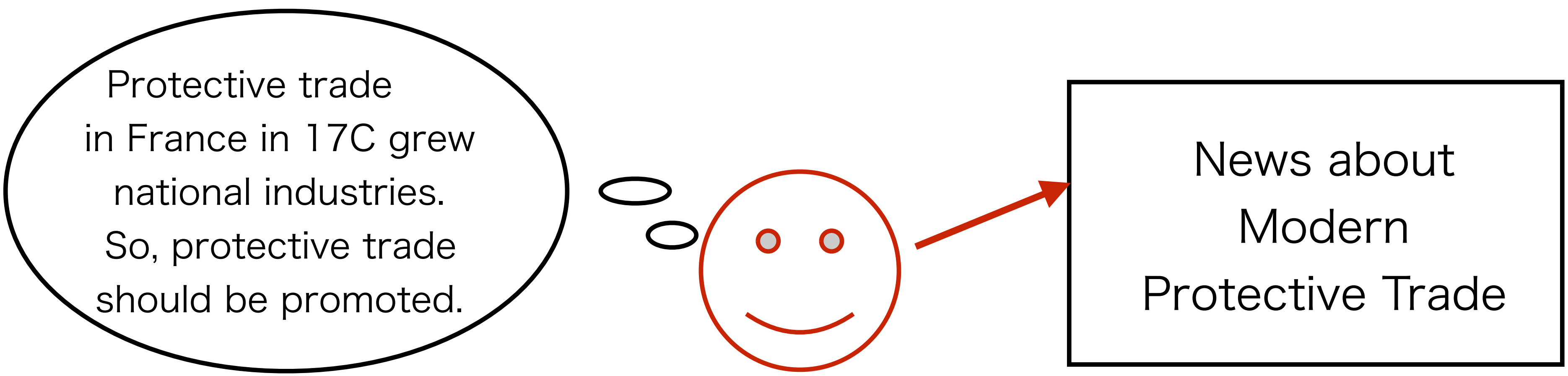
Towards Enhancing Historical Analogy: Clustering Users Having Different Aspects of Events

Ryohei IKEJIRI（Assistant Professor, The University of Tokyo, Japan）
Ryo YOSHIKAWA（Assistant Professor, Nagoya Bunri University, Japan）
Yasunobu SUMIKAWA（Assistant Professor, Tokyo University of Science, Japan）

Background

■Historical analogy

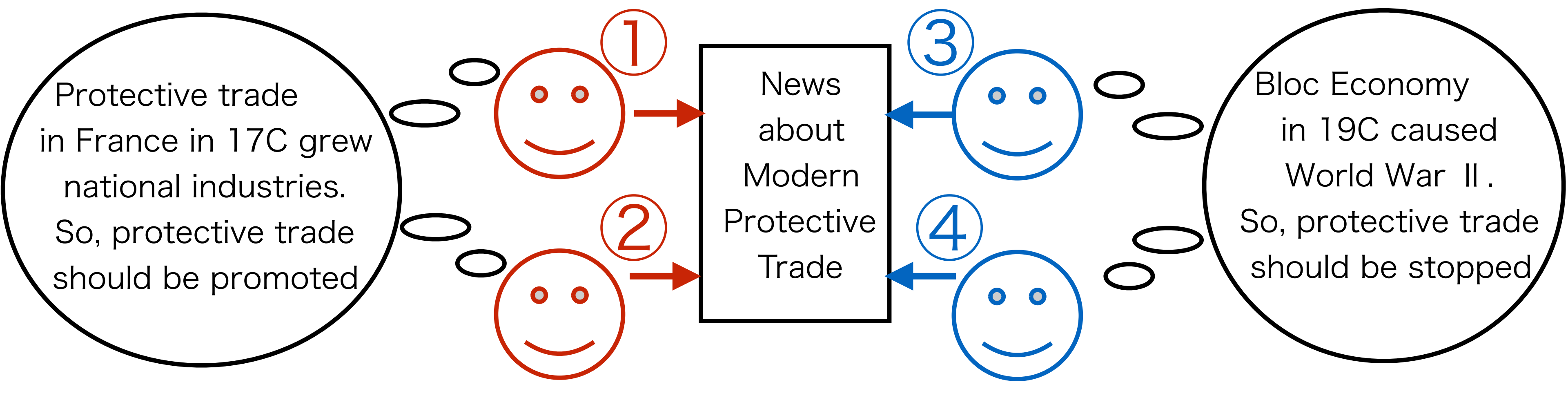
- HA =Applying historical causation to solve modern social issues
- Learning HA is effective in solving modern social issues (Staley 2002)
- Learning HA is regarded as important in history education



Background

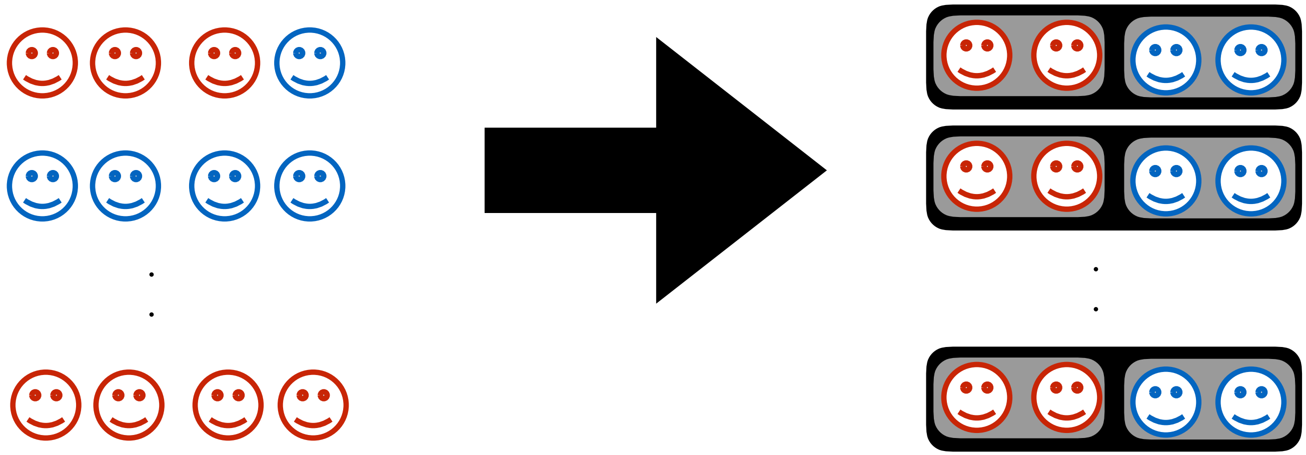
■Towards Enhancing Historical analogy

- Careful discussion is needed for using HA (Fischer 1970)
- Checking the validity of historical analogy in a pair who have same aspects is enhancing historical analogy (Ikejiri 2011)
- Group discussion between two pairs who have different aspects is enhancing historical analogy (Ikejiri et al. 2016) (Holyoak 1980)



Purpose

- We propose a novel clustering algorithm to promote historical analogy through group discussions
- Our algorithm has two objectives to create groups:
  1) finding two users(=pair) who focus on the same aspects of an event
  2) aggregating two pairs (=group) that have different aspects in the same event



Related Works

- Traditional clustering algorithm basically makes groups by similar data

  →The key contribution of our algorithm is to combine not similar data into a group

- Below clustering algorithms are close to ours
  - Partitioning-based algorithm (ex. k-means)
  - Hierarchy-based algorithm (ex. Birch)
  - Distribution-based algorithm (ex. GMM)
  - Graph-based algorithm (ex. Spectral)

Data Collection

- We have developed a educational system for promoting HA with which each user searches for historical events similar to the selected news

(Ikejiri et al. 2016)

Data Collection

- For searching similar history, we use 13 categories that characterize both modern society and historical events on definition of Encyclopedia of Historiography

In Politics Field

1. reign
2. diplomacy
3. war

In Economy Field

4. Production
5. Commerce

In Culture Field

6. study
7. religion
8. literature and thought
9. technology

In Society Field

10. popular movement
11. community
12. disparity
13. environment

Data Collection

- We regard the categories selected by each user as reflecting each aspects of historical analogy

West African Ebola virus epidemic

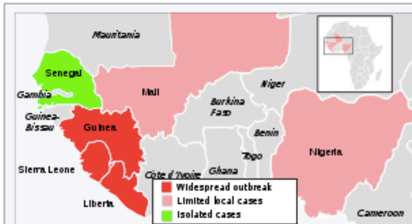
From Wikipedia, the free encyclopedia

The most widespread outbreak of Ebola virus disease (EVD) in history began in 2013 and continued until 2016, causing major loss of life and socioeconomic disruption in the West African region, mainly in the countries of Guinea, Liberia, and Sierra Leone. The first cases were recorded in Guinea in December 2013; later, the disease spread to neighboring Liberia and Sierra Leone,[12] with minor outbreaks occurring elsewhere. It caused significant mortality, with the case fatality rate reported at slightly above 70%,[12][13][14][note 1] while the rate among hospitalized patients was 57–59%.[15] Small outbreaks occurred in Nigeria and Mali,[7][16] and isolated cases were recorded in Senegal,[17] the United Kingdom and Sardinia.[14][18] In addition, imported cases led to secondary infection of medical workers in the United States and Spain but did not spread further.[19][20] The number of cases peaked in October 2014 and then began to decline gradually, following the commitment of substantial international resources. As of 8 May 2016, the World Health Organization (WHO) and respective governments reported a total of 28,616 suspected cases and 11,310 deaths[21] (39.5%), though the WHO believes that this substantially understates the magnitude of the outbreak.[22][23]

On 29 March 2016, the WHO terminated the Public Health Emergency of International Concern status of the outbreak.[24][25][26] Subsequent flare-ups occurred; the last was declared over on 9 June 2016, 42 days after the last case tested negative on 28 April 2016 in Monrovia.[27]

The outbreak left about 17,000 survivors of the disease, many of whom report post-recovery symptoms termed post-Ebola syndrome, often severe enough to require medical care for months or even years. An additional cause for concern is the apparent ability of the virus to "hide" in a recovered survivor's body for an extended period of time and then become active months or years later, either in the same individual or in a sexual partner.[28] In December 2016, the WHO announced that a two-year trial of the rVSV-ZEBOV vaccine appeared to offer protection from the strain of Ebola responsible for the West Africa outbreak. The vaccine has not yet been given regulatory approval, but it is considered to be so effective that 300,000 doses have already been stockpiled.[29][30]

West African Ebola virus epidemic



Simplified Ebola virus epidemic situation map

Date December 2013 – June 2016[1][2]

Casualties

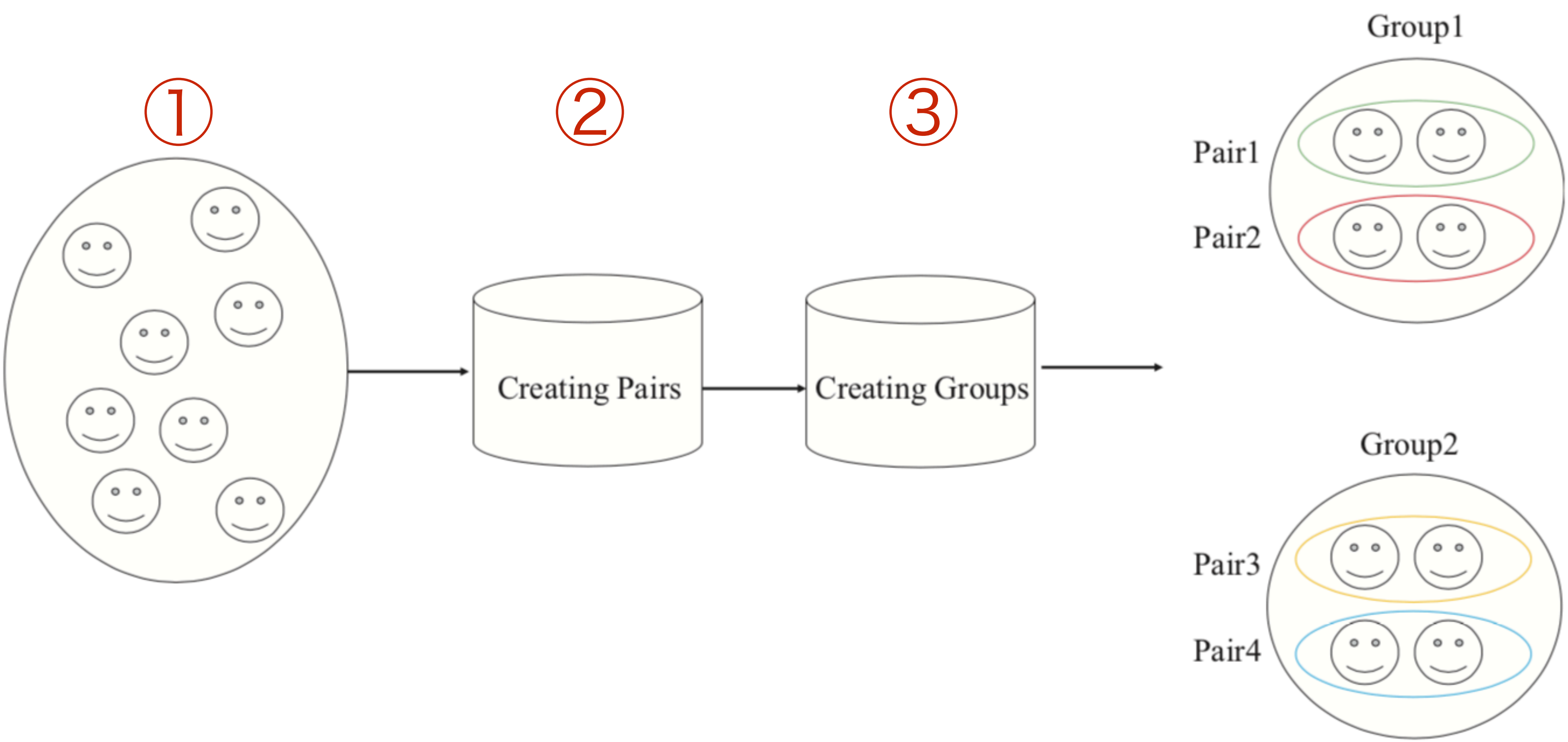
- *Note: current estimates suggest that between 17 percent and 70 percent of Ebola cases were unreported.*[3]

Country	Cases	Deaths	Last update On 9 June 2016 by WHO
Liberia	10,666	4,806	outbreak ended 9 June 2016[2]
Sierra Leone	14,122	3,955	outbreak ended 17 March 2016[4]
Guinea	3,804	2,536	outbreak ended 1

Environment, Study, Technology

Methodology

■Overview of our algorithm



①Feature Vector Creation

- We take event categories selected by users
- We convert the categories to a feature vector whose elements are represented by 0 or 1
- We create a feature vector for each user

student	reign	diplo macy	war	produ ction	comme rce	study	religi on	literature & thought	techn ology	popular movement	commun ity	dispar ity	environ ment
Student 1	0	1	0	0	1	0	0	0	0	0	0	0	0
Student 2	1	0	0	0	0	0	1	0	0	1	0	0	0
…													
student N	0	1	0	0	1	0	1	0	0	0	0	0	0

②Creating Pairs

A) Measuring similarity between two users

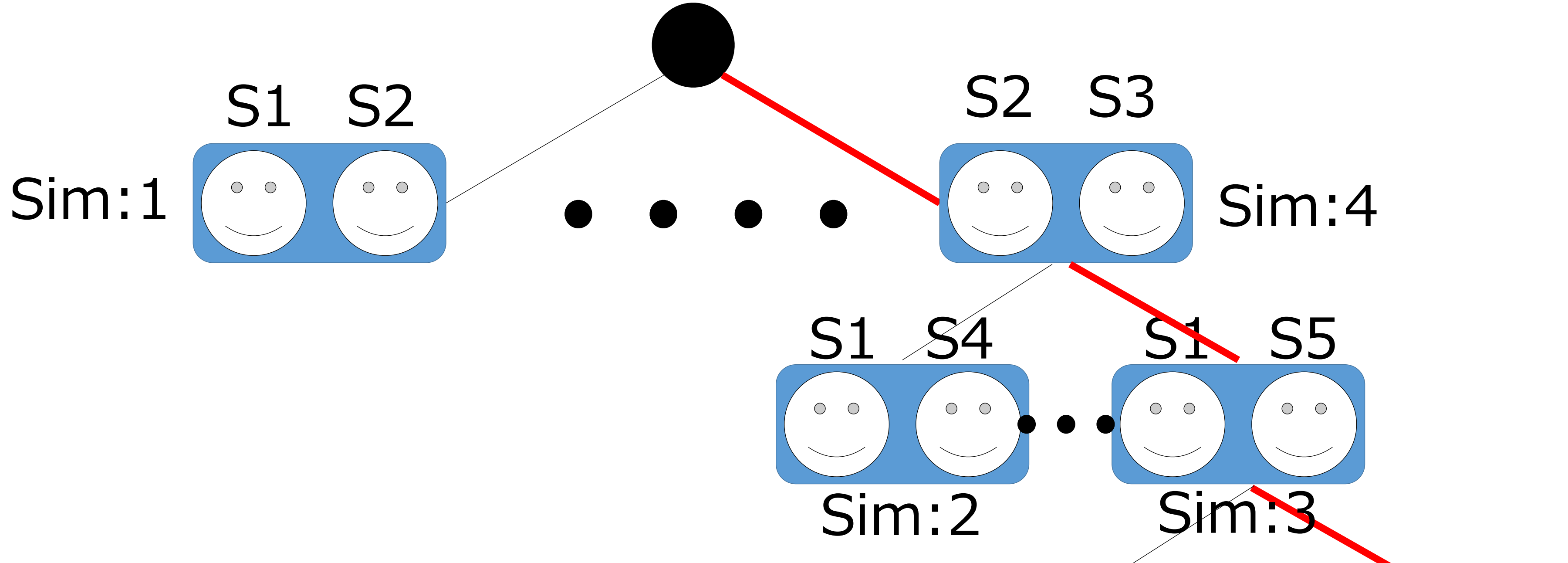
- Pair Similarity is measured by counting how many common categories between two users
- All Pair Similarities are calculated

student	reign	diplo macy	war	produ ction	comme rce	study	religi on	literature & thought	techn ology	popular movement	commun ity	dispar ity	environ ment
Student 1	0	1	0	0	1	0	0	0	0	0	0	0	0
Student 2	1	0	0	0	0	0	1	0	0	1	0	0	0
…													
student N	0	1	0	0	1	0	1	0	0	0	0	0	0

②Creating Pairs

B) Creating a Set of Pairs

- We solve the maximum problem that is essentially Knapsack problem
- Our algorithm find and fix a pair with max Pair Similarity
- The same processing is repeated with the remaining students



③Creating Groups

A) Measuring Similarity between two pairs

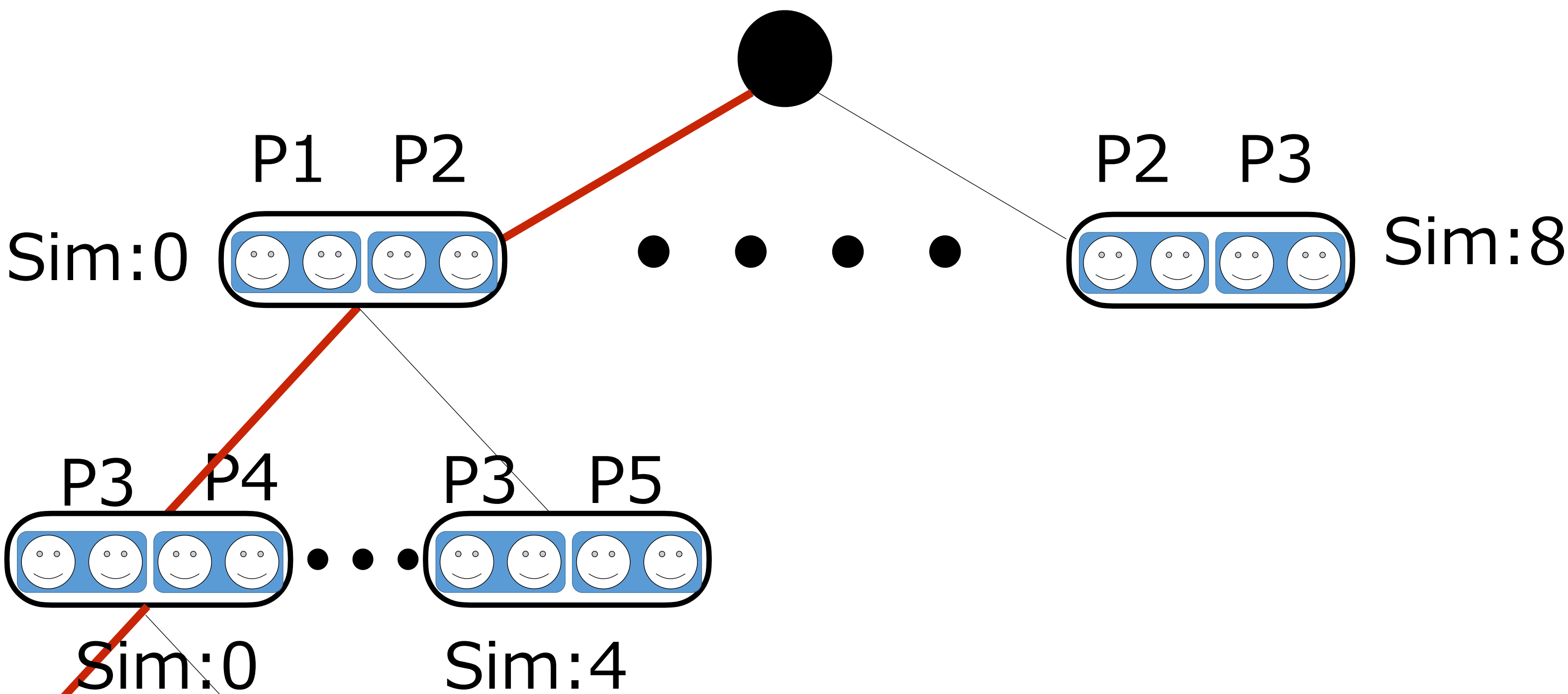
- A feature vector for a pair is created considering pair-level selected categories
- Group Similarity is measured by counting how many common categories between two pairs
- All Group Similarities are calculated

pair	reign	diplomacy	war	production	commerce	study	religion	literature & thought	technology	popular movement	community	disparity	environment
pair 1	0	2	0	0	2	0	0	0	0	0	0	0	0
pair 2	1	0	0	0	0	0	2	0	0	1	0	0	0
...													
pair N	0	2	0	0	2	0	2	0	0	0	0	0	0

③Creating Groups

B) Creating a Set of Groups

- We solve the minimum problem that is essentially Knapsack problem
- Our algorithm find and fix a group with minimum Group Similarity
- The same processing is repeated with the remaining pairs



Experimental Evaluation

■Setup

・One present event is prepared (Labor Problem)

・40 high school students participated in this experiment
→Each students selected some categories for using HA

■Baselines

・We compared our algorithm with K-means, Birch, GMM, Spectral（We set 10 clusters as a result of dividing the number of users by 4 in all algorithms)

Experimental Evaluation & Result

①Size of clusters

→kmeans, Birch, and GMM algorithms fail to include more than 2 users in a few clusters

		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
kmeans	Size	2	11	8	4	3	2	2	2	5	1
	MinDist	1.732	0.0	0.0	1.414	1.414	0.0	1.0	1.732	0.0	0.0
	Inner-cluster	1.732	0.0	35.113	11.295	5.382	0.0	1.0	1.732	5.656	-
Birch	Size	3	7	2	2	2	4	16	2	1	1
	MinDist	1.414	0.0	1.732	1.732	1.732	0.0	0.0	1.414	0.0	0.0
	Inner-cluster	4.878	21.999	1.732	1.732	1.732	7.292	68.709	1.414	0.0	0.0
GMM	Size	12	8	3	2	4	2	1	5	2	1
	MinDist	0.0	0.0	0.0	1.414	1.414	1.414	0.0	0.0	1.732	0.0
	Inner-cluster	19.052	35.431	2.828	1.414	11.799	1.414	-	5.656	1.732	-
Spectral	Size	3	11	6	4	3	2	3	2	3	3
	MinDist	1.414	0.0	1.414	0.0	0.0	0.0	0.0	1.0	1.732	1.732
	Inner-cluster	4.878	0.0	30.381	0.0	0.0	0.0	2.0	1.0	5.464	5.464
Proposed	Size	4	4	4	4	4	4	4	4	4	4
	MinDist	0.0	0.0	0.0	0.0	1.414	1.732	0.0	1.732	0.0	0.0
	Inner-cluster	11.948	11.922	9.797	11.922	11.032	13.512	11.103	12.385	0.0	0.0

Experimental Evaluation & Result

②Quality of Clustering
(= We use Calinski and Harabaz score)

—Average Minimum distances in a cluster
—Inner-cluster（=Average sum of distances of all combinations in each cluster)

Algorithm	Quality	Ave. MinDist	Inner-cluster
k-means	9.834	0.729	6.191
Birch	10.022	0.802	10.948
GMM	9.018	0.597	7.932
Spectral	9.714	0.729	4.918
Proposed	*1.740*	*0.487*	*9.362*

Conclusion

■Conclusion

- Our clustering algorithm makes groups by combining not only similar users but also not similar pairs
- Experimental results proved that only our algorithm creates suitable groups

■Future Works

- analyzing how well users can discuss with our algorithm
- proposing more sophisticated grouping algorithm
- analyzing robustness of our algorithm