HITIR's Update Summary at TAC2008 Extractive Content Selection Using Evolutionary Manifold-ranking and Spectral Clustering

Reporter: Ph.d candidate He Ruifang
rfhe@ir.hit.edu.cn

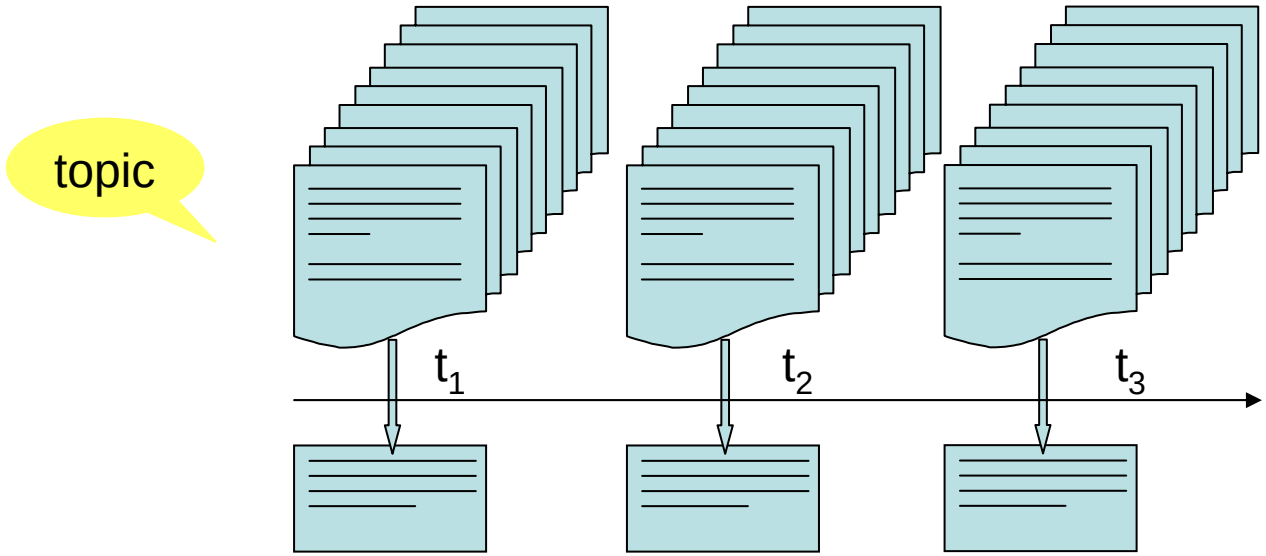
Information Retrieval Lab
School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China

Evaluation rank

- Three top 1 in PYRAMID
  - *average modified(pyramid) score*
  - *average numSCUs*
  - *macro-average modified score with 3 models of PYRAMID*
- 13^{th} *in ROUGE-2*
- 15^{th} *in ROUGE-SU4*
- 17^{th} *in BE*

Update summary introduction

- Aims to capture evolving information of a single topic changing over time
- Temporal data can be considered to be composed of many time slices



Question analysis

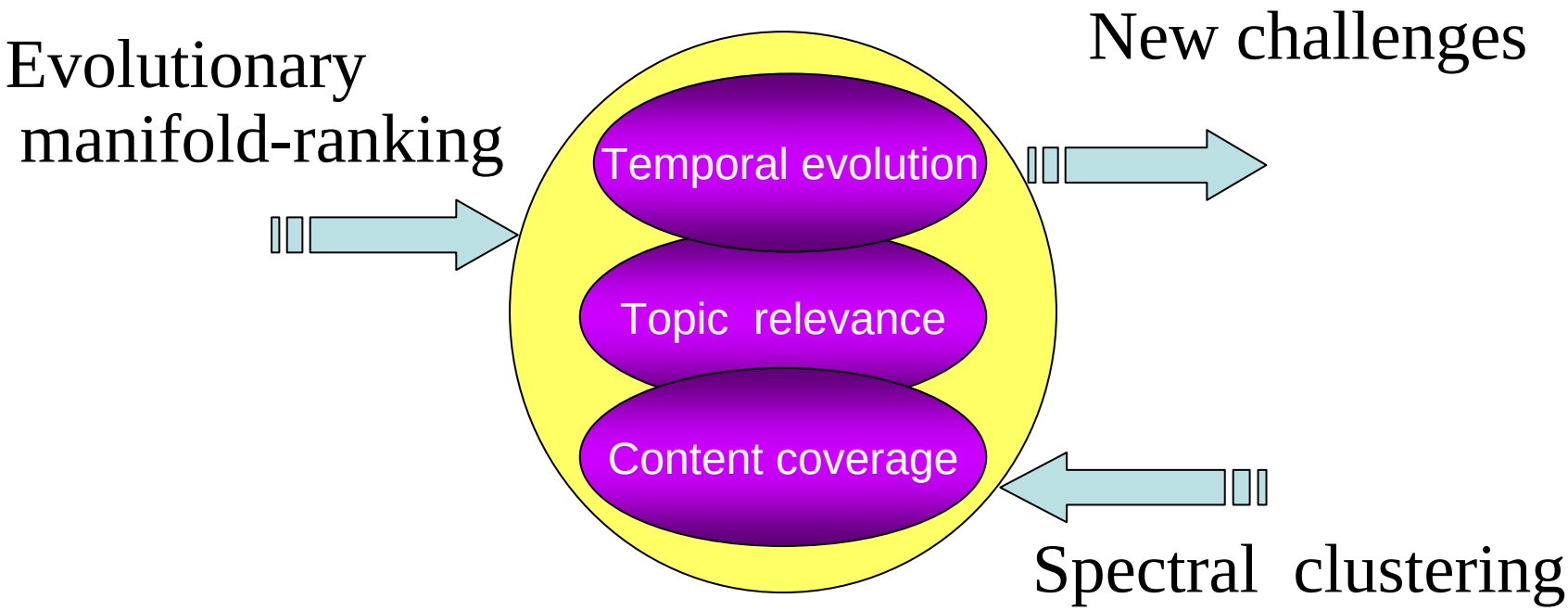
- from view of data
  - First difference: data has the temporal evolution characteristic
    - Deal with dynamic document collection of a single topic in continuous periods of time
- from view of users
  - Second difference: user needs have evolution characteristic
    - Hope to incrementally care the important and novel information relevant to a topic

Challenges for update summary (extractive or generative)

- Content selection
  - Importance
  - Redundancy
  - Content coverage
- Language quality
  - Coherence
  - Fluency

- Just focus on the extractive content selection
- How to model the importance and the redundancy of topic relevance and the content converge under the evolving data and user needs?

Explore the new manifold-ranking framework under the context of temporal data points!



Combine evolutionary manifold-ranking with spectral clustering to improve the coverage of content selection!

Evolutionary manifold-ranking

- Manifold-ranking ranks the data points under the intrinsic global manifold structure by their relevance to the query

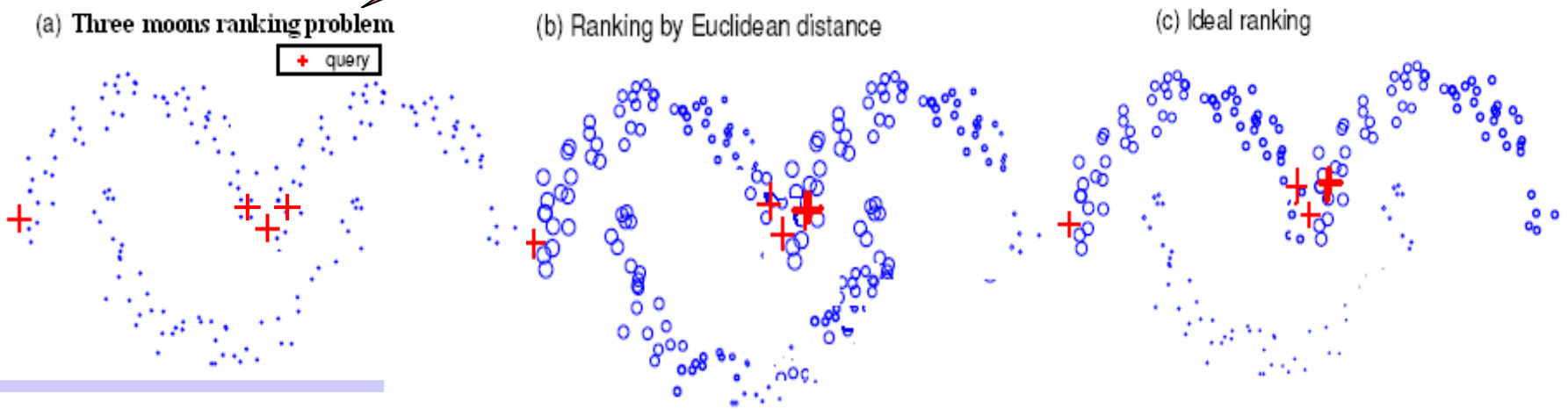
- Difficulty: not model the temporally evolving characteristic, as the query is static !

- Assumption of our idea
  - Data points evolving over time have the long and narrow manifold structure

Motivation of our idea

- Relay point of information propagation
  - Dynamic evolution of query
  - Relay propagation of information
- Iterative feedback mechanism in evolutionary manifold-ranking
  - The summary se evious time slices
  - The first sent ces of documents in current time slice

Relay point of information propagation



Manifold-ranking: Notation

- *n* sentences→data points
- *t* query→label
- One Affinity Matrix for data points: (W, D, S)
  - W: original similarity matrix
  - D: diagonal matrix
  - S: normalized matrix

- Labeling Matrix: $Y = [Y_1^T, \cdots, Y_n^T]^T$ $n \times n$
- Vectorial Function (ranking): $F = [F_1^T, \cdots, F_n^T]^T$ $n \times n$
- Learning task: $\{(W, D, S); Y\} \rightarrow F$

Regularization framework

$$Q(F)=\frac{1}{2}\left(\sum_{i,j=1}^{n}W_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}F_i-\frac{1}{\sqrt{D_{jj}}}F_j\right\|^2+\mu\sum_{i=1}^{n}\left\|F_i-Y_i\right\|^2\right)$$

$$F^*=\arg\min_F Q(F)$$

Fitting constraint

Smoothness constraint

Iterative form

$$F(t+1)=\alpha SF(t)+(1-\alpha)Y$$

Closed form

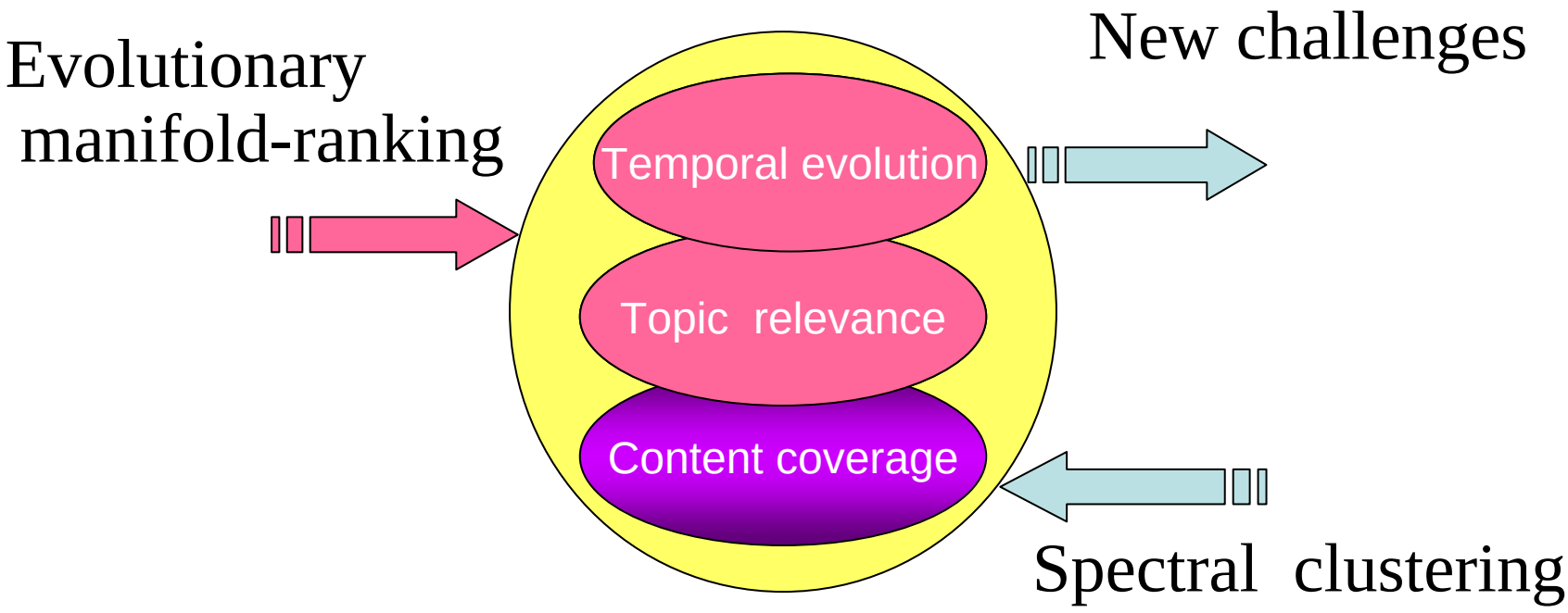
$$F^*=(I-\alpha S)^{-1}Y$$

Evolutionary manifold-ranking framework

Iterative feedback mechanism

- New iterative form $F(t+1) = \alpha S F(t) + (\beta Y_1 + \gamma Y_2 + \eta Y_3)$
- Closed form $F^* = (I - \alpha S)^{-1}(\beta Y_1 + \gamma Y_2 + \eta Y_3)$

- Labeling Matrix:
  - the original query $Y_1 = [Y_{11}^T, \cdots, Y_{1n}^T]^T$
  - the summary sentences from previous time slices $Y_2 = [Y_{21}^T, \cdots, Y_{2n}^T]^T$
  - the first sentences of documents in current time slices $Y_3 = [Y_{31}^T, \cdots, Y_{3n}^T]^T$


Evolutionary
manifold-ranking
Temporal evolution
Topic relevance
Content coverage
New challenges
Spectral clustering

Normalized Spectral clustering

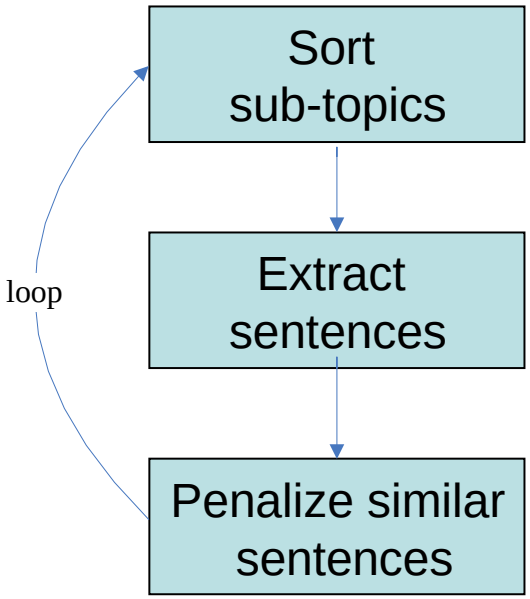
- Why choose the spectral clustering?
  - Automatically determine the number of clusters
  - Cluster the data points with arbitrary shape
  - Converge to the globally optimal solution

- Center object of spectral clustering
  - Graph Laplacian transformation
  - Select normalized random walk Laplacian
    - Have good convergence

Basic idea of spectral clustering

- Good property
  - the number of clusters is determined by the multiplicity of the eigenvalue 0 of normalized random walk Laplacian matrix
- Post processing
  - the properties of eigenvector
  - K-means

Sentence selection

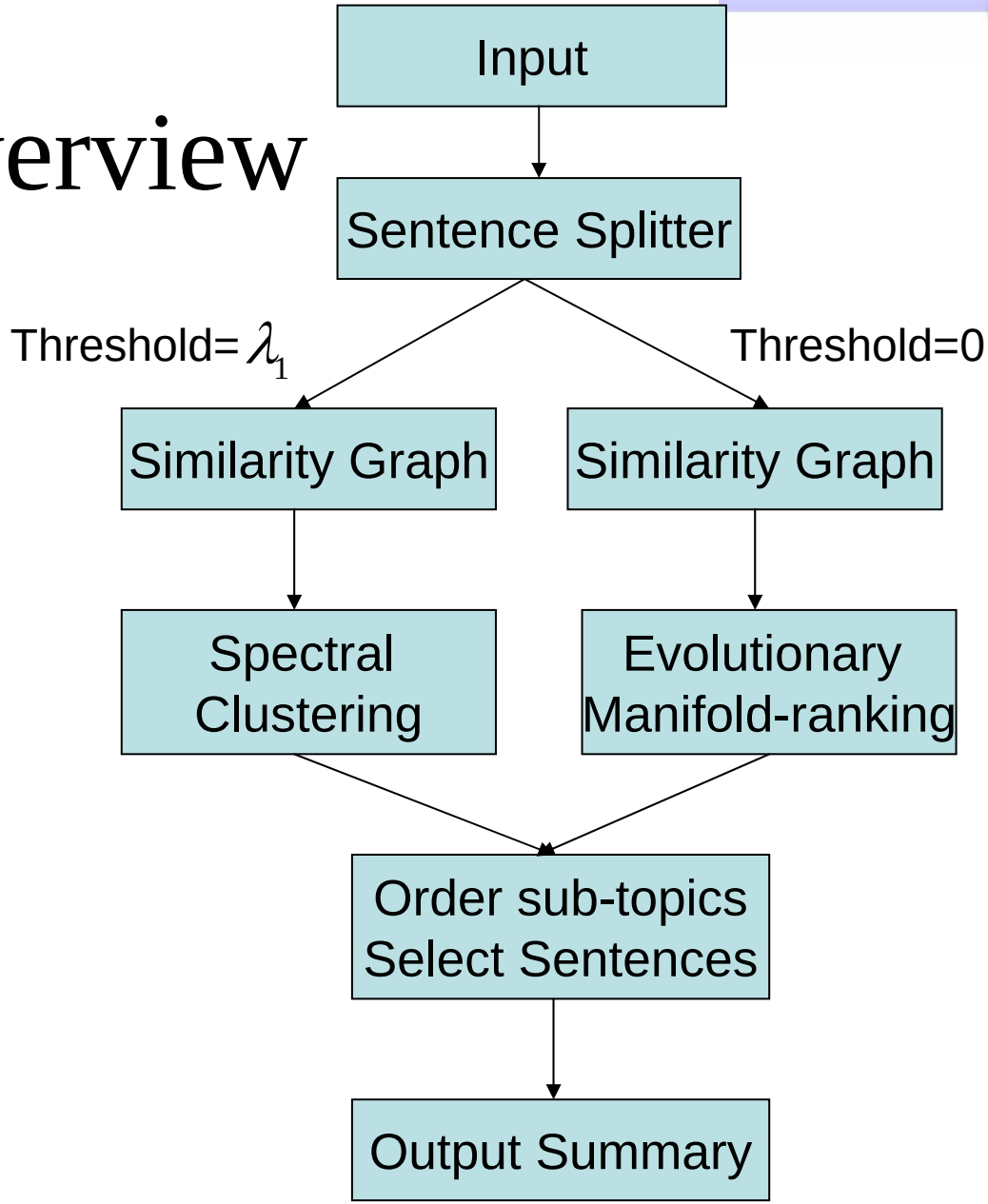


no sub-topics →a greedy algorithm

System design schemes

System No.\Priority	Spectral clustering (post-processing)		
	Properties of eigenvector	k-means	
Evolutionary manifold-ranking	11(1)	41(2)	62(3)

System overview



Evaluation rank

Table 1. The Update Summary Evaluation Results(*Rank*|*Score*)

Evaluation Principle	11	41	62
average modified (pyramid) score	1|0.336	4|0.318	
average numSCUs	1|4.781	2|4.469	
average numrepetitions	9|1.042	9|1.042	
macroaverage modified score with 3 models	1|0.331	3|0.313	
average linguistic quality	27|2.406	33|2.323	
average overall responsiveness	5|2.542	8|2.479	
ROUGE-2	13|0.08854	19|0.08353	15|0.08729
ROUGE-SU4	15|0.12477	19|0.12073	16|0.12250
BE	21|0.05134	31|0.04813	17|0.05228

- *three top 1*
  - *average modified(pyramid) score*
  - *average numSCUs*
  - *macro-average modified score with 3 models of PYRAMID*
- 13th *in ROUGE-2*
- 15th *in ROUGE-SU4*
- 17th *in BE*

Personal viewpoint

- ROUGE and BE →content selection of generative summary
  - Relatively short SCU
- PYRAMID→ content selection of extractive summary
  - Long SCU

- Hope: extend the number of time slices of evolving data

- Conclusion
  - Use normalized spectral clustering and evolutionary manifold-ranking to model the new characteristics of update summary
  - Develop the extractive content selection method for language independence
- Future work
  - Develop high level models
  - Better optimization method of parameters
- Common topic
  - Further explore the appropriate evaluation method for update summary

Thank you!

Any question?