

Identifying the Evolution of Disasters and Responses with Network-Text Analysis

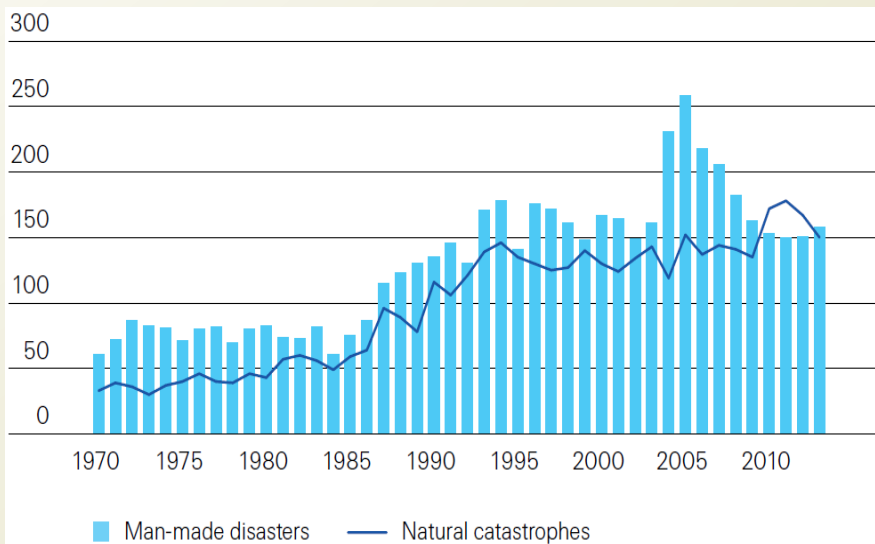
Kyungwoo Song, Do-Hyeong Kim, Su-Jin Shin, Il-Chul Moon
Dept. of Industrial and Systems Engineering, KAIST
Socio-Economic Systems Laboratory

Introduction

Motivation

Importance of Disaster [1]

<Number of events satisfying sigma criteria*>



- Generally, the number of disaster events grows steadily

(*1. Total economics losses over 96m\$ or
2. 50 injured and homeless 2000 people etc.)

Complexity of Disaster

Fukushima Daiichi nuclear disaster [2]



- Nuclear disaster at the Fukushima Nuclear Power Plant on 11 March 2011
- Earthquake+Tsunami+Nuclear
- 15884 confirmed death
- Expected expenses more than \$105 billion

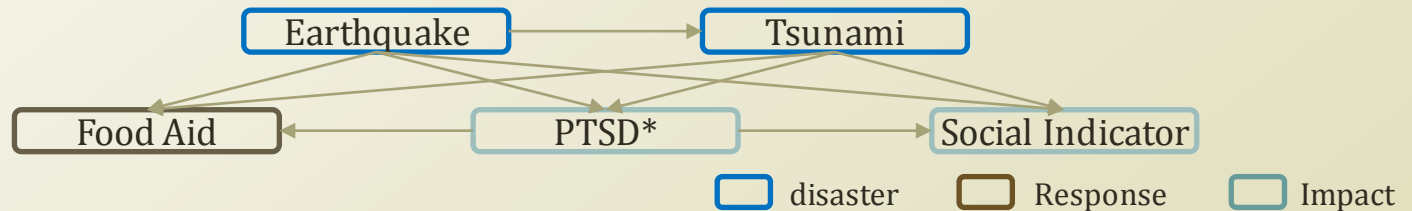
eBay Data Breach [3]



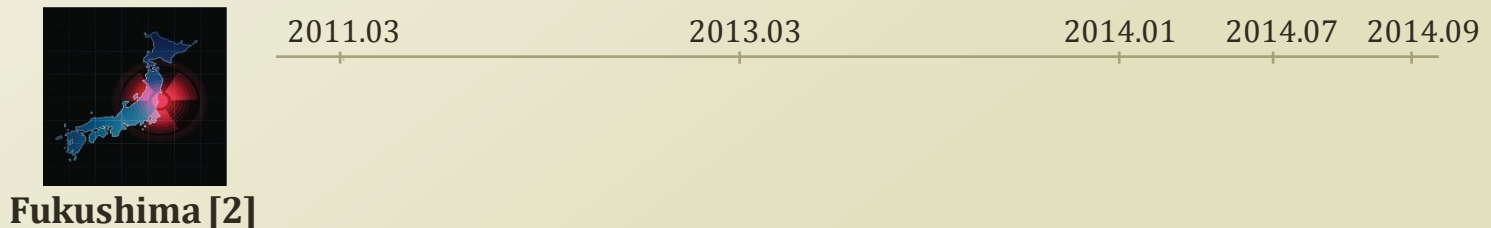
- Between late February and early march 2014
- Anthropogenic Intentional
- 145 million people of usernames, passwords, phone numbers and physical addresses are leaked

Motivation

- Analysis of overall disaster trend is essential
 - Disaster, response and their impact are intertwined.



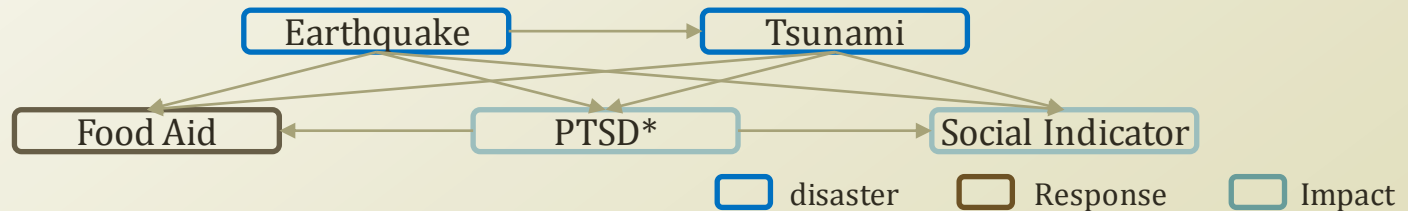
- Long-term Trend
 - The effect and response policy of complex disaster occurs in a long-term period



(*Post-Traumatic Stress Disorder)

Motivation

- Analysis of overall disaster trend is essential
 - Disaster, response and their impact are intertwined.



- Long-term Trend
 - The effect and response policy of complex disaster occurs in a long-term period



Fukushima [2]

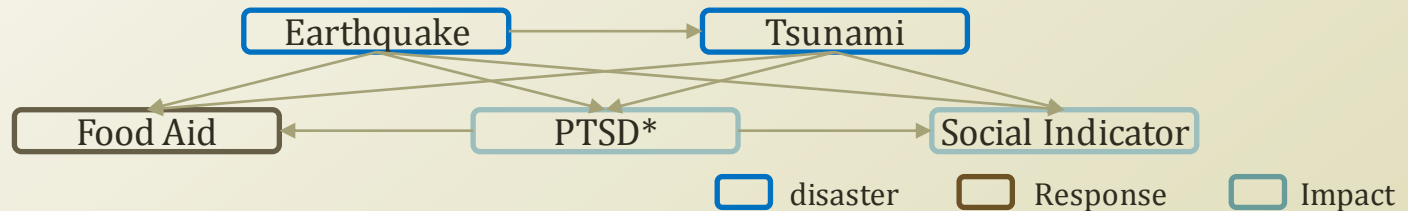
① The unit 3 reactor building explodes



(*Post-Traumatic Stress Disorder)

Motivation

- Analysis of overall disaster trend is essential
 - Disaster, response and their impact are intertwined.



- Long-term Trend
 - The effect and response policy of complex disaster occurs in a long-term period



Fukushima

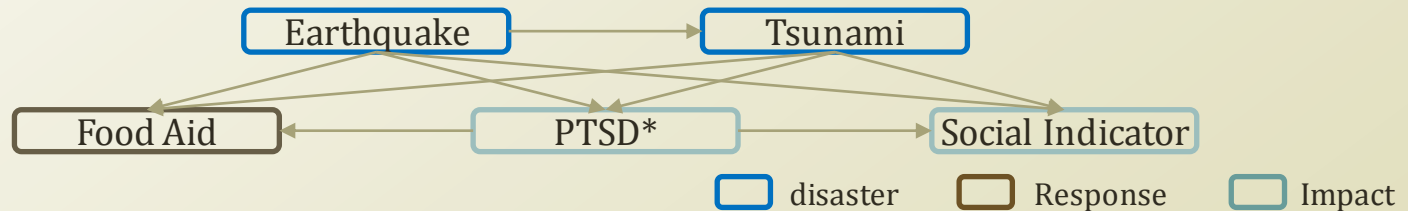
- 1 The unit 3 reactor building explodes
- 2 Government admits Fukushima have leaked radioactive water



(*Post-Traumatic Stress Disorder)

Motivation

- Analysis of overall disaster trend is essential
 - Disaster, response and their impact are intertwined.

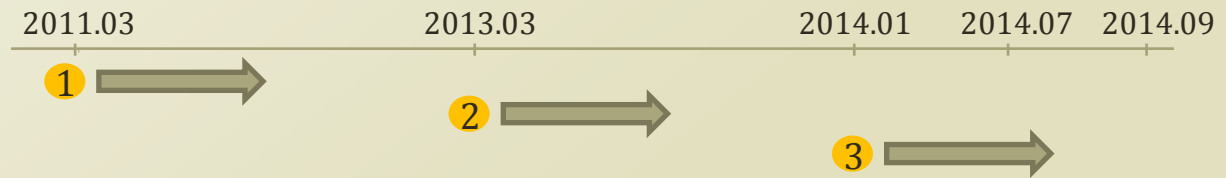


- Long-term Trend
 - The effect and response policy of complex disaster occurs in a long-term period



Fukushima

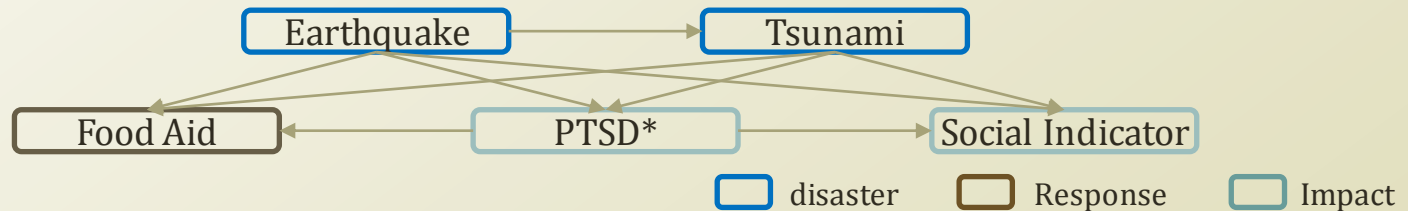
- 1 The unit 3 reactor building explodes
- 2 Government admits Fukushima have leaked radioactive water
- 3 Campaign to contain the spread of contaminated sediments



(*Post-Traumatic Stress Disorder)

Motivation

- Analysis of overall disaster trend is essential
 - Disaster, response and their impact are intertwined.



- Long-term Trend
 - The effect and response policy of complex disaster occurs in a long-term period



Fukushima

- 1 The unit 3 reactor building explodes
- 2 Government admits Fukushima have leaked radioactive water
- 3 Campaign to contain the spread of contaminated sediments
- 4 Planning a controlled nuclear meltdown



(*Post-Traumatic Stress Disorder)

Previous Research

Major flood disasters in Europe: 1950–2005, (J. I. Barredo, 2006) [4]

- Investigation of the flooding over 56 years in the EU region
- Analysis of the statistical long-term trend of financial damage and casualties

Trends in mental illness and suicidality after Hurricane Katrina (R. C. Kessler, 2008) [5]

- Investigation about the trends of PTSD after Katrina hurricane
- Surveyed over the two-year period.

➔ Area-Restricted trend analysis
Context-Restricted Analysis

Discovering Evolutionary Theme Patterns from Text (Q. Mei and C. Zhai, 2005) [6]

- Analysis of the changes of online news articles about Asia Tsunami with from 2004 to 2005 with text-mining techniques
- The analysis utilized a probabilistic mixture model to extract topics of the articles

Trends of Probable Post-Traumatic Stress Disorder in New York City after the September 11 Terrorist Attacks(S.Galea, 2003) [7]

- Surveyed posttraumatic stress disorder (PTSD) of New-York citizens after 911 terror.
- Measured by one, four, and six months

➔ Period-Restricted trend analysis
Word level-Restricted analysis
Sequential-Restricted analysis

We analyze the overall disaster trend with text-mining technique
(Text-mining technique is helpful to analyze the large and long term dataset and its context)

Research Objective

- Provide insights into the future disasters and responses
- Analysis of knowledge in the disaster field as a whole
 - Word analysis
 - Major words through all years('06~'13)
 - Fast increasing and decreasing words
 - Topic analysis
 - Fast increasing & decreasing Topics
 - Trend in fast increasing & decreasing topics
 - ex) Damage estimation is a fast increasing topic
 - From PTSD to social environment analysis in an estimation trend

Experiment Dataset

- Research articles explain the disaster, its cause, response, damage and so on carefully
- Necessity of dataset considering disasters and response

Dataset Introduction

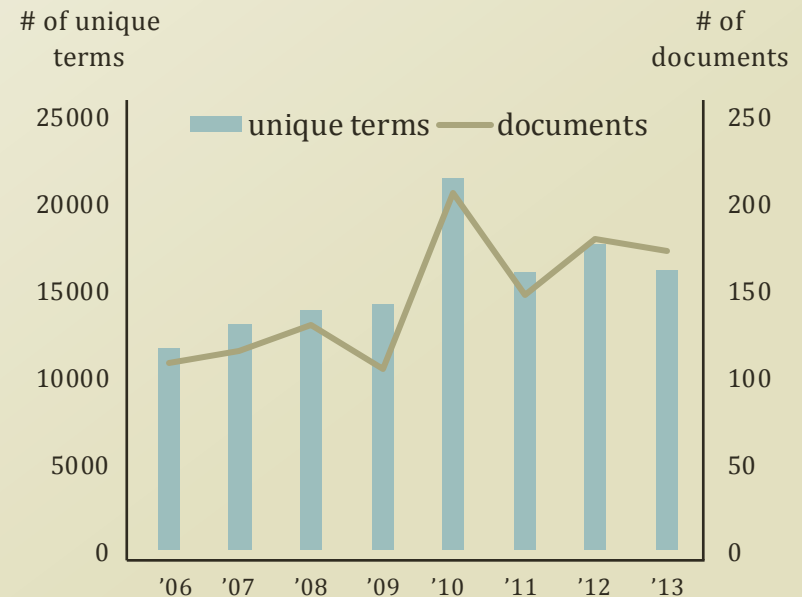
³⁾Disasters

- Area: Disasters causes and the natural disasters
- Leading journal in the complex and natural disasters
- Major peer-reviewed quarterly journal

⁴⁾ISCRAM (Information Systems for Crisis Response and Management)

- Area: disaster response and the man-made disasters
- Community for exchange and development of information system for crisis.
- Social & Technical and practical aspects of system

Descriptive Statistics Table



Total number of documents ► 1,158

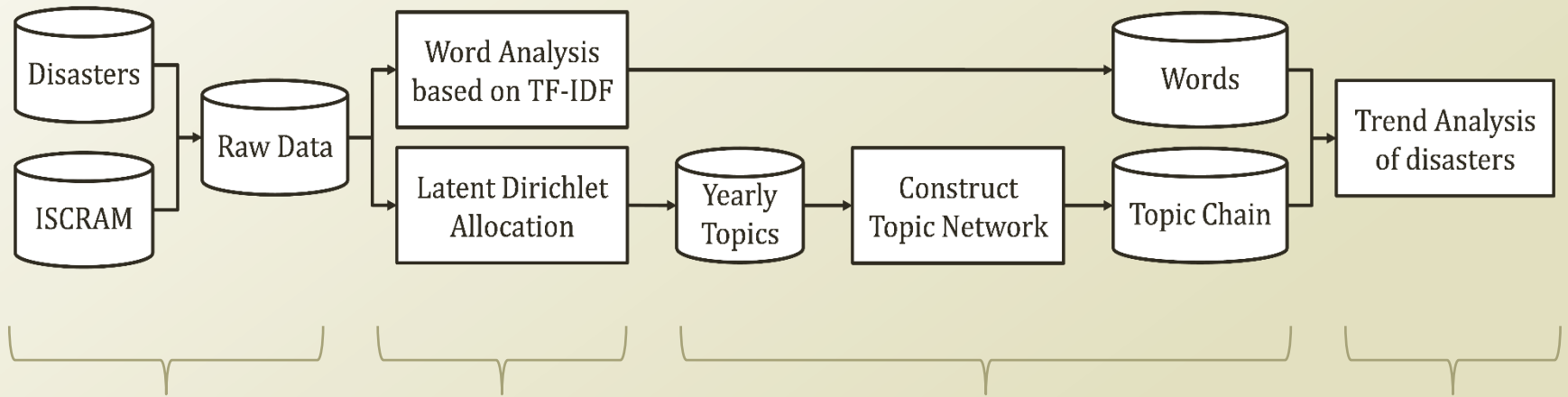
Total number of unique terms ► 64,295

*Source: ³⁾ [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1467-7717/homepage/ProductInformation.html](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-7717/homepage/ProductInformation.html)

⁴⁾ <http://www.iscramlive.org/portal/mission>

Research Framework

Experimental Framework for Disaster Trend Analysis



Dataset

Word and topic level analysis

Word and topic evolution over time

Trend analysis

- Preprocessing
- Merge the two corpora
- TF-IDF
- LDA
- Simple Linear Regression
- Spectral Clustering
- Word trend result
- Topic trend result

Methodology

Word-level:TF-IDF

- We want to know the important and highest frequency term in the corpus excluding the word like 'the, as, and' which is used a lot in all of the document
- TF-IDF is a measurement indicating the importance of words in the information retrieval field (high TF-IDF -> important word)

$$TF-IDF = TF \times IDF$$

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

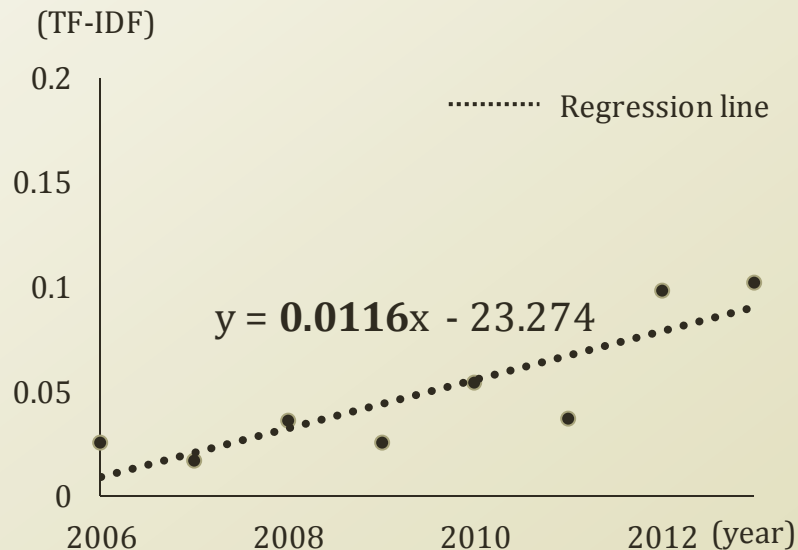
$tf_{i,j} = \# \text{ of word } i \text{ in } j^{\text{th}} \text{ document}$
 $df_i = \# \text{ of documents containing word } i$

- TF-IDF helps us to find the word because of multiplication of TF and Inverse of DF(IDF) (Common place word like 'the, as, and' cannot get a high tf-idf value because of high df score)

Dynamic Analysis-Word

- Our goals :
 1. Analyze disasters trend (time-series analysis)
 2. Analyze word trend as dynamically via simple linear regression based on TF-IDF*

<Annual TF-IDF value of word 'interoperability' >



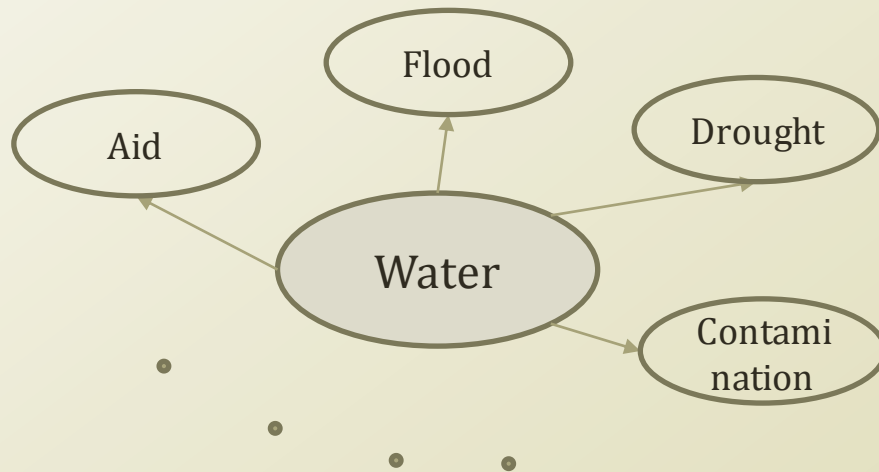
- The word, “**interoperability**”, starts appearing frequently
 - The information between disaster organizations need to :
 - Communicate
 - Diffuse information
 - Operate seamlessly

(*Post-Traumatic Stress Disorder)

Dynamic Analysis-Word

- Restriction of Word Analysis :
 1. Hard to interpretation
 2. One dimensional information in gauging the relation of words to the disasters and response

< Multiple usages of the word >



< Simple ups and downs >

- Easy to know ups and downs
- Hard to know specific trend
 - Word importance should be supported by its context
 - Some words might be related together, and similarly

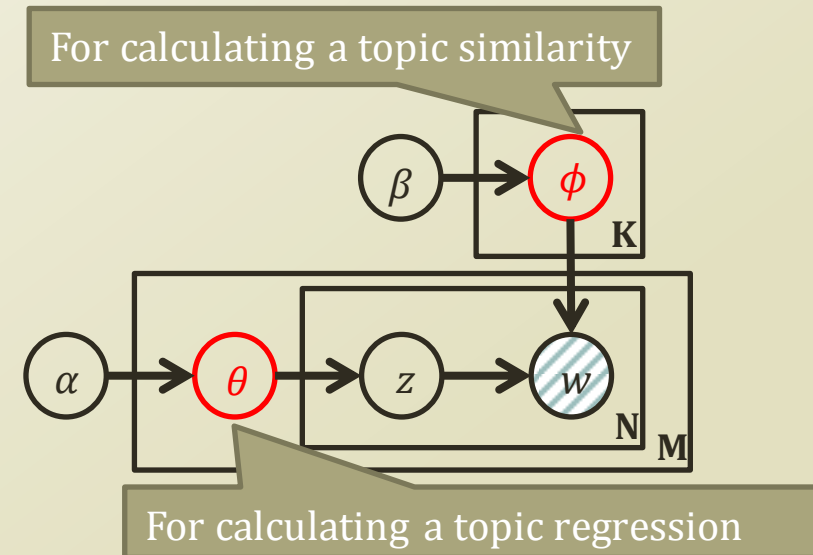
(*Post-Traumatic Stress Disorder)

LDA(Latent Dirichlet Allocation)[8]

Notation

Symbol	Meaning
α	Parameter for Dirichlet distribution
β	Parameter for Dirichlet distribution
θ_i	Topic proportion in document i
ϕ_k	Distribution of words in topic k
$w_{i,j}$	j^{th} word in i^{th} document
$z_{i,j}$	Topic assignment for $w_{i,j}$
K	Number of topics
M	Number of documents
N	Number of words

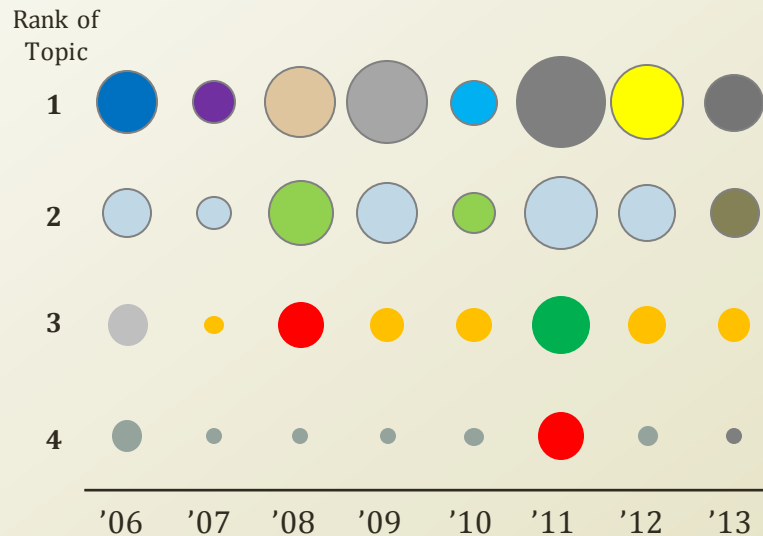
Graph representation



- For each topic k
 - Draw $\phi_k \sim \text{Dir}(\beta)$
- For each document i
 - Draw $\theta_i \sim \text{Dir}(\alpha)$
 - For each word j
 - Draw $z_{i,j} \sim \text{Multinomial}(\theta)$
 - Draw $w_{i,j}$ from $p(w_{i,j} | z_{i,j}, \phi)$

Dynamic Analysis-Topic

Topic proportion by year



Size of circle : topic proportion in the year

Color of circle : type of topic

ex) Blue ● : topic about **earthquake**,

Red ● : topic about **tsunami**

- Difficulty of Analyzing topic evolution
- Hardness of topic similarity determination
 - Topic about earthquake in 2006 and topic about tsunami(triggered by earthquake) in 2008
- Necessity of topic clustering regardless of year

Dynamic Analysis-Topic

1. Draw ϕ values for each word and topic

	w^1	w^2	...	w^v
Topic1	0.02	0.27		0.15
Topic2	0.13	0		0.23
⋮				

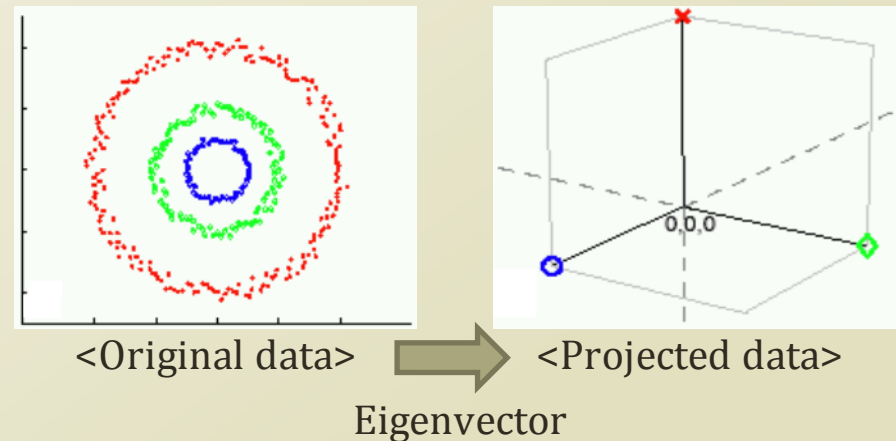
2. Calculate topic similarity between topics

$$S_{i,j} = \frac{\sum_{i=1}^v w_j^i \times w_k^i}{\sqrt{\sum_{i=1}^v (w_j^i)^2} \times \sqrt{\sum_{i=1}^v (w_k^i)^2}}$$

$w_j^i = \phi$ val of i^{th} wor in j^{th} top d i

$S_{i,j} = \text{simila}$ r betw i^{th} top an j^{th} top

3. Spectral Clustering [9]





- For topic clustering regardless of the year
- Clustering based on similarity without a year
- Clustering technique based on eigenvalue and eigenvector
- Projection the data to lower dimension by using eigenvector to cluster easily

Result and Conclusion

Result-Word analysis

WORD TF-IDF VALUE THROUGH ENTIRE CORPUS (2006' ~ 2013')

Rank	TF-IDF	Word	Rank	TF-IDF	Word
1	0.977	flood	16	0.723	participants
2	0.960	tweets	17	0.705	organizations
3	0.921	earthquake	18	0.704	decision
4	0.882	simulation	19	0.685	game
5	0.815	ontology	20	0.685	tsunami
6	0.802	user	21	0.685	community
7	0.797	network	22	0.680	users
8	0.772	health	23	0.680	twitter
9	0.763	Risk	24	0.676	scenario
10	0.761	Team	25	0.673	exercise
11	0.750	vulnerability	26	0.666	agent
12	0.750	fire	27	0.646	children
13	0.737	incident	28	0.637	web
14	0.740	model	29	0.630	security
15	0.725	emergency	30	0.629	media

- Top 30 keywords from 2006 to 2013.
- Estimation of the major issues in the period.
-  : “flood”, “earthquake”, “fire”, “tsunami”, indicate the importance of natural disasters.
-  : “tweets”, “simulation”, “ontology” suggests the application of IT to the disaster responses.

Result-Word analysis

Fast Increasing word

Rank	Coeff.	Word
1	0.0414	tweets
2	0.0240	exercise
3	0.0234	game
4	0.0229	twitter
5	0.0182	media
6	0.0134	health
7	0.0127	tweet
8	0.0125	recovery
9	0.0120	seed
10	0.0116	interoperability

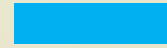


: Word about **SNS and SMS**

- SNS becomes popular in disaster area. (ex. SNS can be used to announcement or response of disaster)
- On the other hand, frequency of SMS becomes lower.
- Twitter replaces the role and function of SMS.

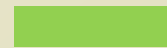
Fast Decreasing word

Rank	Coeff.	Word
1	-0.018	Cash
2	-0.016	Climate
3	-0.007	Measles
4	-0.005	Households
5	-0.005	Agent
6	-0.004	Vaccine
7	-0.004	Heritage
8	-0.004	SMS
9	-0.004	Livestock
10	-0.004	Mortality



: Word about **simulation**

- Simulation is widely used in the disaster area (ex. Simulation is used to training for disaster response)

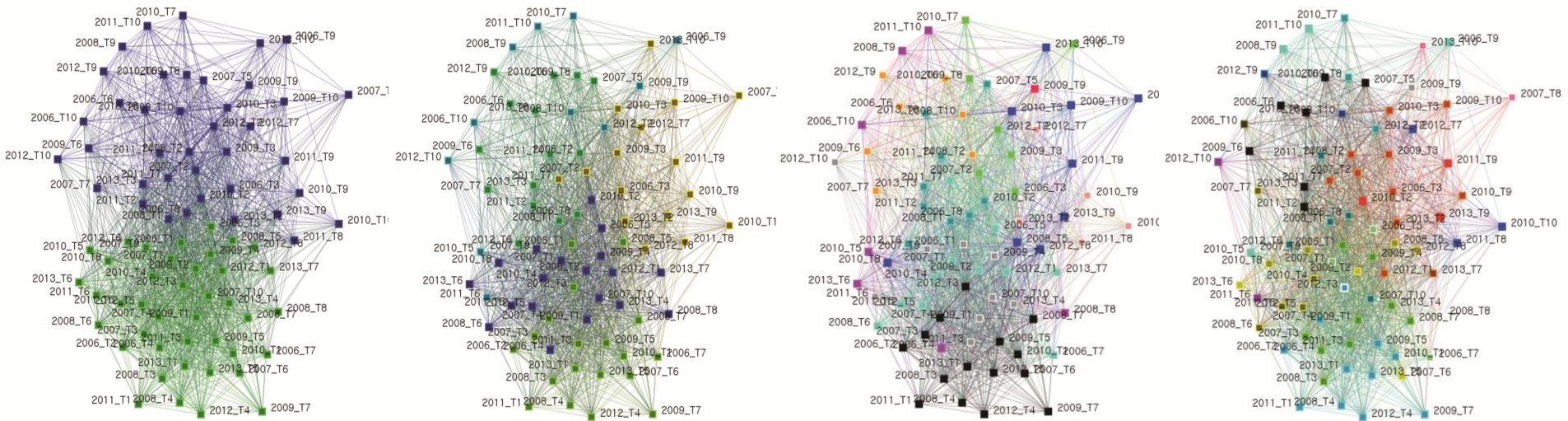


: Word about **measles** [10]

- Measles becomes less common disease.
- The number of measles changed from 370 thousand(in 2007) to 220thousand(in 2012)

Result-Topic Analysis

Graphical representation of spectral clustering by changing number of clusters



<# of clusters: 2>

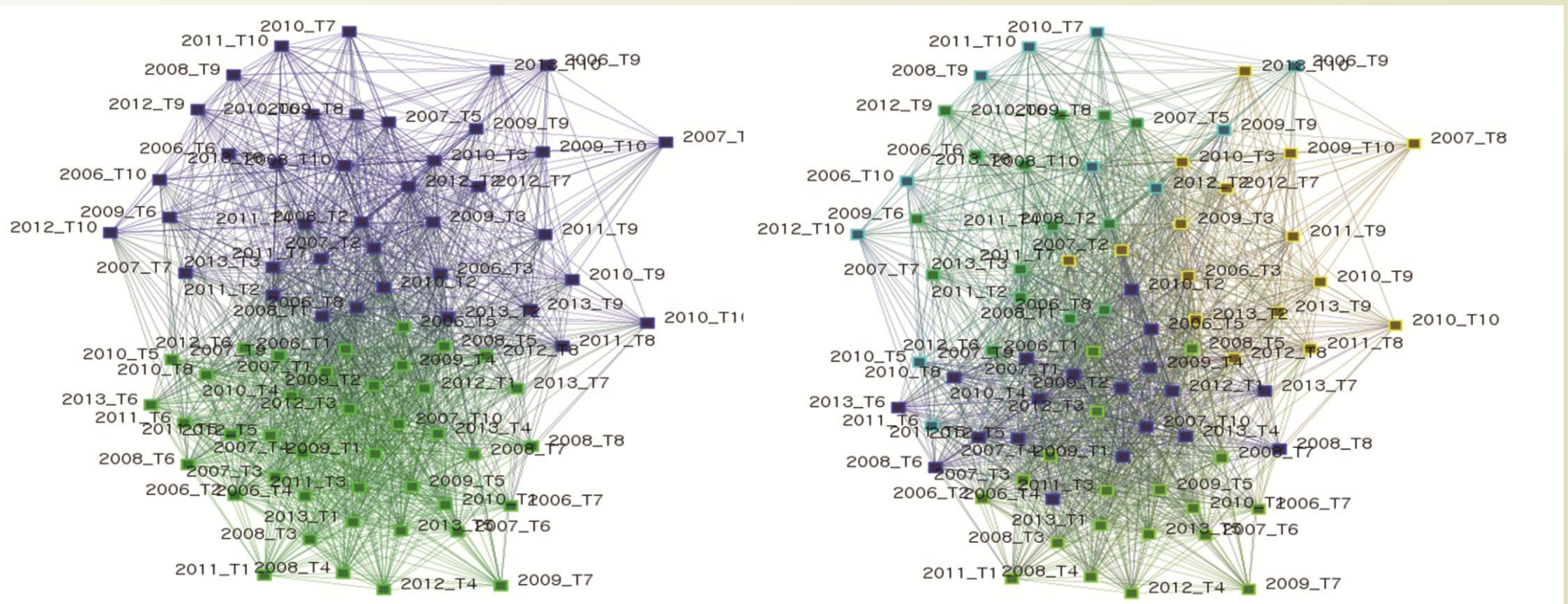
<# of clusters: 5>

<# of clusters: 10>

<# of clusters: 20>

Result-Topic Analysis

Graphical representation of spectral clustering by changing number of clusters



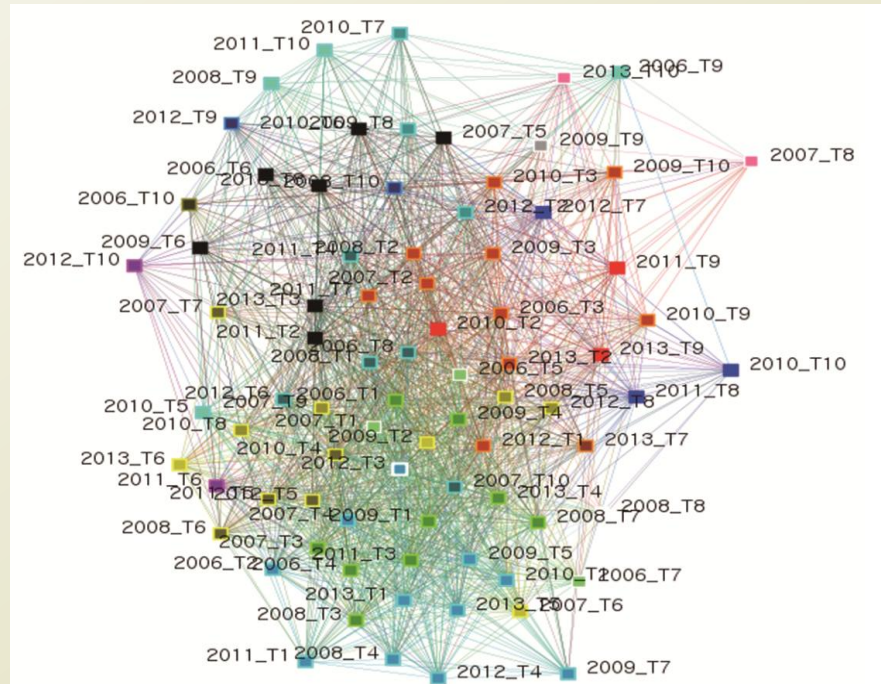
<# of clusters: 2>

<# of clusters: 5>

- Set as two or five makes the number of topics in each cluster is too many to find the specifics of the topic cluster.

Result-Topic Analysis

Graphical representation of spectral clustering by changing number of clusters

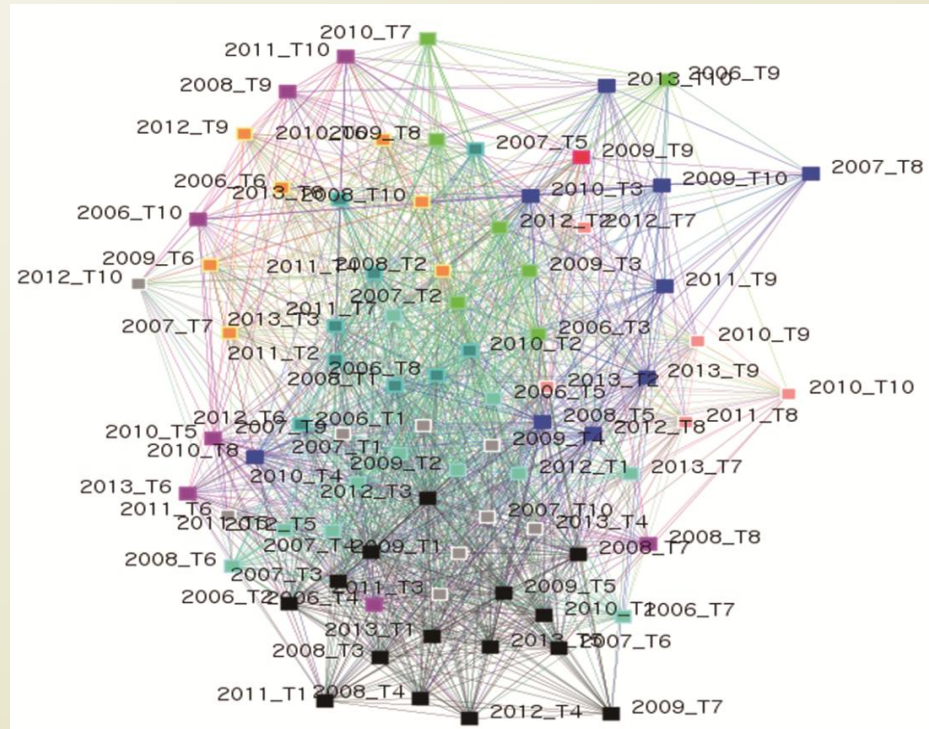


<# of clusters: 20>

- When the number of cluster is set to 20, there are very few merit of the clustering (The number of topics in each cluster is averagely less than four)

Result-Topic Analysis

Graphical representation of spectral clustering by changing number of clusters



<# of clusters: 10>

- Set the cluster number as 10 when the result is most interpretable.

Result-Topic Analysis

Simple Regression in each cluster

Cluster	Coeff.	Major word in a each cluster				
1	-0.001	Information	communication	Management	Organizations	Emergency
2	0.016	disaster	factor	risk	people	impact
3	0.01	aid	organizations	security	food	relief
4	-0.007	flood	task	modeling	coordination	networks
5	0.016	health	security	agencies	response	refugees
6	-0.019	children	population	household	impact	health
7	0.002	response	support	decision	simulation	models
8	-0.008	emergency	scenario	rescue	technology	simulation
9	-0.001	scenarios	sector	power	damage	families
10	-0.015	risk	people	vulnerability	building	hazards

**Fast
Increasing
Cluster**

**Fast
Decreasing
Cluster**

Result-Topic Analysis

Simple Regression in each cluster

Cluster	Coeff.	Major words in a each cluster				
1	-0.001	Information	communication	Management	Organizations	Emergency
2	0.016	disaster	factor	risk	people	impact
3	0.01	aid	organizations	security	food	relief
4	-0.007	flood	task	modeling	coordination	networks
5	0.016	health	security	agencies	response	refugees
6	-0.019	children	population	household	impact	health
7	0.002	response	support	decision	simulation	models
8	-0.008	emergency	scenario	rescue	technology	simulation
9	-0.001	scenarios	sector	power	damage	families
10	-0.015	risk	people	vulnerability	building	hazards

- Fast increasing cluster is cluster 2 which consider the estimation method of damages of disasters

Result-Topic Analysis

Simple Regression in each cluster

Cluster	Coeff.	Major words in a each cluster				
1	-0.001	Information	communication	Management	Organizations	Emergency
2	0.016	disaster	factor	risk	people	impact
3	0.01	aid	organizations	security	food	relief
4	-0.007	flood	task	modeling	coordination	networks
5	0.016	health	security	agencies	response	refugees
6	-0.019	children	population	household	impact	health
7	0.002	response	support	decision	simulation	models
8	-0.008	emergency	scenario	rescue	technology	simulation
9	-0.001	scenarios	sector	power	damage	families
10	-0.015	risk	people	vulnerability	building	hazards

- Fast increasing cluster is cluster 2 which consider the estimation method of damages of disasters
- Cluster 6 which consider PTSD and physical damage of victim becomes less important.

Result-Topic Analysis

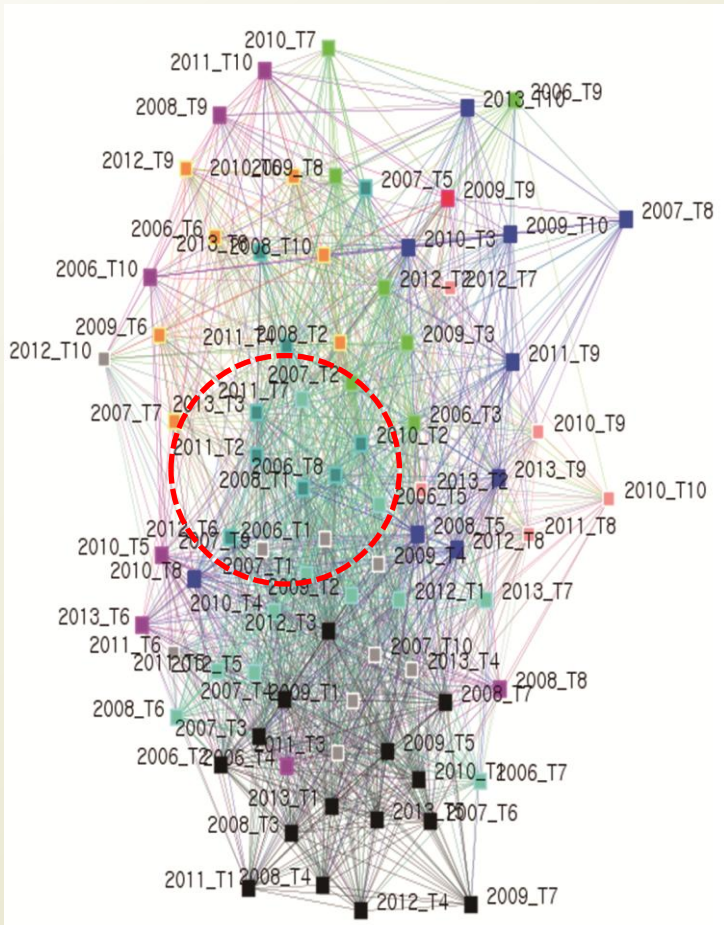
Simple Regression in each cluster

Cluster	Coeff.	Major words in a each cluster				
1	-0.001	Information	communication	Management	Organizations	Emergency
2	0.016	disaster	factor	risk	people	impact
3	0.01	aid	organizations	security	food	relief
4	-0.007	flood	task	modeling	coordination	networks
5	0.016	health	security	agencies	response	refugees
6	-0.019	children	population	household	impact	health
7	0.002	response	support	decision	simulation	models
8	-0.008	emergency	scenario	rescue	technology	simulation
9	-0.001	scenarios	sector	power	damage	families
10	-0.015	risk	people	vulnerability	building	hazards

- Fast increasing cluster is cluster 2 which consider the estimation method of damages of disasters
- Cluster 6 which consider PTSD and physical damage of victim becomes less important.
 - Instead of estimation focusing on PTSD, another damage estimation method becomes more popular

Result-Topic Analysis

Cluster 2: Measurement of the disaster damage



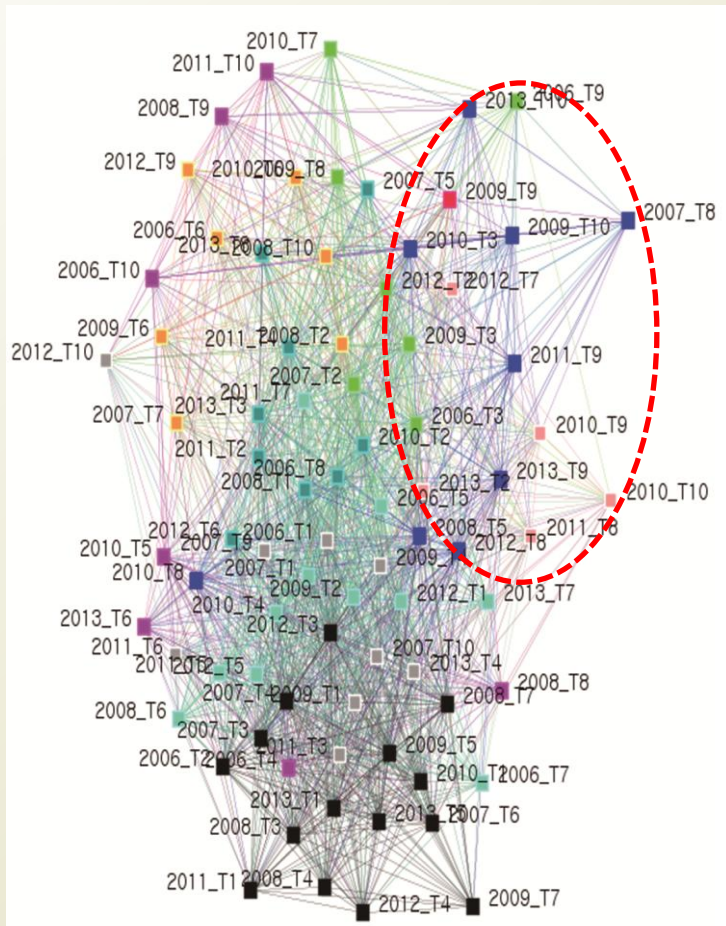
Main node of Cluster 2

Node	Words
2006_T8	children, earthquake , response, disasters, disaster, emerg ency, management, evaluation , emergencies, responses, na ture, health , labour, post , systems, reliability, resettlement, range, production, strategy,
2007_T5	disaster, earthquake , housing, damage, cent, earthquakes, disasters, flood, planning, activities, community, risk, survi vors , post, studies, period, organisations, factors , buildings, reconstruction,
2011_T4	disaster , disasters, vulnerability, business , earthquake, i mpact, building, cent, community, security, preparedness, b usinesses, hurricane, respondents, survey, variables, govern ment , management, source, impacts ,
2013_T8	disaster , community, vulnerability, cent, damage , manage ment, disasters, people, post, reconstruction, risk , event, hu rrricane, reduction, factors, income , communities, family, ka trina, population,

- Estimation of survivors and children in 2006, 2007
- Relation between the disaster damage and the societal environment of the disaster in recent times

Result-Topic Analysis

Cluster 3: Logistics effort



Main node of Cluster 3

Node	Major 20 Words
2007_T8	food, aid, sudan, wfp, security, conflict, darfur, livelihoods, people, access, country, insecurity, protection, region, livelihood, production, peace, organisations, war, ngo,
2008_T5	information, disaster, security, coordination, activities, ngos, trust, people, organisations, media, crisis, activity, approach, support, action, agencies, field, relief, aid, ngo,
2011_T9	people, tsunami, relief, coordination, ngos, aid, aceh, media, agencies, twitter, staff, issues, stress, focus, programme, standards, issue, sa, workers, assistance,
2013_T9	media, people, earthquake, aid, food, reports, china, governance, participation, countries, world, assistance, women, police, humanitarianism, emergency, report, role, information, awareness,
2013_T10	government, protection, aid, children, conflict, rights, child, sudan, services, household, town, households, governments, famine, structures, violations, resettlement, livelihood, chad, shelter,

2007,2008

Food supply and
organizational
cooperation

2011

Advanced
technology

2013

Humanitarian aid

Conclusion

Summary

- Identifying the evolution of disasters and responses with text-mining based on Disasters and ISCRAM paper

Application

Conclusion

Summary

- Identifying the evolution of disasters and responses with text-mining based on Disasters and ISCRAM paper
- The information diffusion with IT and the organizational interoperation have been the quickly expanding topics in the community

Application

Conclusion

Summary

- Identifying the evolution of disasters and responses with text-mining based on Disasters and ISCRAM paper
- The information diffusion with IT and the organizational interoperability have been the quickly expanding topics in the community
- Shift of the key interests: from the simple after action review of disaster responses to the societal aspects of disaster damage

Application

Conclusion

Summary

- Identifying the evolution of disasters and responses with text-mining based on Disasters and ISCRAM paper
- The information diffusion with IT and the organizational interoperability have been the quickly expanding topics in the community
- Shift of the key interests: from the simple after action review of disaster responses to the societal aspects of disaster damage

Application

- The utilized analysis can be applied to different sources of texts, such as news articles and social media.

Conclusion

Summary

- Identifying the evolution of disasters and responses with text-mining based on Disasters and ISCRAM paper
- The information diffusion with IT and the organizational interoperability have been the quickly expanding topics in the community
- Shift of the key interests: from the simple after action review of disaster responses to the societal aspects of disaster damage

Application

- The utilized analysis can be applied to different sources of texts, such as news articles and social media.
- The target corpus and the focus of the analyses can be adapted while reusing the methodology.

Conclusion

Summary

- Identifying the evolution of disasters and responses with text-mining based on Disasters and ISCRAM paper
- The information diffusion with IT and the organizational interoperation have been the quickly expanding topics in the community
- Shift of the key interests: from the simple after action review of disaster responses to the societal aspects of disaster damage

Application

- The utilized analysis can be applied to different sources of texts, such as news articles and social media.
- The target corpus and the focus of the analyses can be adapted while reusing the methodology.
- For example, we can analyze the change of public perspective about disaster and response

Reference

- [1] Natural catastrophes and man-made disasters in 2013, Swiss Re Economic Research & Consulting
- [2] 2011 Japan Earthquake - Tsunami Fast Facts, CNN Library, 2014
- [3] eBay Data Breach -- The 'Inexcusable' Impact on 233 Million Customers, Leo Sun, 2014
- [4] J. I. Barredo, "Major flood disasters in Europe: 1950–2005," *Nat. Hazards*, vol. 42, no. 1, pp. 125–148, Nov. 2006.
- [5] R. C. Kessler et al., "Trends in mental illness and suicidality after Hurricane Katrina," *Mol. Psychiatry*, vol. 13, no. 4, pp. 374–84, Apr.2008.
- [6] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in ... *conference on Knowledge discovery in data mining*, 2005, pp. 198–207.
- [7] S. Galea, "Trends of Probable Post-Traumatic Stress Disorder in New York City after the September 11 Terrorist Attacks," *Am. J. Epidemiol.*, vol. 158, no. 6, pp. 514–524, Sep. 2003
- [8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [9] H. Shinnou and M. Sasaki, "Spectral Clustering for a Large Data Set by Reducing the Similarity Matrix Size," in *In proceedings of the Sixth International Language Resources and Evaluation(LREC08)*, 2008, pp.201–204.
- [10] Progress in creating a world free of measles, rubella and congenital rubella syndrome, Measels & Rubella Initiative, 2013

APPENDIX

Dynamic Analysis-Topic

Result of Spectral clustering

<u>Cluster 1</u>	<u>Topic proportion</u>	<u>Cluster 2</u>	<u>Topic proportion</u>
2006_T1*	0.4	2006_T8	0.07
2006_T4	0.2	2007_T5	0.09
2008_T2	0.5	2008_T1	0.16
2008_T3	0.3	2010_T2	0.15
2008_T4	0.02	2011_T2	0.15
2010_T3	0.02	2011_T4	0.12
<u>Cluster 3</u>	<u>Topic proportion</u>	2012_T6	0.09
2006_T2	0.4	2013_T3	0.13
2006_T5	0.2	2013_T8	0.07

⋮

Topics about
damage estimation

Linear regression



- Damage estimation is the increasing topics of interests in the community.

(*2006_T1 : 1st topic in 2006)

Spectral Clustering

Spectral Clustering

- Clustering the data based on the eigenvalue and eigenvector
- Calculate Laplacian matrix and its smallest k eigenvalue and corresponding eigenvector matrix V
- Projection the data to lower dimension by using eigenvector to cluster easily

$$L = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}} \text{ whe } D_{ii} = \sum_{j=1}^n S_{ij}$$

$$V = [v_1, \dots, v_k] = D^{\frac{1}{2}} E \text{ whe } E = \begin{bmatrix} e_1 & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & e_k \end{bmatrix}$$

e_i : vect whi eleme n ar al 1