". SIGMOD'05
- Graphs
  - On going!
- Time Series
  - SpADe: On Shape-based Pattern Detection in Streaming Time Series." ICDE'07
- High Dimensional Data
  - LDC: Enabling Search By Partial Distance In A Hyper-Dimensional Space.ICDE'2004
  - Similarity Search: A Matching Based Approach, VLDB'06
  - Finding k-Dominant Skylines in High Dimensional Space, SIGMOD'06

About My Research(IV)

- Research mentality very much affect by the following two quotes
- “We can’t choose reviewers but we can choose to write good papers”
  - Raymond Ng, UBC
- “If you think your idea is going to be easily published by someone else tomorrow, then probably it is not too innovative”
  - Philip Long, Google

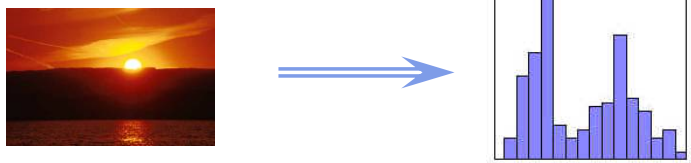
Outline



- Traditional approach to similarity search
- Deficiencies of the traditional approach
- Our proposal: the n-match query
- Algorithms to process the n-match query
- Experimental results
- Future work and Conclusion

Similarity Search : Traditional Approach

- **Objects represented by multidimensional vectors**



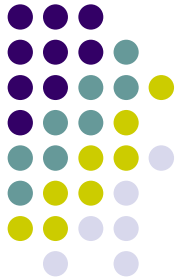
Elevation	Aspect	Slope	Hillshade (9am)	Hillshade (noon)	Hillshade (3pm)	…
2596	51	3	221	232	148	
…						

- **The traditional approach to similarity search: kNN query**

  Q = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)

ID	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	Dist
P1	1.1	1	1.2	1.6	1.1	1.6	1.2	1.2	1	1	**0.93**
P2	1.4	1.4	1.4	1.5	1.4	1	1.2	1.2	1	1	**0.98**
P3	1	1	1	1	1	1	2	1	2	2	**1.73**
P4	20	20	21	20	22	20	20	19	20	20	**57.7**
P5	19	21	20	20	20	21	18	20	22	20	**60.5**
P6	21	21	18	19	20	19	21	20	20	20	**59.8**

Deficiencies of the Traditional Approach

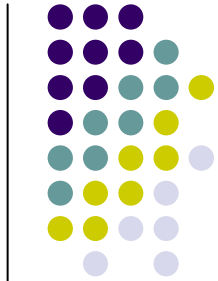


- **Deficiencies**
  - Distance is affected by a few dimensions with high dissimilarity
  - Partial similarities can not be discovered
- **The traditional approach to similarity search: kNN query**

  Q = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)

ID	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	Dist
P1	1.1	**100**	1.2	1.6	1.1	1.6	1.2	1.2	1	1	**99.0**
P2	1.4	1.4	1.4	1.5	1.4	**100**	1.2	1.2	1	1	**99.0**
P3	1	1	1	1	1	1	2	**100**	2	2	**99.0**
P4	20	20	21	20	22	20	20	19	20	20	**57.7**
P5	19	21	20	20	20	21	18	20	22	20	**60.5**
P6	21	21	18	19	20	19	21	20	20	20	**59.8**

Thoughts



- Aggregating too many dimensional differences into a single value result in too much information loss. Can we try to reduce that loss?
- While high dimensional data typically give us problem when in come to similarity search, can we turn what is against us into advantage?
- Our approach: Since we have so many dimensions, we can compute more complex statistics over these dimensions to overcome some of the “noise” introduce due to scaling of dimensions, outliers etc.

The *N*-Match Query : Warm-Up