Expert knowledge similarity measurement using network graph edit distance

Qingyun Liao

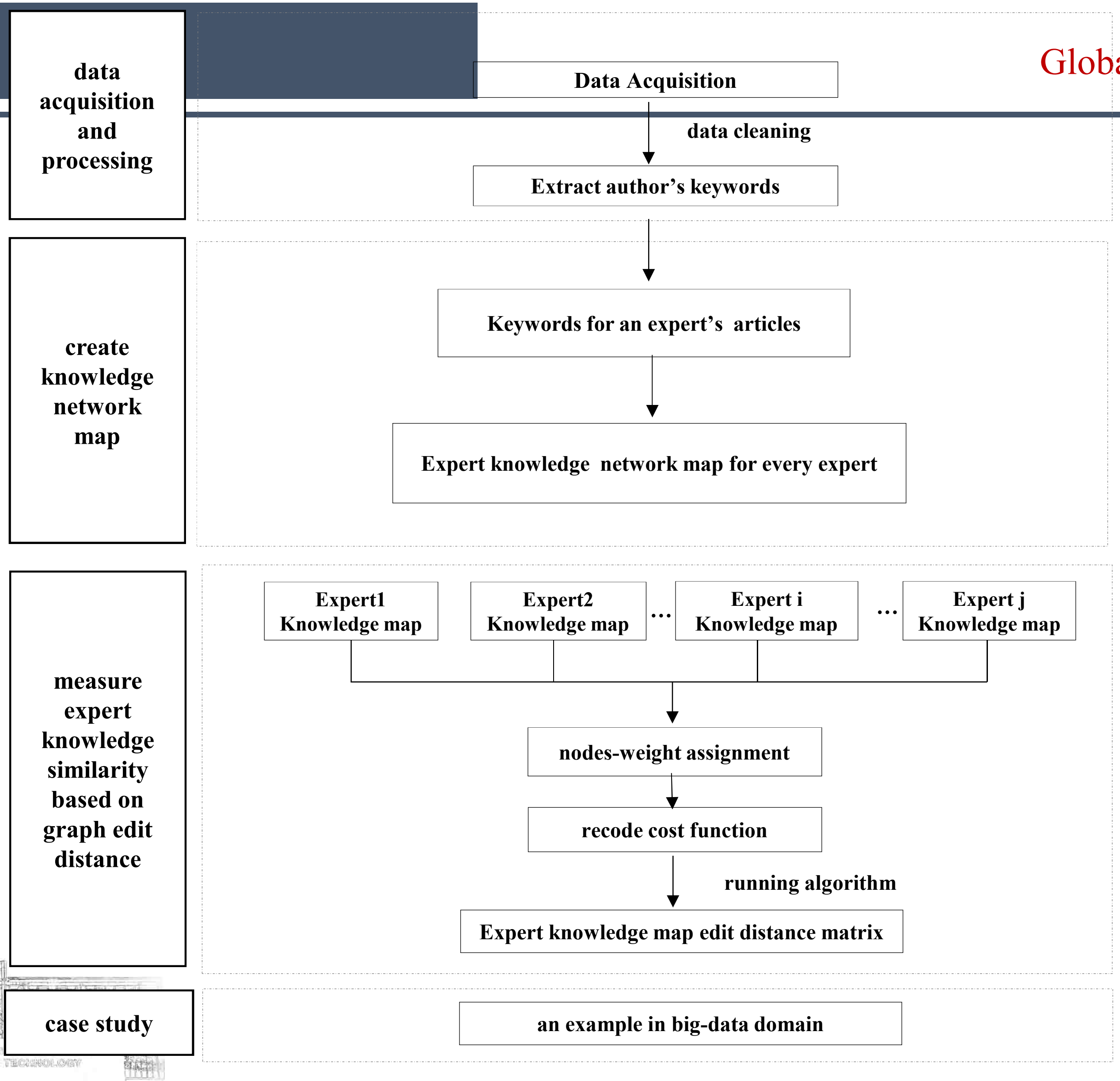
School of Management and Economics, Beijing Institute of Technology,
Beijing, China

Co-author
Qi lv , Ying Wang, Xuefeng Wang, Dong Wan

Background

Drawbacks	Improvements
Hard to understand: the description of expert (vector) is hard to understand for people who are not familiar with multidimensional vector .	Expert knowledge graph: Graph is an vivid and easy to understand expression, it can be wildly applied.
Information omitted: when we use cosine similarity to calculate the similarity of the vector, some important information is omitted	Expert knowledge graph , not only keywords but also includes knowledge construction/distribution.
Sparse matrix: the more experts, the larger keywords matrix is and the smaller the proportion of keywords it involves for each expert. It is likely that each expert will use a vector of thousands of dimensions, but only a few is non-zero, and sparseness will affect the confidence of the results.	Calculate Expert-pair matrix In this article, we construct keywords matrix for expert-pair

Methods



data acquisition and processing

Data Acquisition

data cleaning

Extract author’s keywords

create knowledge network map

Keywords for an expert’s articles

Expert knowledge network map for every expert

measure expert knowledge similarity based on graph edit distance

Expert1 Knowledge map

Expert2 Knowledge map

…

Expert i Knowledge map

…

Expert j Knowledge map

nodes-weight assignment

recode cost function

running algorithm

Expert knowledge map edit distance matrix

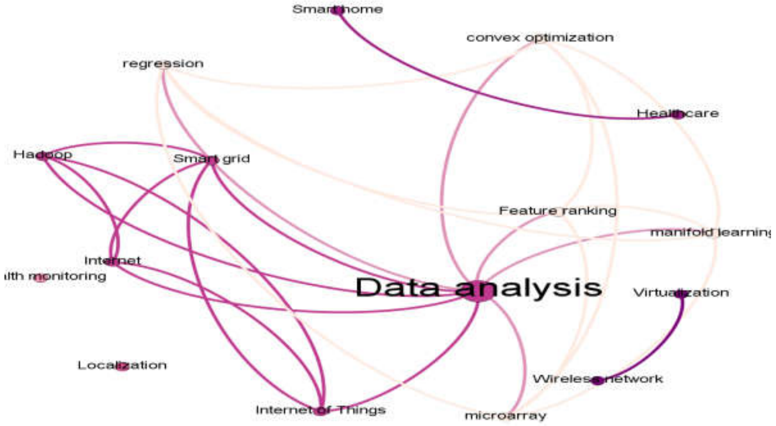
case study

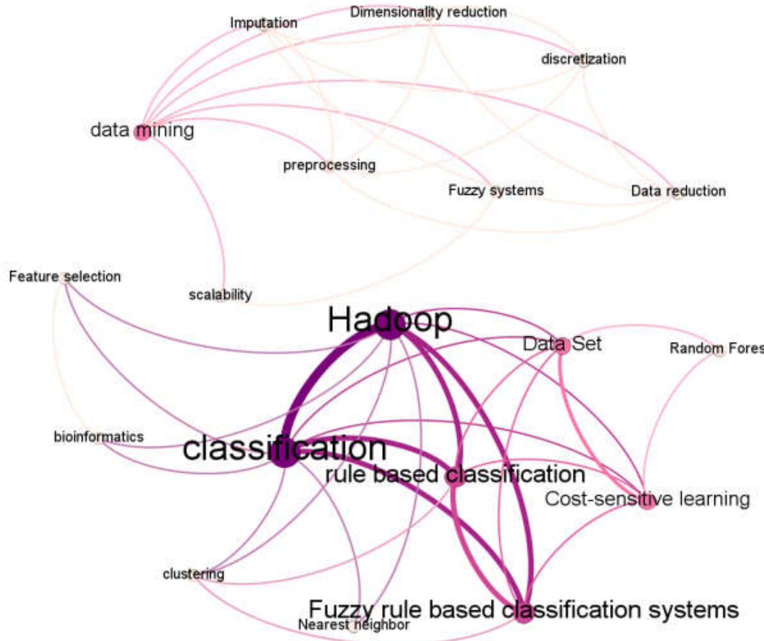
an example in big-data domain

Methods

➢Representing Expertise within Expert Knowledge Map

- ➢we introduce the concept of expert knowledge map to represent expertise.
- ➢Focusing on a domain that we want to study.
- ➢searching in Web of Science and download.
- ➢use Vantage Point (VP) to clean raw data.
- ➢Generate "authors" to "keywords" matrix.
- ➢import "authors" to "keywords" matrix in gephi.
- ➢export expert's knowledge maps for each author.





Methods

Brief view of Graph Edit Distance

Definitions of the graph and the attributed graph

- A graph is denoted by G = (V, E)
- If both nodes and edges in a graph have attributes, the graph is an attributed graph denoted by G = (V, E, α, β), *where* $\alpha: V \rightarrow L_N$ *and* $\beta \rightarrow L_E$ *are node and edge labeling functions.* L_N *and* L_E *are restricted to labels consisting of fixed-size tuples, that is,* $L_N = R^p, L_E = R^q, p, q \in N \cup \{0\}$.

Definition and computation of Graph Edit Distance

- source graph G_1 = (V_1, E_1, α_1, β_1) and the target graph G_2 = (V_2, E_2, α_2, β_2), to transform G_1 into G_2 using some edit operations.
- This method can cope with arbitrary labels on both nodes and edges as well as with directed or undirected edges.
- 3 Basic edit operations ： insertions, deletions, and substitutions， suitable for both nodes and edges.

Methods

- Cost function
- one introduces a cost c(e) for every edit operation e, measuring the strength of the edit operation.
  - between two similar graphs, there should exist an inexpensive edit path, representing low-cost operations
  - for dissimilar graphs an edit path with high cost is needed.
- Graph edit distance method allow users to define cost function, which makes it one of the most flexible dissimilarity models available for graphs.

Methods

- **Example of graph edit distance**
- Given two graphs， name every node like “u1”, “ v2”. Number in circle is dot attribute label.
- G1={{u1,u2,u3,u4},{(u1,u2),(u2,u3),(u3,u4),(u2,u4)},{u1:1, u2:2, u3:1, u4:3}}

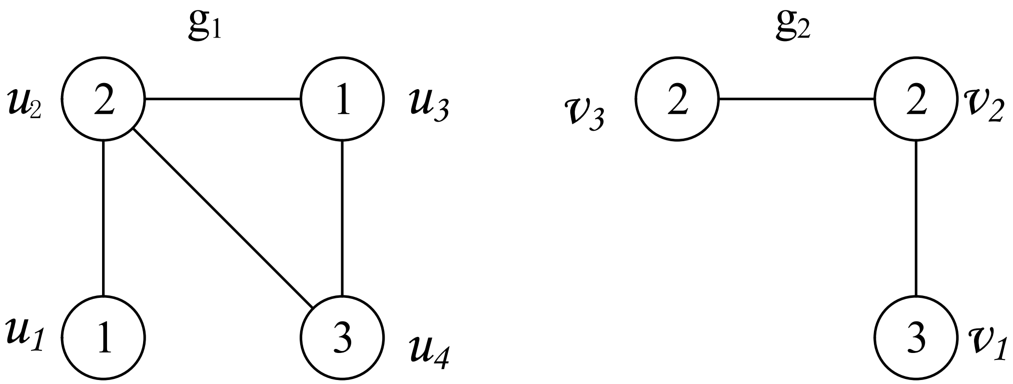


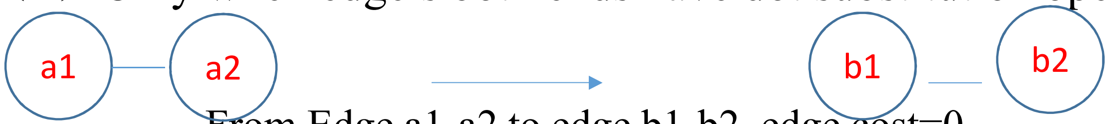
Figure 2. examples of source-graph and target-graph

Methods

- Basic rules
- 1.take all nodes distortion in to account first
  - because edge distortion are usually accompanied by nodes distortion.
- 2.When it comes to nodes edit operations, consider node-substitutions firstly, then insertions and deletions.
- 3.For every edit operation there is a defined cost computation: node-insertion and node-deletion are usually defined as a constant value while node-substitution value varies by the difference between nodes attribute label.
  - example: In figure 2, if we change node u1 to v1, the cost is 2, that is $cost = |1 - 3| = 2$. Certainly, value of node attribute can be define by users.

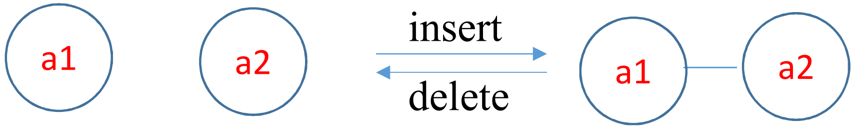
Methods

- 4.edge deletion/insertion/disappearance
- Edge edit operation has a constant cost.
  - （1）Only when edge’s both ends have dot substitution operation the edge cost=0;

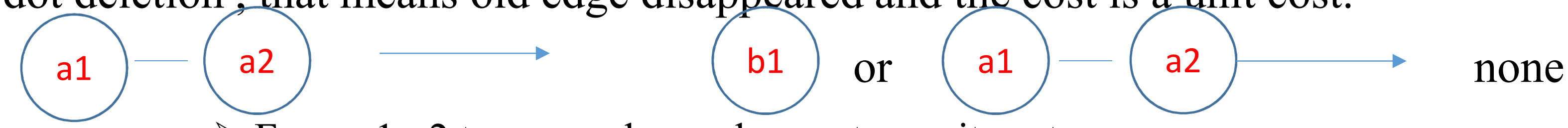


From Edge a1-a2 to edge b1-b2, edge cost=0

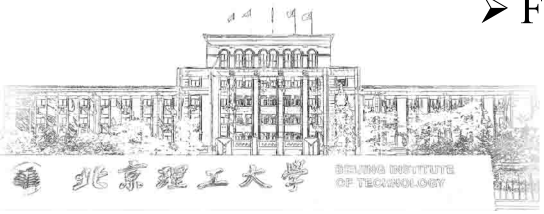
  - （2）If we insert/delete an edge ,edge cost= a unit cost



From none to edge a1-a2/ from edge a1-a2 to none edge cost= a unit cost

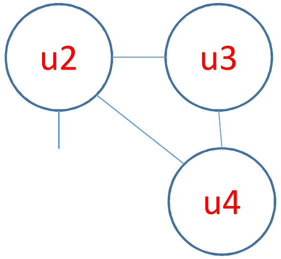
  - （3）If one end have dot deletion operation , the other have dot substitution or both ends have dot deletion , that means old edge disappeared and the cost is a unit cost.
  - 

- From a1-a2 to none edge , edge cost=a unit cost

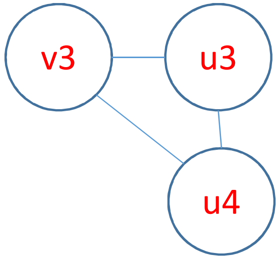


Methods

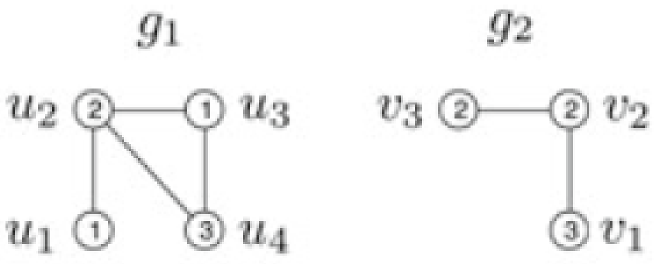
- Here is an edit path λ (g1, g2) between two undirected and unlabeled graphs g1 and g2 is illustrated. Obviously, this edit path is defined by
- $\lambda = \{(u_1 \rightarrow \varepsilon), (u_2 \rightarrow v_3), (u_3 \rightarrow v_2), (u_4 \rightarrow v_1)\}$.
- This particular edit path implies the following edge edit operations:
- $\{((u_1, u_2) \rightarrow \varepsilon), ((u_2, u_3) \rightarrow (v_3, v_2)), ((u_3, u_4) \rightarrow (v_2, v_1)), ((u_2, u_4) \rightarrow \varepsilon)\}$.

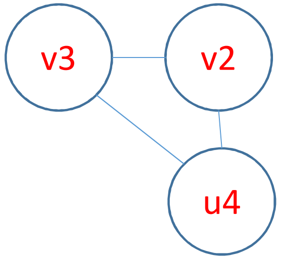


Delete u1,
Cost = 1

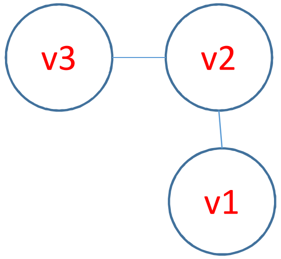


1.Substitute u2 to v3, cost of substitution=|2-2|=0
2.Edge u1-u2 disappear cost=1.
3.Total cost=1+0=1

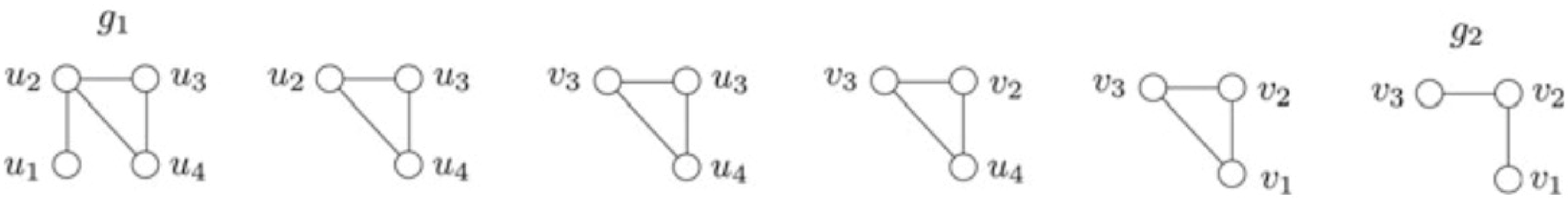




Substitute u3 to v2, cost of substitution=|2-1|=1
Cost = 1



1.Substitute u4 to v1, cost of substitution=|3-3|=0
2.Delete edge u2-u4,edge cost=1
3.Total cost= 0+1=1



Methods

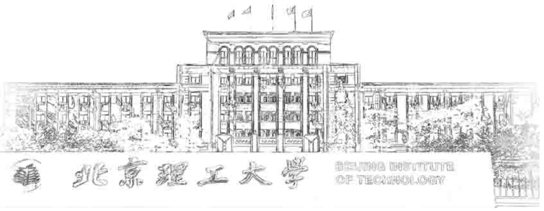
- **Computation of Graph Edit Distance**
- **Computation of Graph Edit Distance can be divided in 3 steps.**
  - **Step 1: code graphs in GXL(Graph Exchange Language) files. The point of GXL files is to define attributes of nodes, including node name (keyword), node records and edge weight at least.**

  - **Step 2: define core parameter. We need to define a cost function of node substitution, a constant cost value of node insertion as well as node deletion , a constant cost value of edge substation , a constant cost value of edge insertion and deletion.**

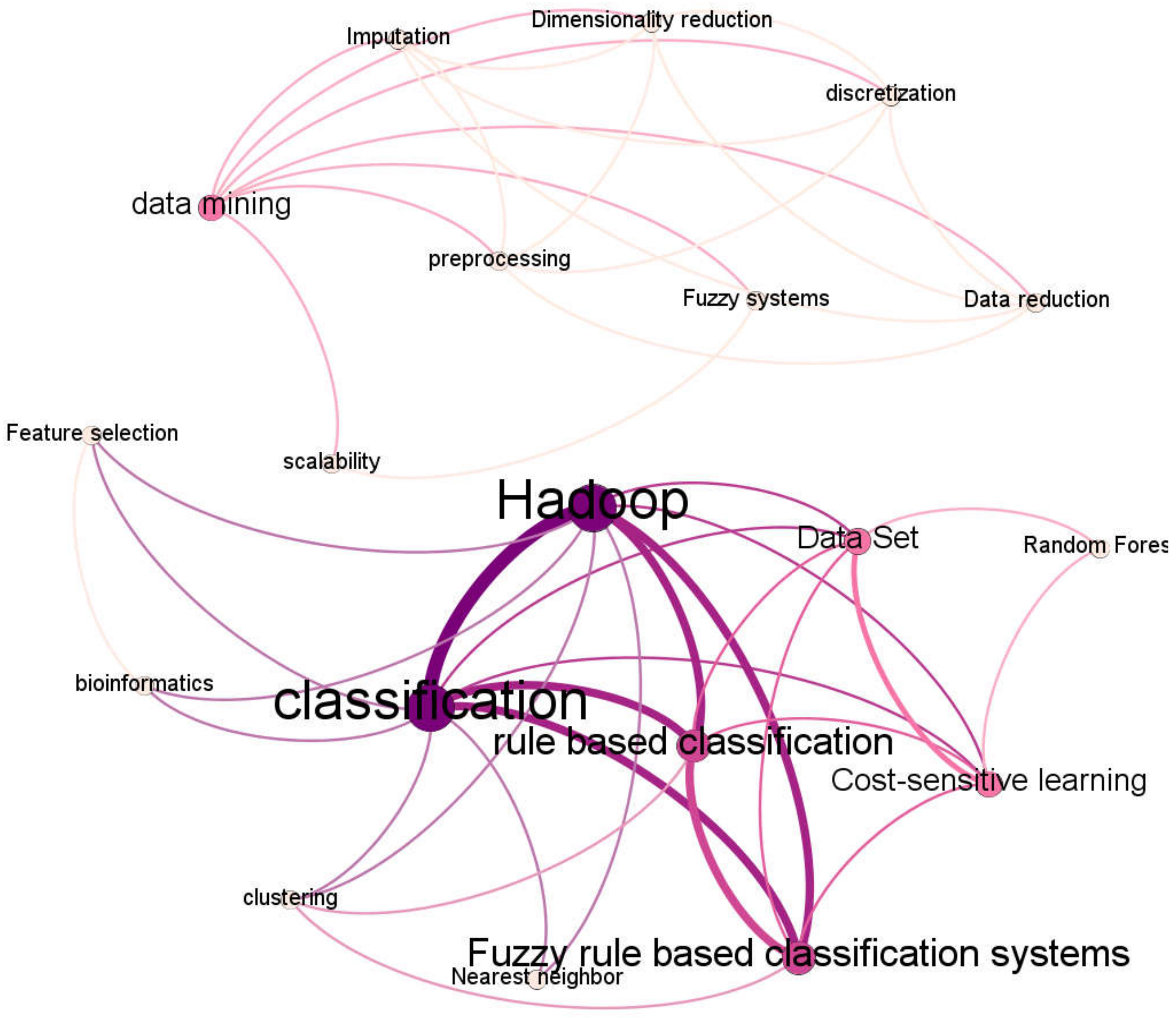
  - **Step 3:calculate graph edit distance.**

Case study

- We choose **big data domain** to do case study.
- **Search strategy as follows:**
- TS= ((“Big Data” or Bigdata) OR (((Big Near/1 Data or Huge Near/1 Data) OR “Massive Data” OR “Huge Information” OR “Big Information” OR “Large-scale Data” OR “Semi-Structured Data” OR “Unstructured Data”) AND (“analytic*” OR “analyz*” OR “analys*”))).What’s more set Time windows from 2008 to 2016, select SCI-E database, and choose article-type records. Finally, we get 4867 records.
- From raw data, we get 11974 author keywords， After data cleaning we finally get 1110 keywords.

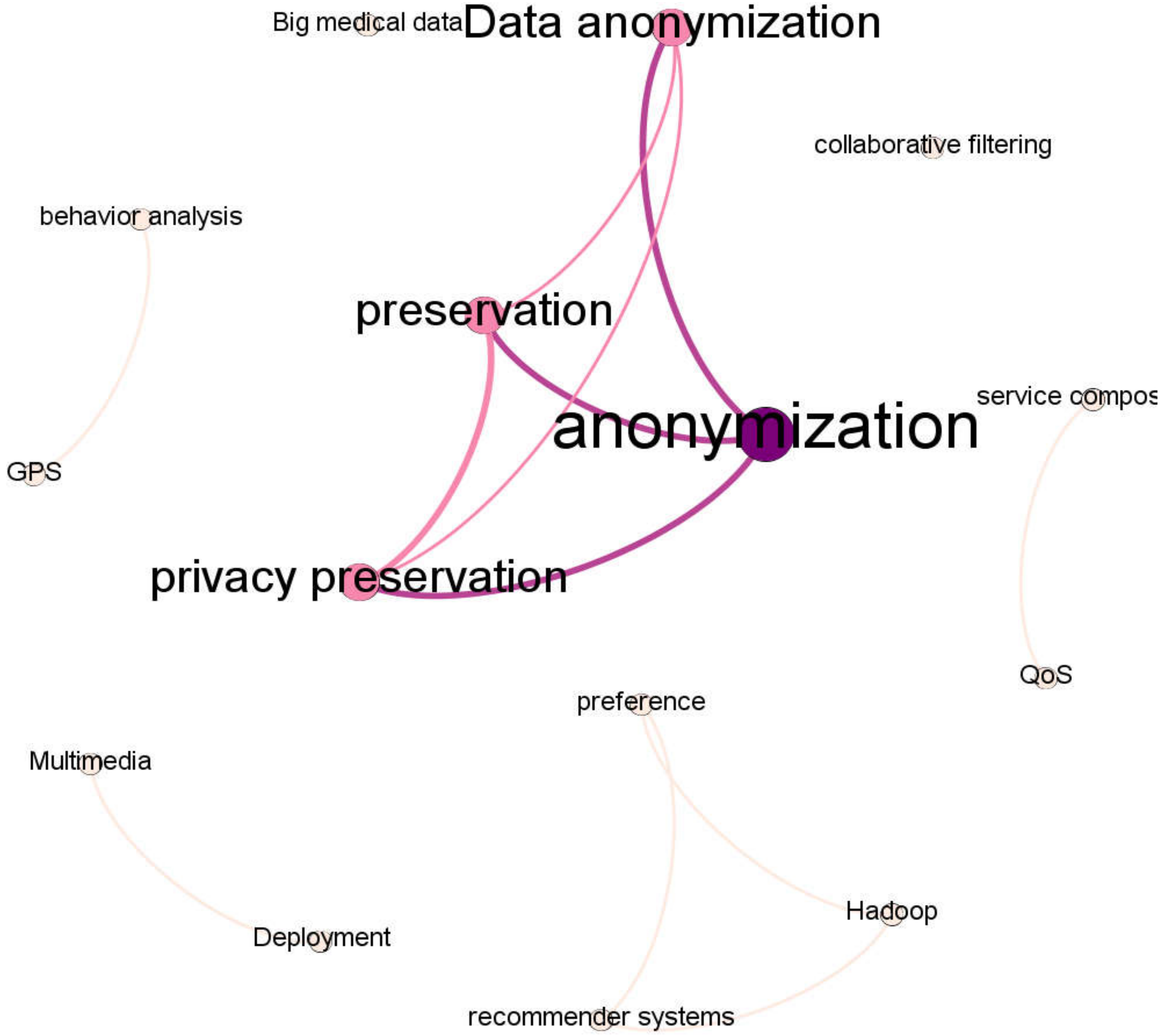


- ***Figure 3. Herrera, Francisco expert knowledge graph***



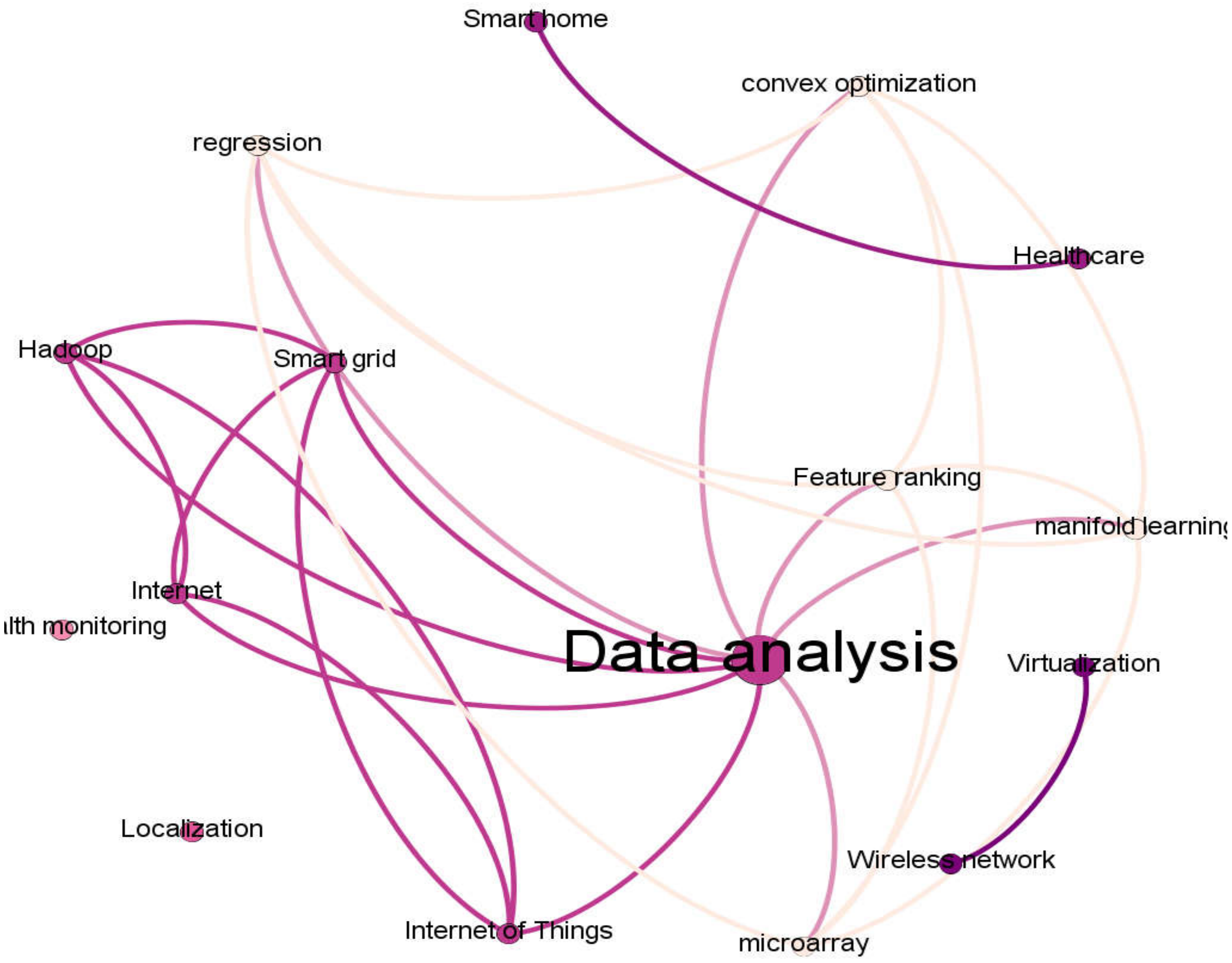
Herrera, Francisco is form Universidad de Granada in spain, majoring computer science and artificial intelligence

Figure 4. Wanchun. Dou's expert knowledge graph



Wanchun, Dou is from china, professor of Nanjing university.

Figure 5. Min, Chen’s expert knowledge graph



Min Chen is a professor in the School of Computer Science and Technology at HUST.

Case study

Min ,Chen， 16 keywords	Wanchun , Dou, 15 keywords	Herrera, Francisco， 19 keywords
Data analysis	Anonymization	Classification
Convex optimization	Data anonymization	Hadoop
Feature ranking	Preservation	Fuzzy rule based classification systems
Hadoop	Privacy preservation	Rule based classification
Health monitoring	Behavior analysis	Cost-sensitive learning
Healthcare	Big medical data	Data mining
Internet	Collaborative filtering	Data Set
Internet of Things	Deployment	Bioinformatics
Localization	GPS	Clustering
manifold learning	Hadoop	Data reduction
Microarray	Multimedia	Dimensionality reduction
Regression	Preference	Discretization
Smart grid	QoS	Feature selection
Smart home	Recommender systems	Fuzzy systems
Virtualization	Service composition	Imputation
Wireless network		Nearest neighbor
		Preprocessing
		Random Forest
		Scalability

Case study

- **Outcome**
- **GED between Wanchun and Herrera is 20.66**
- **GED between Wanchun and Min of 16.45.**

- **This outcome shows that graph edit distance between Wanchun and Herrera is alittle bit bigger than Wanchun and Min, which also means , Wanchun’s expertise composition is more similar with Min than Herrera.However, in this article, concrete value of graph edit distance for one expert pair is useless, but when compared with other expert pairs it is meaningful.**

Conclusion and future work

conclusion:

We use text mining technology and visualization tool to create expert knowledge network map. We then applied graph edit distance on on expert knowledge network map to measure expert knowledge similarity.
We find Graph edit distance method is an efficient way on expert knowledge similarity measuring and expert knowledge network map is a vivid expression of expert knowledge.

Contributions:

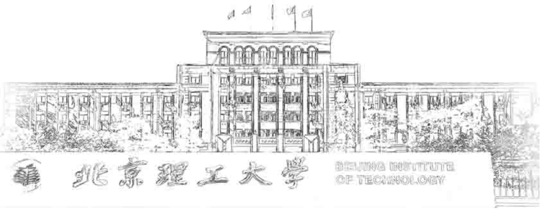
The study of expert knowledge measurement is meaningful for expert finding, expert identification and competitors or partners locating. It is also a fresh Appliance of graph edit distance and a creative way to measure expert similarity.

Limitation:

Large computational complexity

Future work:

1.Find a suitable way to add node weight into cost function
2.Need to do similarity measurement contrast verification



Thanks for your listening!
Question & Comments

Qingyun Liao

livia615118@foxmail.com

School of Management and Economics, Beijing Institute of Technology, Beijing, China