25th XBRL International Conference
Nov. 6-8, 2012

FUJITSU
shaping tomorrow with you

Analysis of XBRL Reports Using Text Mining

Ryo ISHIZAKI
Fujitsu Laboratories Ltd.

Background of the Research

FUJITSU

- The Research of Analysis Technologies in Fujitsu Labs.
  - We have worked on developing of text mining technologies in various domains. In some approaches, we tried to conduct an analysis that utilizes business information described in Annual Securities Reports. However, we could not it well because of some problems of Data availability and Data formatting.
- Progress of Data usability by XBRL
  - Recently, Annual Securities Reports have become available online. And its XBRL specification will be extended to the textual parts in 2013.
- Analysis of XBRL Reports and Discussion
  - So, we tried analyzing Annual Securities Reports. We will show the progress of our research, and we want to

Outline

- Part1: What is Text Mining?
  - Overview of text mining technologies and its applications
- Part2: Application of Text Mining to XBRL Reports
  - Analysis of annual securities reports
- Part3: Discussion

FUJITSU

Part1:
What is Text Mining?

- What is Mining?
- Technologies for Text Mining
- Examples of Text Mining Applications

What is Mining?

FUJITSU

- Data Mining and Text Mining
  - Data Mining: Knowledge discovery from numerical or categorical data
    - e.g.) Basket analysis on POS data in supermarket which reveals that paper diaper and canned beer are often bought together.
  - Text Mining: Knowledge discovery from textual data
    - e.g.) Analysis on Q&A log data in call center to find out consumer's needs, wants, claims and so on.
- Why textual data is important?
  - To find out unexpected (but described) knowledge
    - i.e.) questionnaire: choice (categorical) answer is for confirmatory analysis
      free (textual) answer is for exploratory analysis
- Combination of data and text mining
  - Successful approach is to detect trends and changes with data mining, and to figure out reasons and causes with text

Technologies for Text Mining

FUJITSU

- **Natural Language Processing Technology**
  - To extract the semantic contents from textual description
- **Visualization Technology**
  - To grasp the overview, characteristics and relations

Various information sources

On the Web

In companies

and so on..

Natural Language Processing Technology

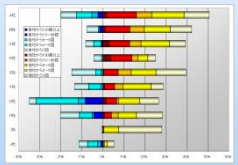
To extract the semantic contents

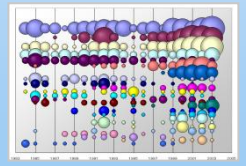
e.g. Marketing
To extract “reputations“ and "complaints" among customers

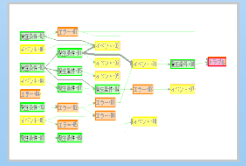
e.g. Patent Analysis
To extract “purpose“ and “object" of a patent

Visualization Technology

Visualize the overview, characteristics and relations

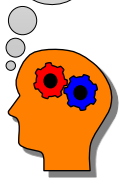






Support

Idea & Realization



Analysts

Natural Language Processing Technology

FUJITSU

Trouble Report

the Titanic accident

【Sequence】 Titanic, which the entire world was keeping its eye on, was thought to be an unsinkable ship. On April 10, 1912, it left the British port of Southampton toward New York in the US on her maiden voyage with about 2,220 passengers and crew on board a month after its original scheduled departure. After starting on the voyage, …

【Cause】 The direct cause of this accident was a collision with an iceberg. The hull consisted of a large number of compartments, …

Term Extraction

Extract words/phrases from textual description

entire world, keep one's eye on, left the British port, maiden voyage, original scheduled departure, …

Term Weighting

Iceberg, visibility → appeared in some reports → keyword

cause, accident → appeared in every report → common w

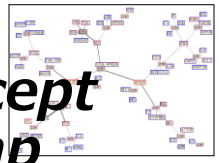
Calculation of Co-occurrences

Iceberg and *collision* → often used together → strong rela

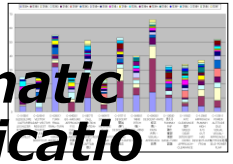
ignore and *warning* → often used together → strong rela

Visualization

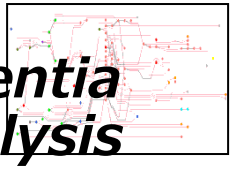
Concept Map



Automatic Classification



Sequential Analysis



Examples of Text Mining Applications

FUJITSU

- Patent Mining
  - To analyze and evaluate patents to build and grow a strong patent portfolio
- Proactive Risk Management
  - To prevent or avoid troubles before occurrence
- Automatic Generation of Near-miss Map
  - To specify areas/spots are strongly related to traffic accidents
- Market Defect Detection
  - To detect sign of malfunctions with products in the market

Summery of Part1:What is Text Mining ?

- Knowledge discovery from textual data
- Core technologies
  - Natural Language Processing Technology to extract the semantic contents from textual data
  - Visualization Technology to grasp the overview, characteristics and relations
- Application to various domains and purposes.

FUJITSU

Part2: Application of Text Mining to XBRL Reports(Annual Securities Reports)

- Overview of Annual Securities Reports
- Comparison Analyses of Description Contents

Sections of Annual Securities Reports

- There are 7 chapters, 25 sections (in general).
- 9 sections are mainly described by textual information.
  - We focused on only these sections in our analysis.

…Sections that describe narrative information

1.Overview of company	1.Summary of business results
	2.Company history
	3.Description of business
	4.Overview of group entities
	5.Information about employees
2.Overview of business	1.Overview of business results
	2.Overview of production, orders received and sales
	3.Issues to address
	4.Business risks
	5.Critical contracts for operation
	6.Research and development activities
	7.Analysis of financial position, operating results …
[illegible]	1.Overview of capital expenditures, etc.
	2.Major facilities
	3.Planned additions, retirements, etc. of [illegible]
4.Information about reporting company	1.Information about shares, etc.
	2.Acquisitions, etc. of treasury shares
	3.Dividend policy
	4.Historical records of share price
	5.Information about officers
	6.Explanation about corporate governance, etc.
5.Financial information	1.Consolidated financial statements, etc.
	2.Financial statements, etc.
6.Overview of operational procedures for shares	
7.Reference information	1.Information about parent company, etc. …
	2.Other reference information

Image of Process for Analysis

FUJITSU

Annual Securities Reports

Company X FY 2007 [volume of] description Existence of

Company Y FY 2007 [volume of] description Existence of

Company X FY 2008 [volume of] description Existence of

Company Y FY 2008 [volume of] description Existence of

1.Preprocessing (Splitting sections)

1	**Overview of**
1-1	Summary of business
1-2	Company history
1-3	Description
1-4	Overview of gr
1-5	Information
2	**Overview**
2-2	Overview of
2-3	Issues to address
2-4	Business risks
2-5	Critical contracts
3-2	Major facilities
3-3	Planned additions,.

Keywords

Sales →5

Acquisitio →2

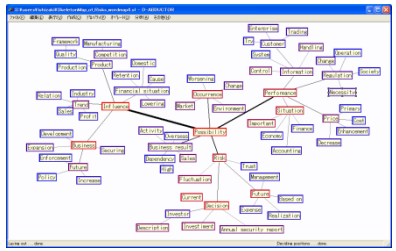
R&D →10

2.Natural Language Processing (Keyword Extraction & Aggregation)

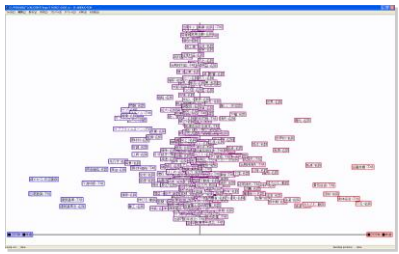
Aggregation Table

			Keywords A	B	C	…
FY 2007	company X	section 1	5	0	1	…
		section 2	4	1	2	…
		section 3	3	0	1	…
		…	…	…	…	…
	company Y	section 1	3	1	3	…
		section 2	4	1	2	…
		section 3	7	0	2	…
		…	…	…	…	…
	…	…	…	…	…	…
FY 2008	company X	section 1	2	2	1	…
		section 2	4	1	2	…
		section 3	3	0	1	…
		…	…	…	…	…

3. Analysis & Visualization



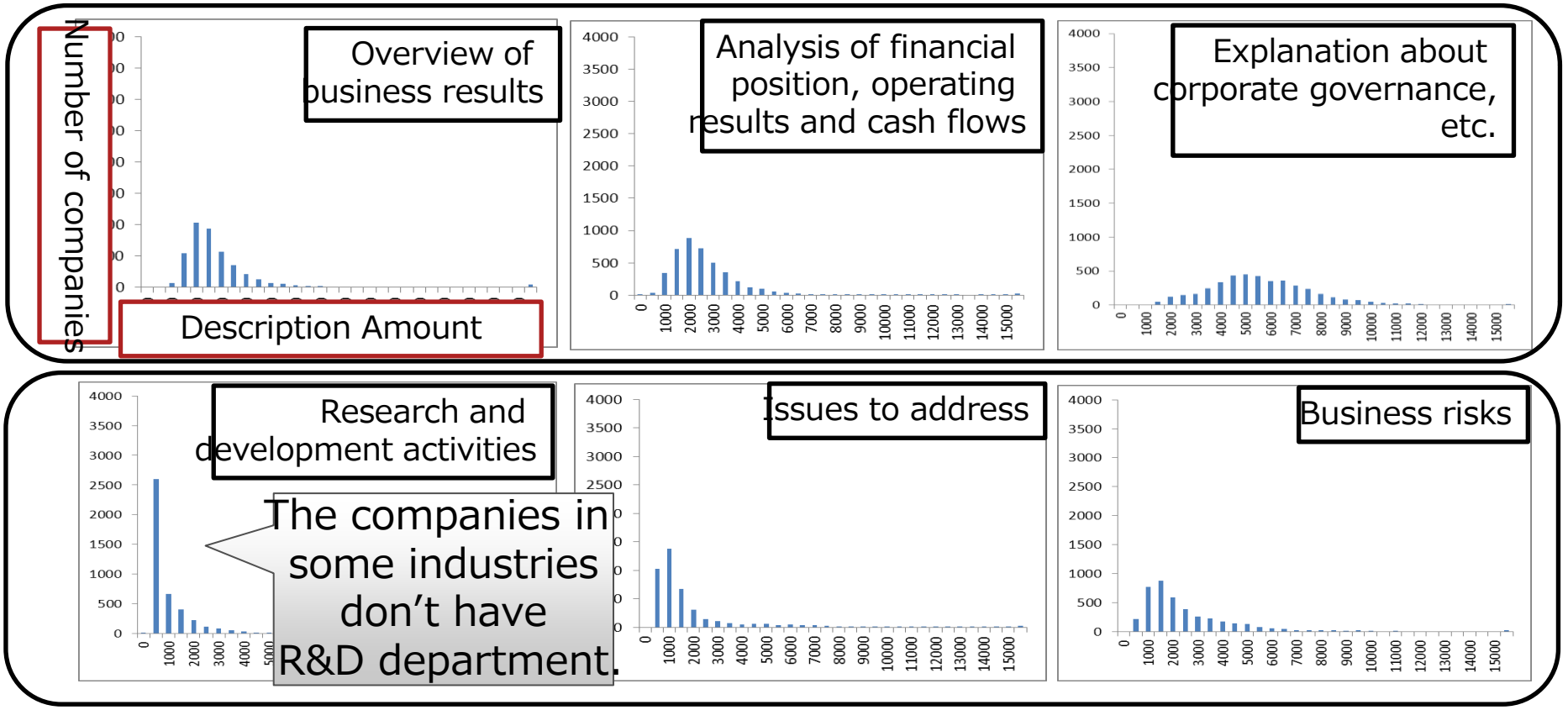
Overview of "Business risks" Analysis among Industries



Comparing Company X and Y
Comparing FY 2007 and 2008

Comparing Sections by Description Amounts

- Upper 3 figures indicate that many companies tend to describe a large amount of text.
- Lower 3 figures indicate that many companies tend to describe a small amount of text.



Analysis 1: Analysis of Differences among Industries

FUJITSU

- Purpose
  - To reveal differences among industries
- Target Text
  - Section “Business risks” that describes about business risks that the company takes
- Text mining tool
  - “Complex Skeleton Map” to visualize relations among keywords represent “Industry” and keywords related with “Risk”

Analysis 1: Analysis of Differences among Industries

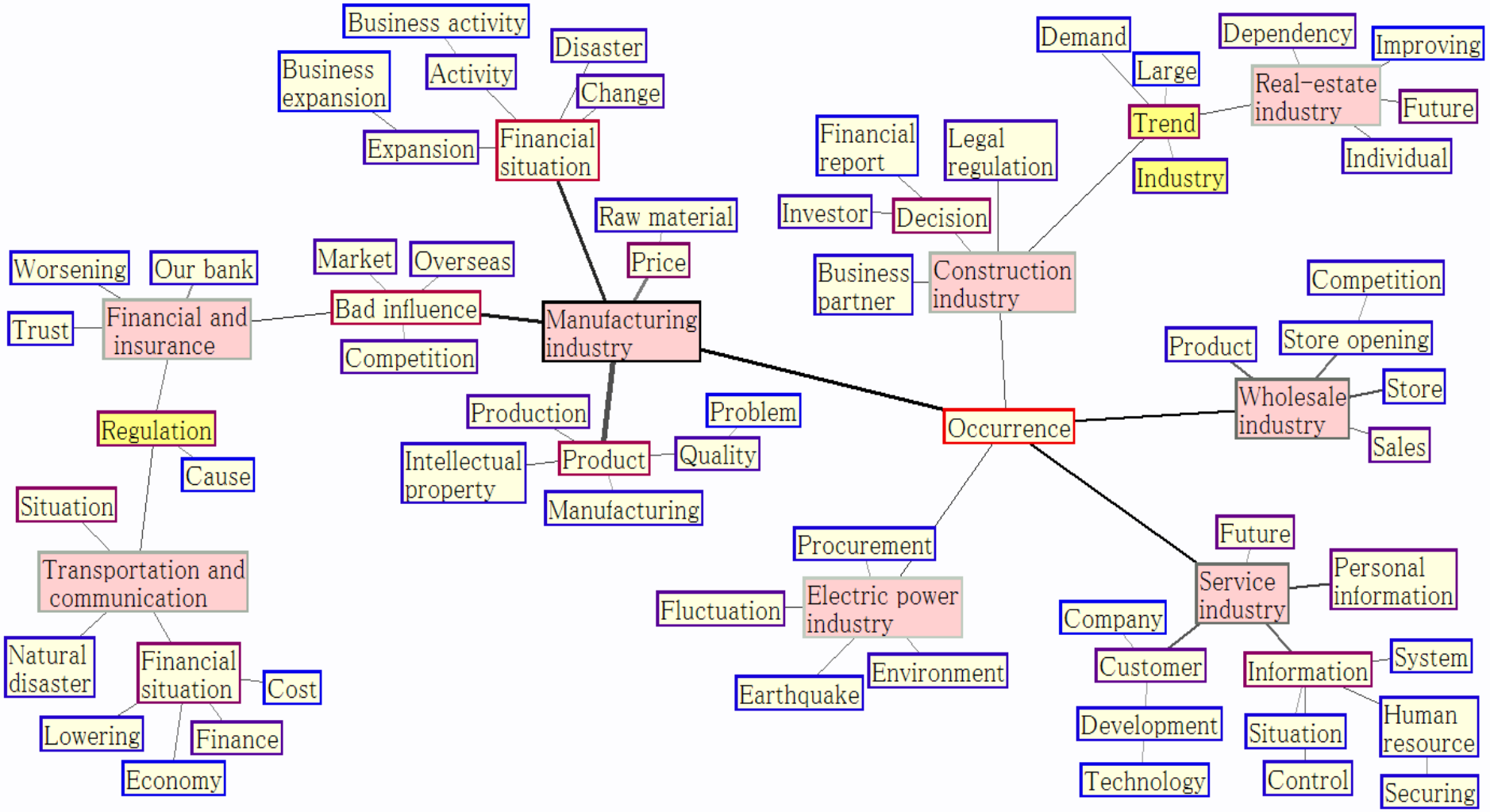
- The characteristic keywords(“Disaster”, ”Facility” in Electric Power) of each industry are ranked lower than generic keywords(“Influence”, “Possibility” and so on).
- It is unclear whether there is any relation among industries.

Characteristic keywords in each industry

	Manufacturing	Service	Wholesales	Transportation &Communication	Finance and insurance	Construction	Real estate	Electric power
1	Influence	Influence	Influence	Influence	Influence	Influence	Possibility	Influence
2	Possibility	Possibility	Possibility	Possibility	Possibility	Possibility	Influence	Possibility
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3	Decision	Business	Decision	Business	Risk	Decision	Business	Business
11	Occurrence	Important	Important	Important	System	Work	Occurrence	System
12	Fluctuation	Investment	Sales	Fluctuation	Fluctuation	Investment	Interest	Decision
13	Sales	Information	Product	Future	Trust	Situation	Building	Disaster
14	Price	Future	Description	Investment	Important	Important	Investment	Information
15	Manufacturing	Control	Control	Regulation	Situation	Price	Description	Facility

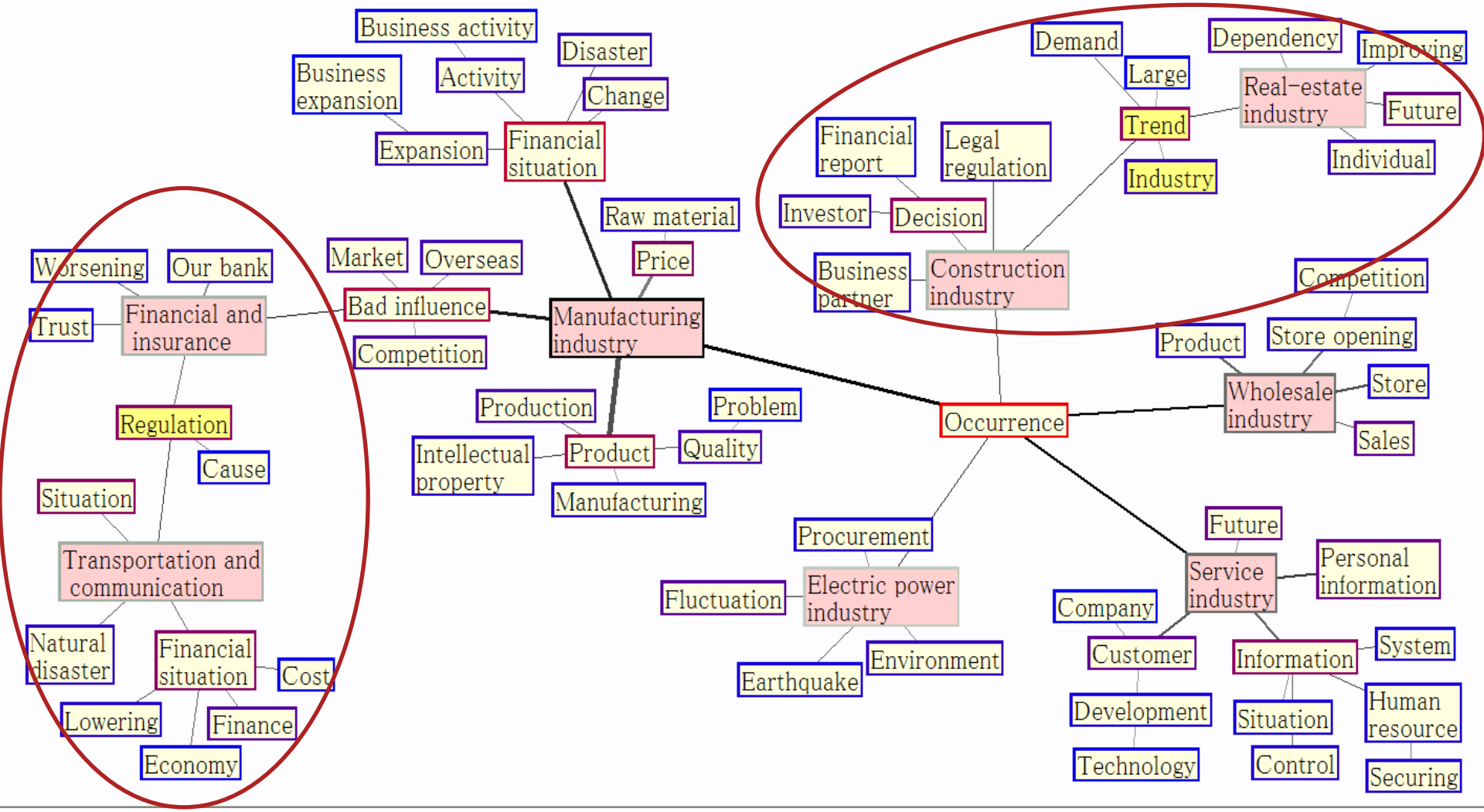
Analysis 1: Analysis of Differences among Industries

- By using “Complex Skeleton Map”, we can figure out relations among industries.



Analysis 1: Analysis of Differences among Industries

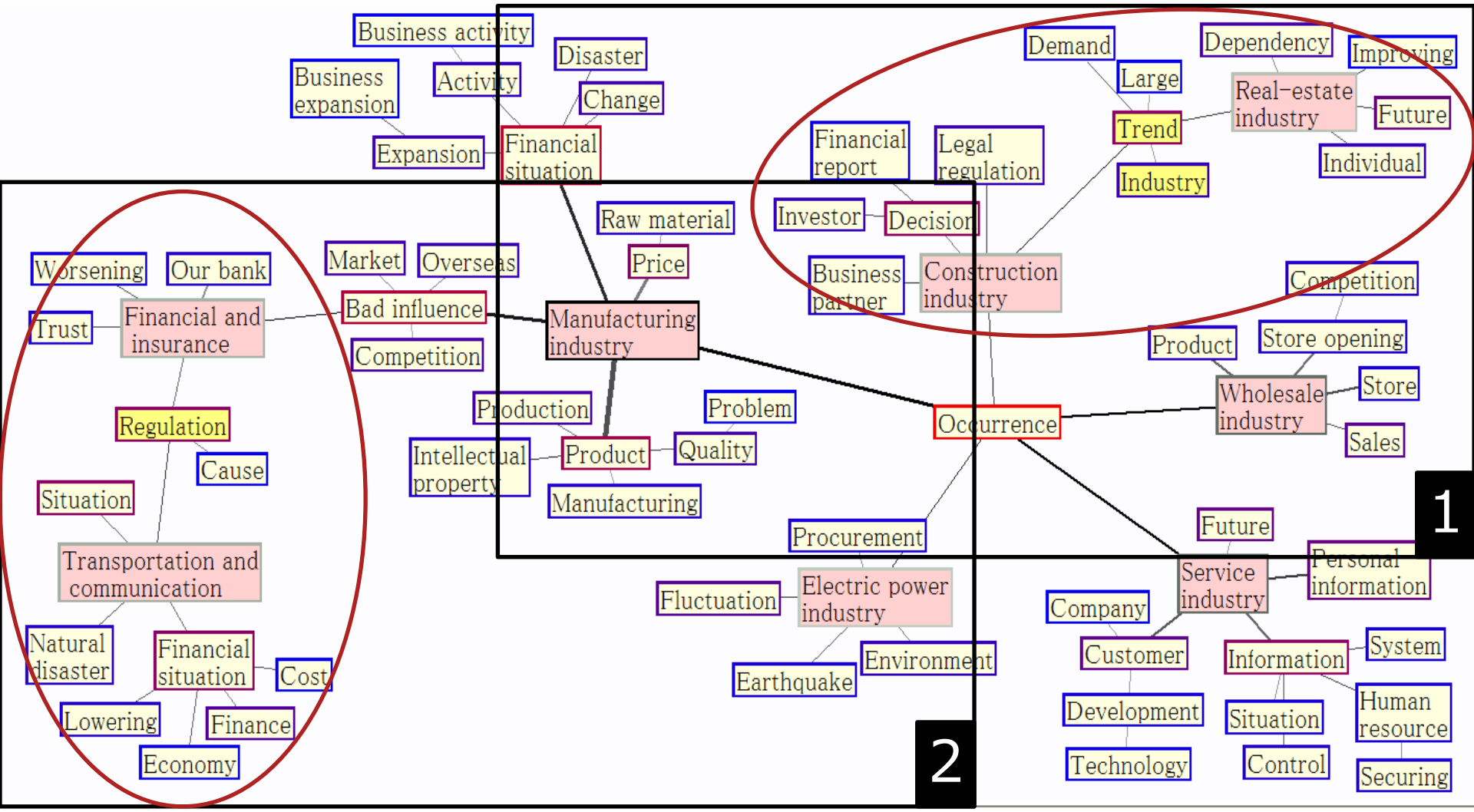
- We can figure out that there are some pairs of industries which have common risks.



Analysis 1: Analysis of Differences among Industries

FUJITSU

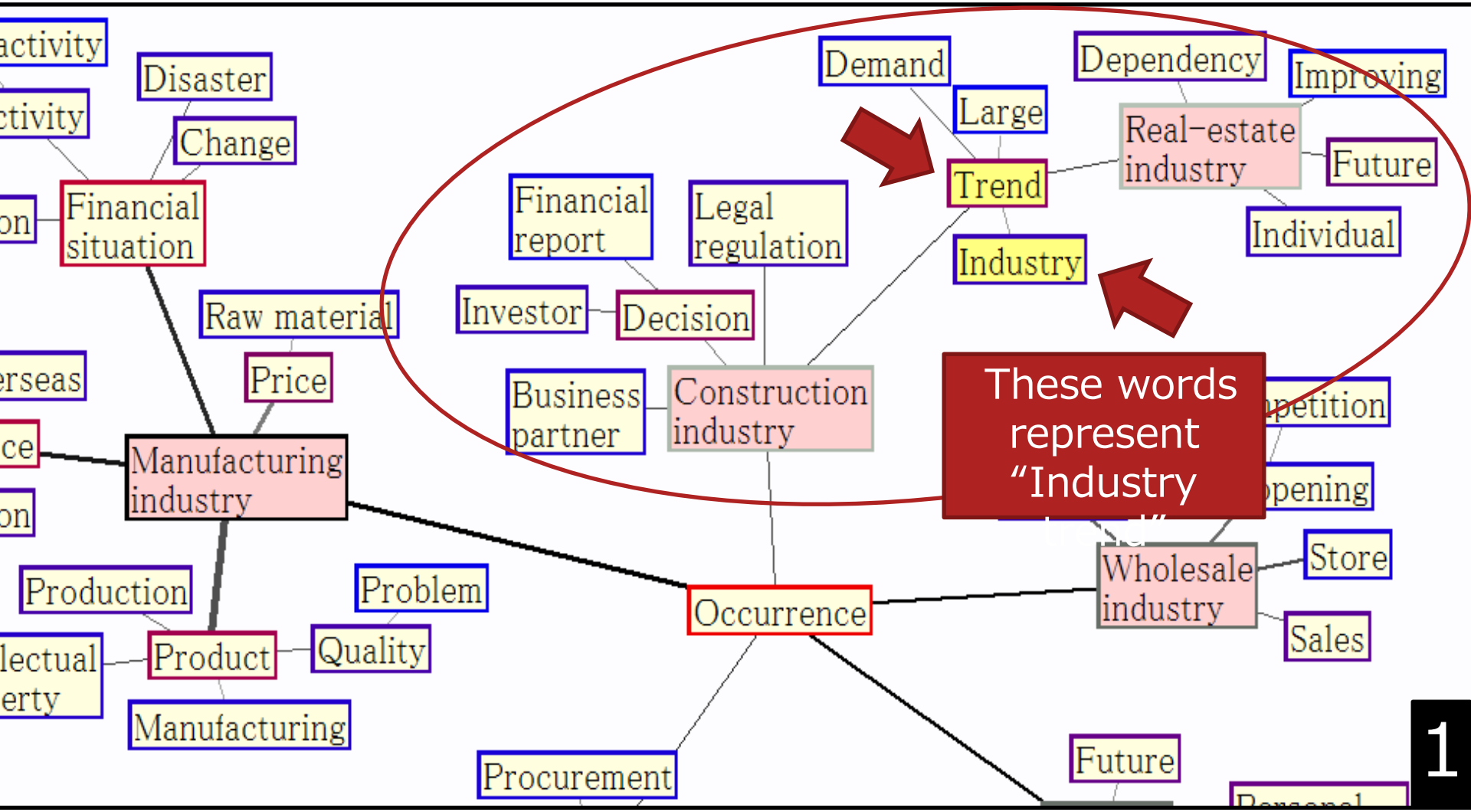
- We can figure out that there are some pairs of industries which have common risks.



Analysis 1: Analysis of Differences among Industries

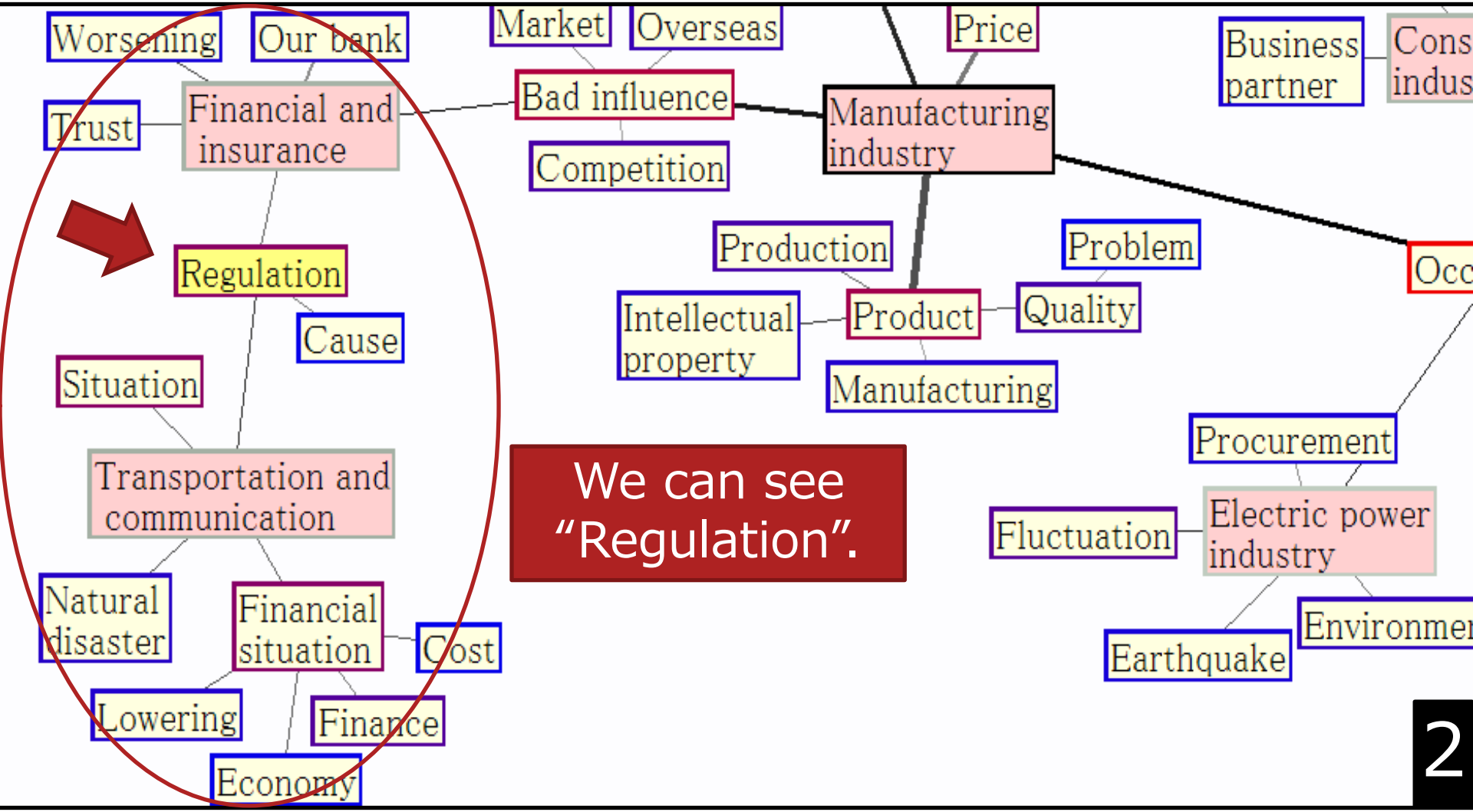
FUJITSU

- "Real estate" and "Construction" are influenced strongly by "industry trends"



Analysis 1: Analysis of Differences among Industries

- "Finance and insurance" and "Transportation" are influenced strongly by "regulations".



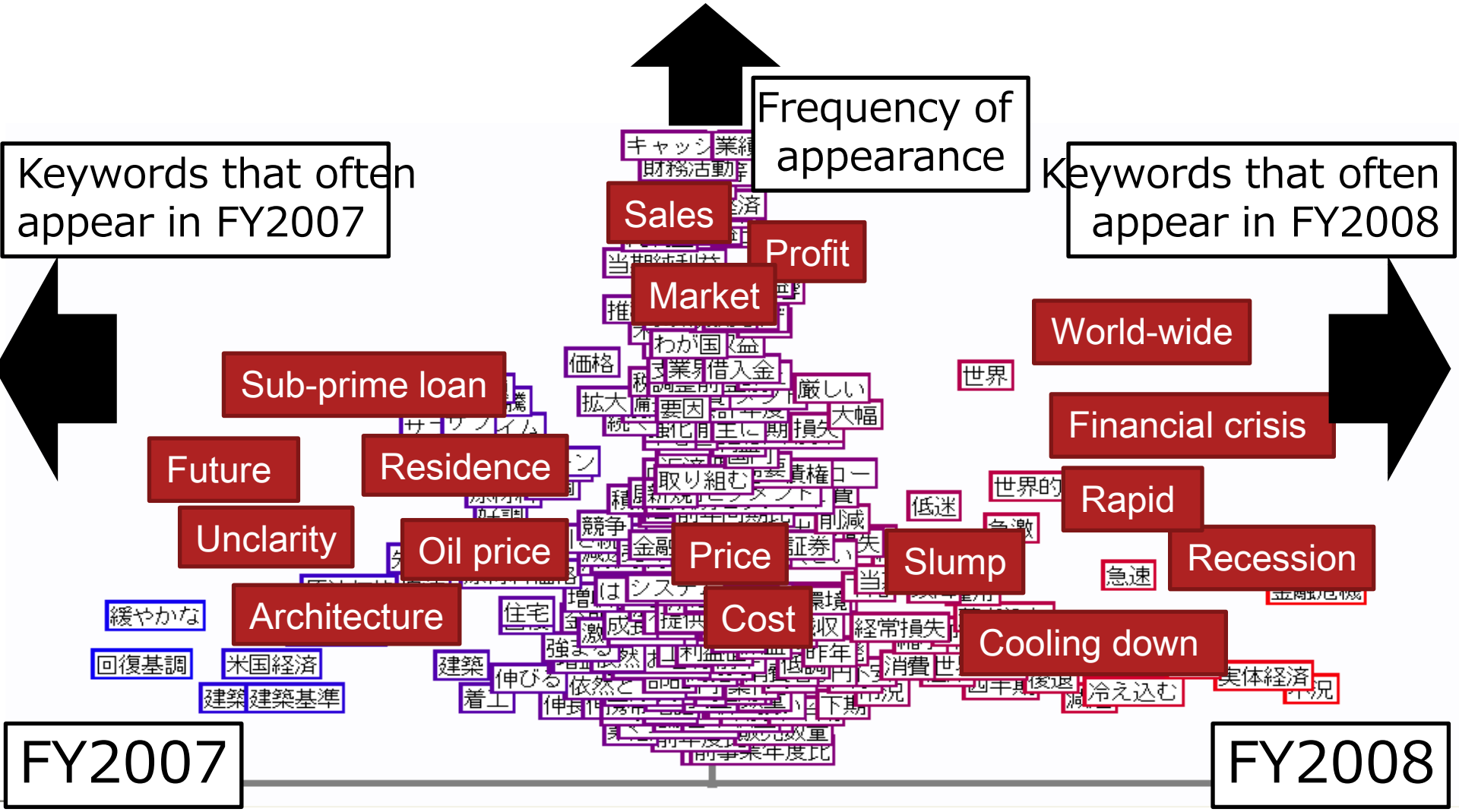
Analysis 2: Analysis of an Event Influence

FUJITSU

- Purpose
  - To reveal the influence of an event
- Target Text
  - Section “Overview of business results” that describes about business circumstance around the company
- Text mining tool
  - “Comparison Map” to compare a text group and another text group

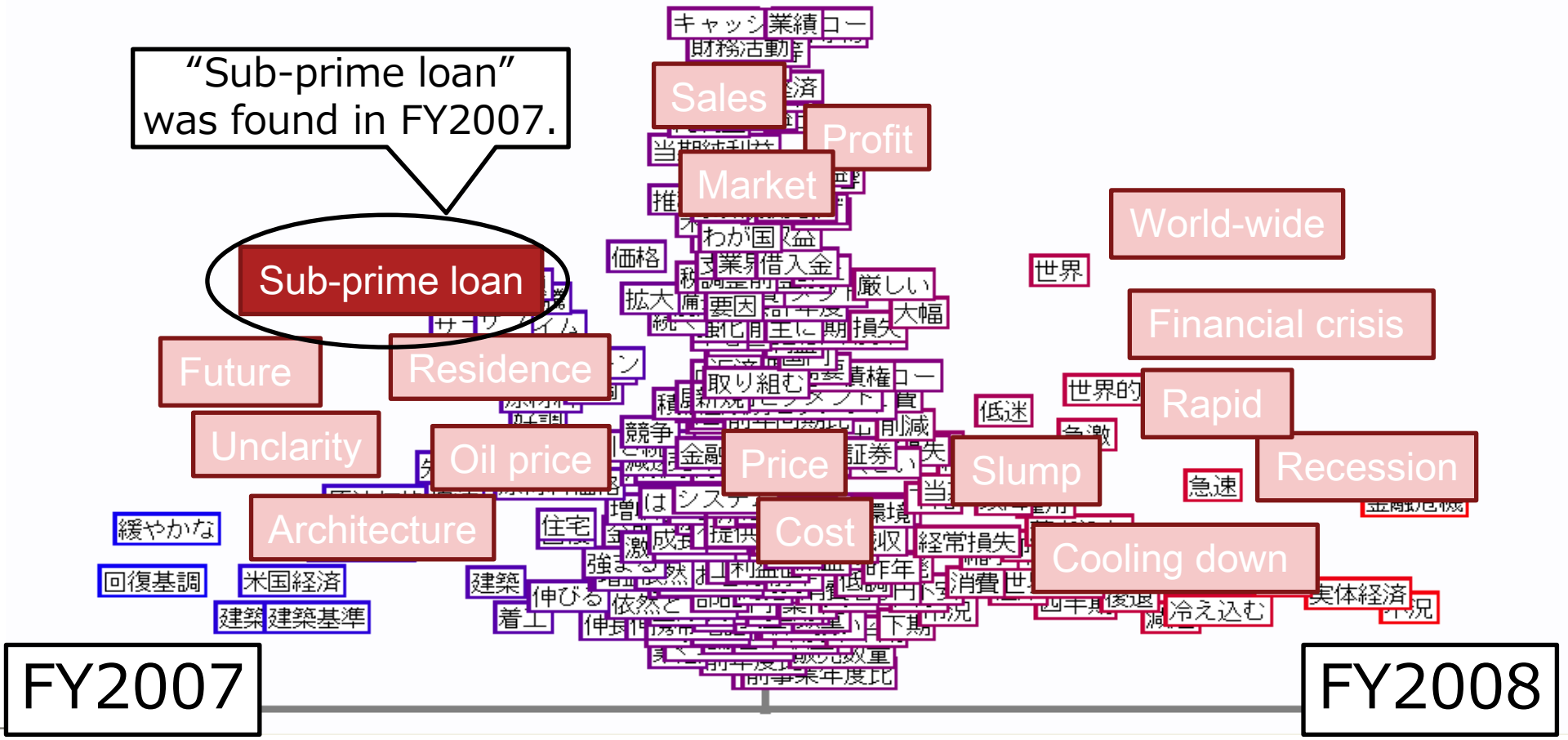
Analysis 2: Analysis of an Event Influence

- By comparing FY2007 and FY2008, we can figure out the influence of “World Finance Crisis”.



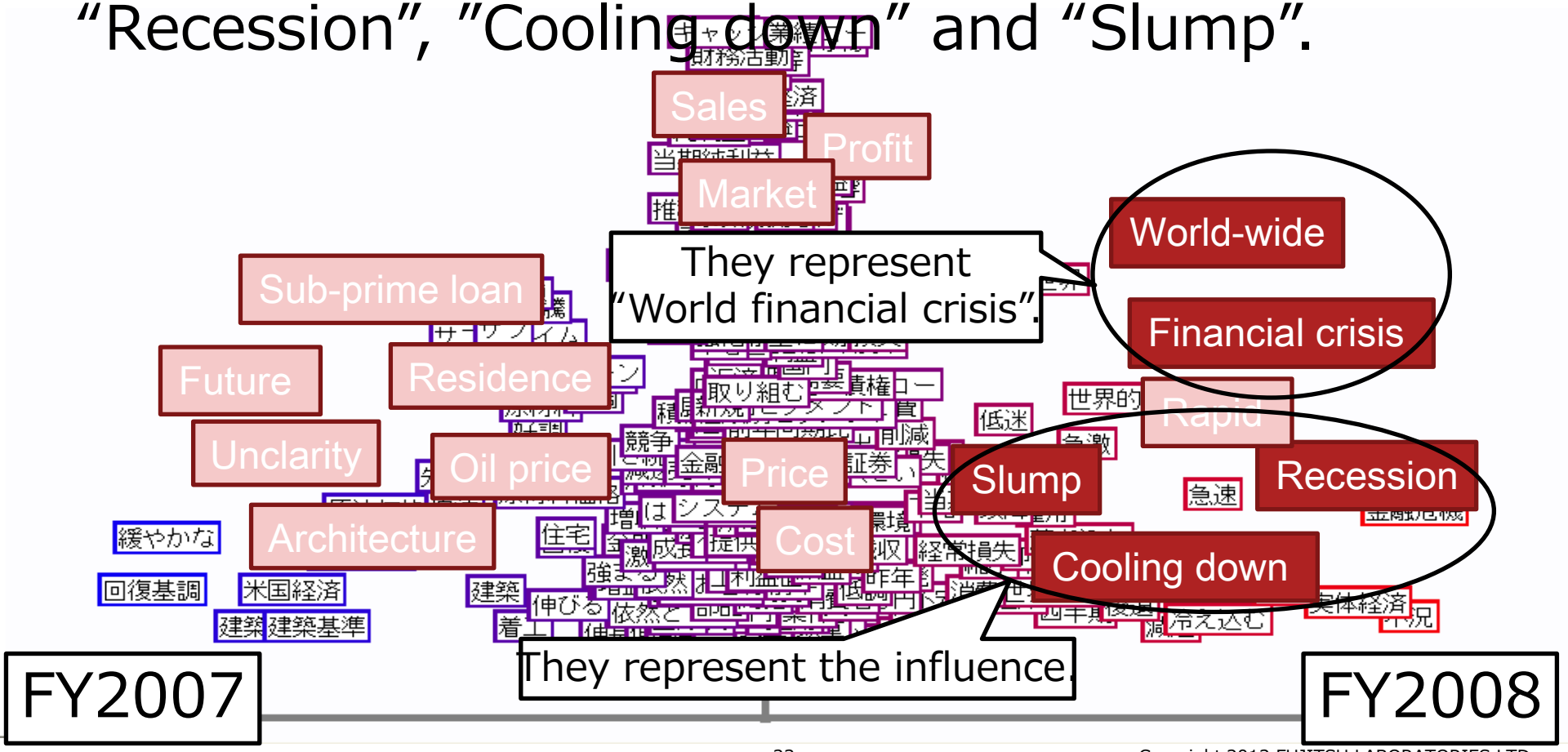
Analysis 2: Analysis of an Event Influence

- "World financial crisis" started from "Sub-prime loan problem" that occurred in FY2007.
- We can find "Sub-prime loan" in keywords of FY2007.



Analysis 2: Analysis of an Event Influence

- “Sub-prime loan problem" spread into “World financial crisis" in FY2008.
- We can figure out the influence of “World Finance Crisis” that was indicated by “Recession”, ”Cooling down” and “Slump”.



Summary of Part2

FUJITSU

- Comparison Analysis of Annual Securities Reports
  - By using “Business risks” and “Complex Skeleton Map”, differences and similarities among industries are revealed well.
  - By using “Overview of business results” and ”Comparison map”, the influence of Global Financial Crisis is revealed well.
- The analyses with the following viewpoints are effective.
  - Analysis that focuses on a specific section such as “Business risks"
  - Analysis of differences and similarities among the companies group such as "industry".
  - Analysis that focuses on "fiscal year"

FUJITSU

PART3: Discussion

Scope Extension of XBRL in FY2013

- The target of XBRL will be extended to all sections.

XBRL Scope from FY 2013

All Sections (Including Text Parts)

Extension

XBRL Scope until FY 2012

Only Primary Financial Statement

1.Overview of company	1.Summary of business results
	2.Company history
	3.Description of business
	4.Overview of group entities
	5.Information about employees
2.Overview of business	1.Overview of business results
	2.Overview of production, orders received and sales
	3.Issues to address
	4.Business risks
	5.Critical contracts for operation
	6.Research and development activities
	7.Analysis of financial position, operating results …
[illegible]	1.Overview of capital expenditures, etc.
	2.Major facilities
	3.Planned additions, retirements, etc. of facilities
4.Information about reporting company	1.Information about shares, etc.
	2.Acquisitions, etc. of treasury shares
	3.Dividend policy
	4.Historical records of share price
	5.Information about officers
	6.Explanation about corporate governance, etc.
5.Financial information	1.Consolidated financial statements, etc.
	2.Financial statements, etc.
6.Overview of operational procedures for shares	
7.Reference information	1.Information about parent company, etc. …
	2.Other reference information

Expectation for the New XBRL Specification

- By splitting sections, we could conduct following 2 analyses that focused on a specific section.
  - Analysis of differences among industries using “Business risks”
  - Analysis of an event using “Overview of business results”
- If we don’t split data, we could only use aggregation table over all sections and only see section-mixed keywords.
- In the preprocessing, we splitted reports with section labels.
  - Investigating the variation of labels
  - Aggregating labels
- We can analyze from a macro perspective.

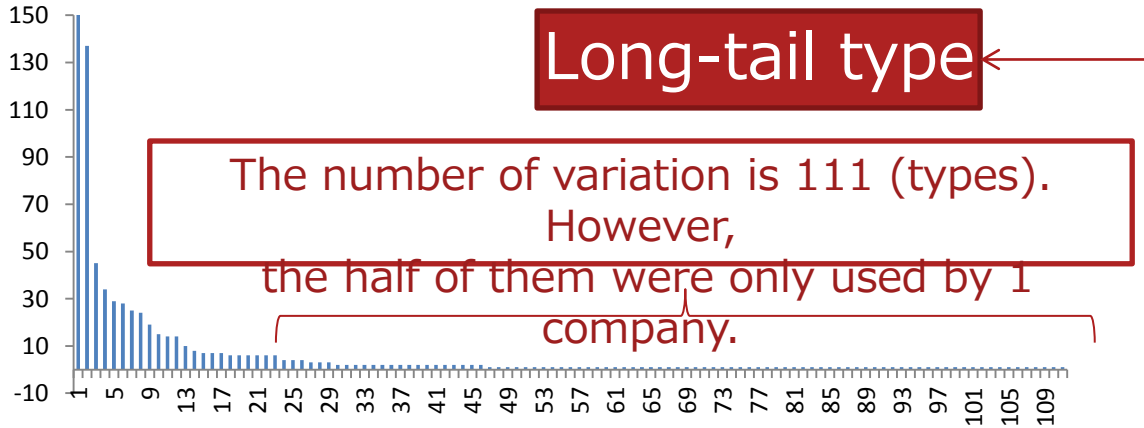
Examples : variation of the labels

Overview of production, orders received and sales	3034
Overview of purchases and sales	137
Overview of sales	45
Overview of sales and purchases	34
Overview of production and sales	29
Overview of purchases, received and sales	28
Overview of received and sales	25
Overview of productions, purchases	15
Overview of received and sales	10
Overview of production and	

Top 10 covers almost

Expectation for the New XBRL Specification

- There is no problem in most cases. However, in the special case such as analysis from a micro perspective (analysis with a small amount of reports), or searching of reports without omissions, there are limitations.
- The distribution over labels is long-tail type, hence label aggregation is not easy. However, the problems will be solved by the new XBRL specification which prepares tags of sections completely.



Examples : variation of the labels

Overview of production, orders received and sales	3034
Overview of purchases and sales	137
Overview of sales	45
Overview of sales and purchases	34
Overview of production and sales	29
Overview of purchases, received and sales	28
Overview of received and sales	25
Overview of productions, purchases and sales	15
Overview of received and sales	10

Analysis Patterns on XBRL Reports

FUJITSU

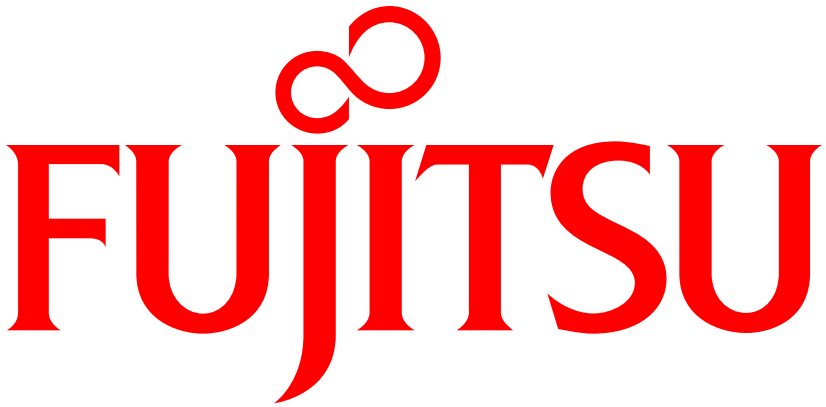
- Analysis that uses only textual data
  - The analysis that focuses on “Business risks”, to figure out what kinds of risks are considered among companies
  - The relation between “Business risks” and “Issues to address”, to reveal whether a company takes measures
- Analysis that combines textual data and numerical data
  - To reveal whether a company gets returns of research investment, with the use of the relation between “Research and development activities” and R&D expense in Financial Statement
- Analysis that combines XBRL reports and other data
  - To rate patents evaluating how much the business

Analysis Patterns on XBRL Reports

FUJITSU

- Analysis with sequential analyzing technology applying to non textual data
  - To reveal the propagation of the impact of a bankruptcy (chain bankruptcy) with the use of dealing relations
- Analysis that uses text of section that does not mainly describe about narrative information
  - To help bank to finance by automatic screening with the use of explanatory notes to Financial Statements

We would like other opinions about analysis patterns of the effective use of XBRL Reports with text mining.



shaping tomorrow with you