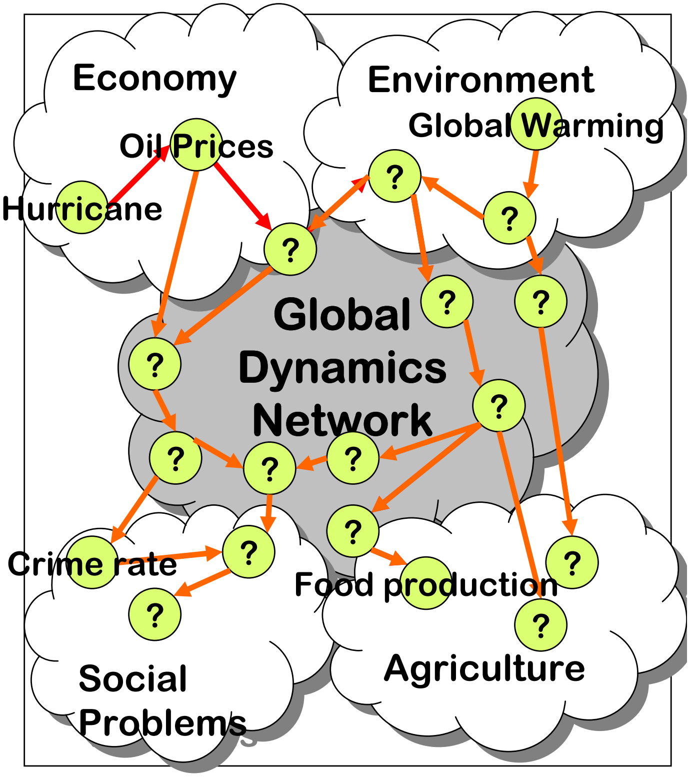
Visualization for Statistical Term Network in Newspaper

Hideki Kawai[†], Kazuo Kunieda[†], Keiji Yamada[†],
Haruka Saito, Masaaki Tsuchida, Hironori Mizuguchi
[†] NEC C&C Innovation Research Laboratories
NEC Common Platform Software Research Laboratories

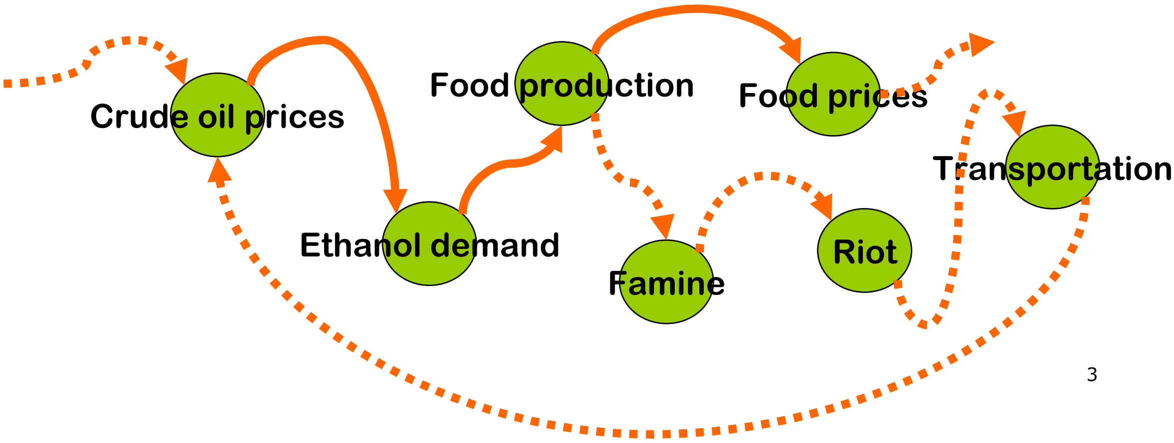
Background

- Complex relations between various problems
- Causal relations
  - Butterfly effect
  - Ripple effect
- Global Dynamics
  - Global Solution
  - Idea Support
- Focusing on Statistical Terms



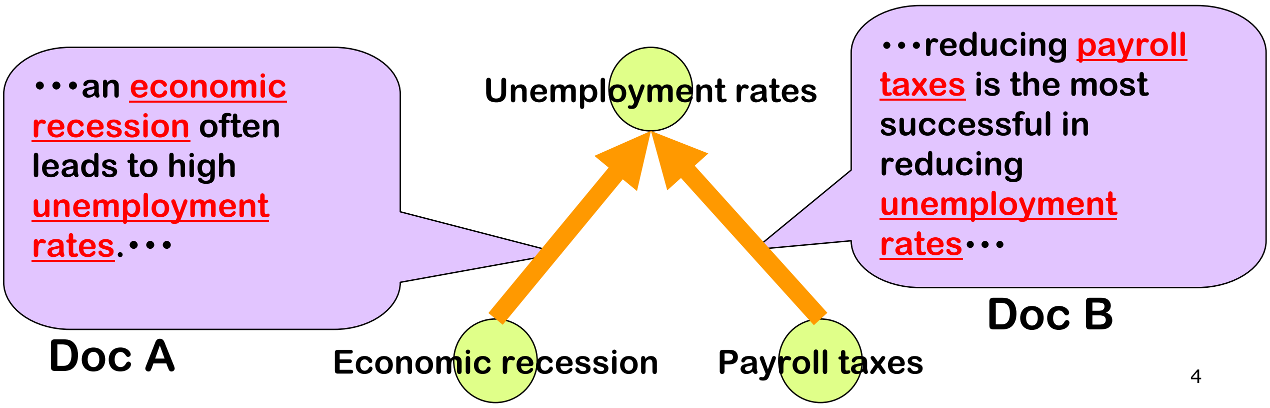
Motivation

- Web contains several billions of text pages
- Text contains information like causal relations
  - Ex1: "the rise in crude oil prices stimulated ethanol demand"
  - Ex2: "farmers to answer to the ethanol demand, leave less land available to food production"



Goal

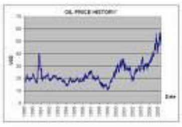
- Visualize Global Dynamics as a Statistical Term Network
  - Node：Statistical terms
  - Edge：Relationship between terms


…an economic recession often leads to high unemployment rates.…

…reducing payroll taxes is the most successful in reducing unemployment rates…

What are we looking for?

- Nodes: Expressions
  - Statistical terms (relative to a movement)



    - Ex. Unemployment rates, birth rates, crime rates, oil prices, corn prices
    - **Suffixes**: Rates, prices, costs
  - Events & Facts

    - Ex. Hurricanes, riot, war
    - Ex. Urbanization, sustainable resources, global warning
    - **Classes**: natural disasters, social, economics

> ➢ Extraction with the help of suffix & class patterns

What are we looking for?

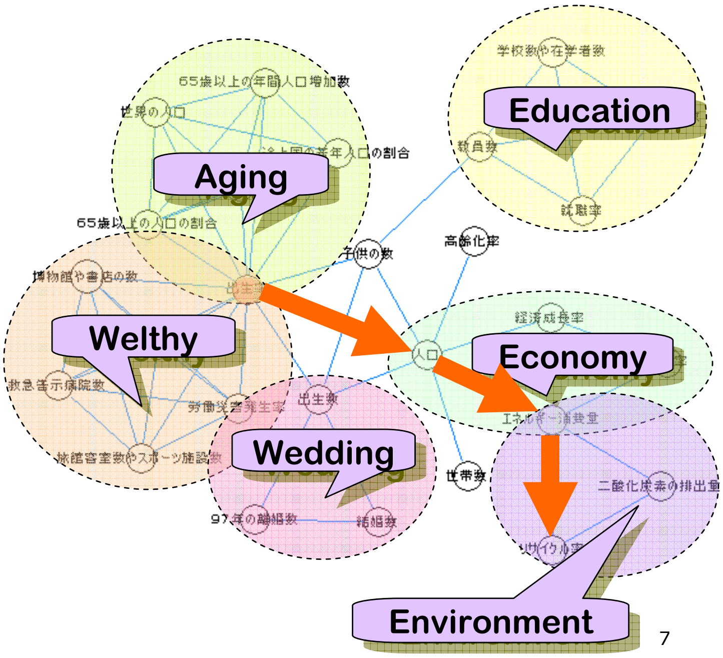
- Edges: Relations
  - Link between a cause and its effect
  - In text: Cause & effect are expressions linked via some verbs.
  - For causal relation, in 80% of case: cause & effect are in the same sentence [1]

> ➢ We will extract sentences containing a couple of expressions

[1] Takashi Inui, *Creating an Annotated Corpus for the Analysis of Causal Relations*

Previous Work [NTCIR-6, 2007]

- Suffix-based Statistical term extraction
  - 「失業<u>率</u> (unemployment <u>rates</u>)」
  - 「原油<u>価格</u> (oil <u>prices</u>)」
  - 「物価<u>指数</u> (consumer price <u>index</u>)」
- Co-occurrence network of statistical terms
  - Manually constructed



Suffix-based Statistical Term Extraction

- 1) Find a suffix of statistical terms
  - ・・・ 増え て おり 、 ビール の 出荷 <u>台数</u> が ・・・ ○
  - ・・・ ある 場合 では テレビ の 生産 <u>台数</u> 競争 ・・・ ×
- 2) Scan leftward from the starting point to find the morpheme which is neither noun nor specific particles
  - ・・・ 増え て おり 、 ビール の 出荷 <u>台数</u> が ・・・

    End ← Start
- 3) Extract morphemes between starting point to ending point as statistical terms
  - ・・・ 増え て おり 、 ビール の 出荷 <u>台数</u> が ・・・

    Statistical Term

Semantic Structure of Statistical Terms

- Base Form
  - Shortest sequence of morphemes having a statistically valid meaning
    - Unemployment rates, Oil prices

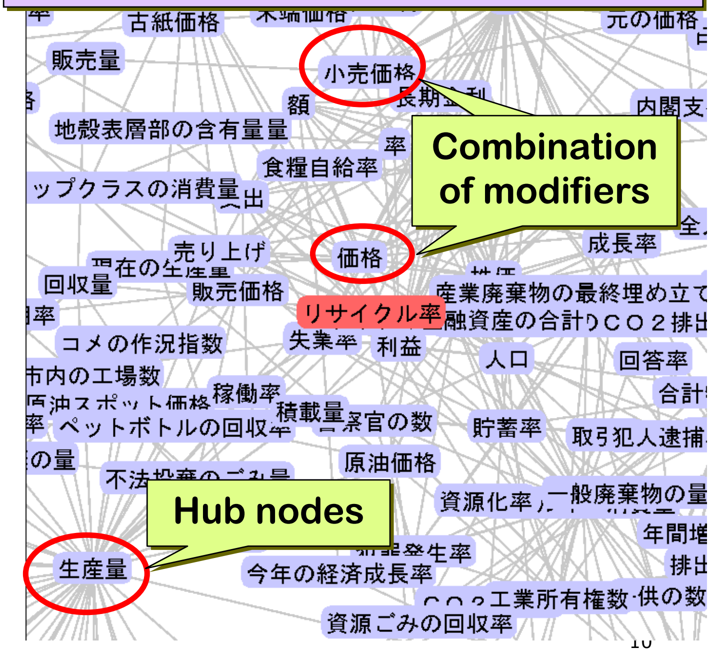
Various combinations:
「unemployment rates」
「domestic unemployment rates」
「American unemployment rates in Mar. 1998」

- Modifiers
  - Object
    - What is measured
      - Beer、PC
  - Subject
    - Who measured
      - Kirin、NEC
  - Time Span
    - When was it measured
      - 1999、Feb.
  - Region
    - Where was it measured
      - Japan、America

Example of Statistical Term Network

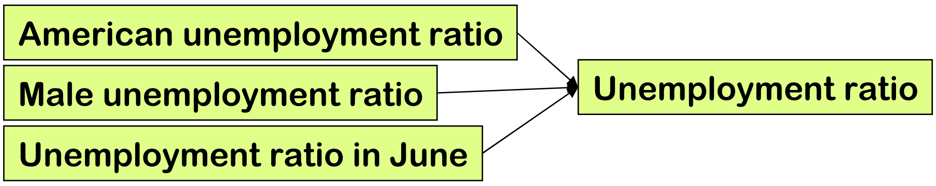
- Too High Density
  - Difficult to see the relationship between statistical terms
- Reasons
  - Hub nodes
    - price、volume of production
  - Combination of modifiers
    - price、retail price

2 hops from “recycle ratio”



Challenges

- Complexity of Network Structure
  - High dense network
    - Clustering generates one big cluster
    - Threshold of co-occurrence does not work because most co-occurrence of statistical terms are only 1 or 2.
- Complexity of Semantic Structure
  - Appropriate combination of modifiers is not clear
    - Can we integrate?



American unemployment ratio

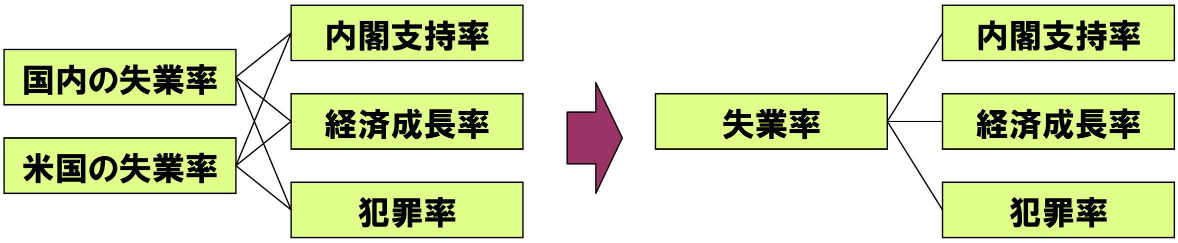
Male unemployment ratio

Unemployment ratio in June

Unemployment ratio

Our Approach

- Simplify the Network Structure
  - Limit the degree of nodes
    - Link only top ω terms
    - Main structure can be observed
- Simplify the Semantic Structure
  - Integrate statistical terms which have common base form and co-occurring terms



国内の失業率

米国の失業率

内閣支持率

経済成長率

犯罪率

失業率

内閣支持率

経済成長率

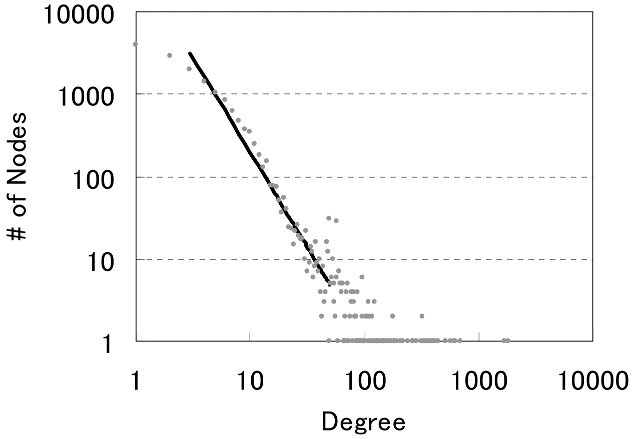
犯罪率

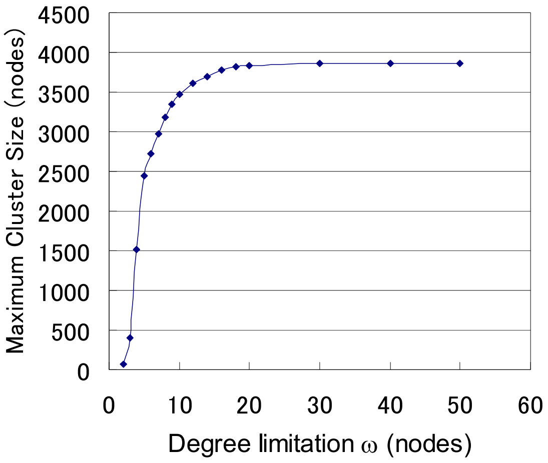
Experimental Settings

- Suffix dictionary
  - 86 statistical terms tagged in MuST corpus
- Mainichi News 1998-1999 (Japanese)
- Extracted terms：8,600 words
- Degree parameter ω
  - Investigated Maximum cluster size with ω
- Visualization Tool
  - prefuse (http://prefuse.org/)

Degree Limitation Parameter ω

- Degree distribution of statistical terms follows power law
- Smaller Limitation Parameter ω divides the original network into small pieces
- We chose ω = 10 for the visualization.





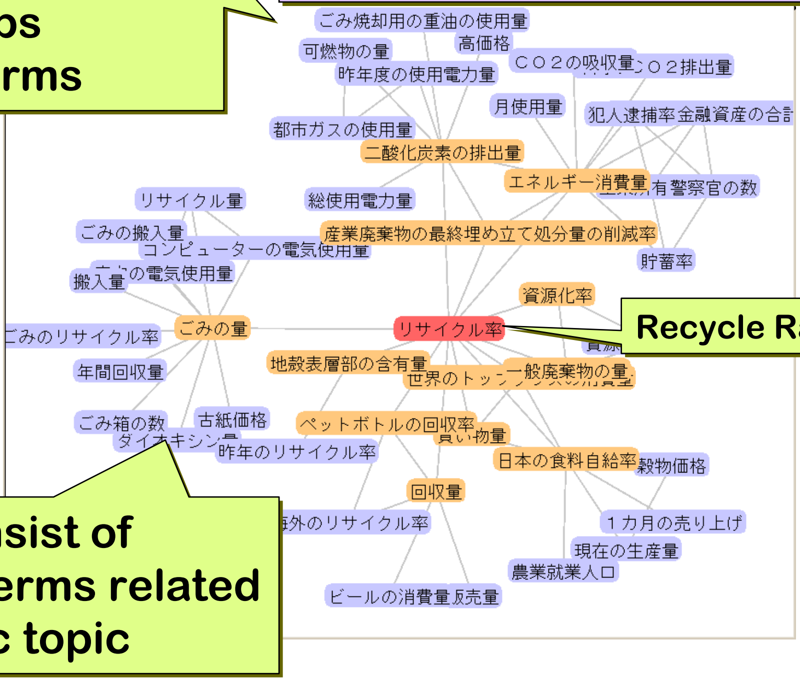
Result of Semantic Structure Simplification

Original Expression	Simplified Expression
わが国の温室効果ガスの総排出量 Domestic total greenhouse gas emission	**温室効果ガスの総排出量** Total greenhouse gas emission
96年の温室効果ガスの総排出量 Total greenhouse gas emission in 1996	
12月のパソコン販売台数 PC sales volume in December	**パソコン販売台数** PC sales volume
秋葉原の電気街のパソコン販売台数 PC sales volume in Akihabara	
埼玉県所沢市の野菜の価格 Vegetable prices in Tokorozawa city Saitama	**野菜の価格** Vegetable Prices
すべての野菜の価格 All vegetable prices	

Result of Network Structure Simplification

Easier to observe relationships between terms

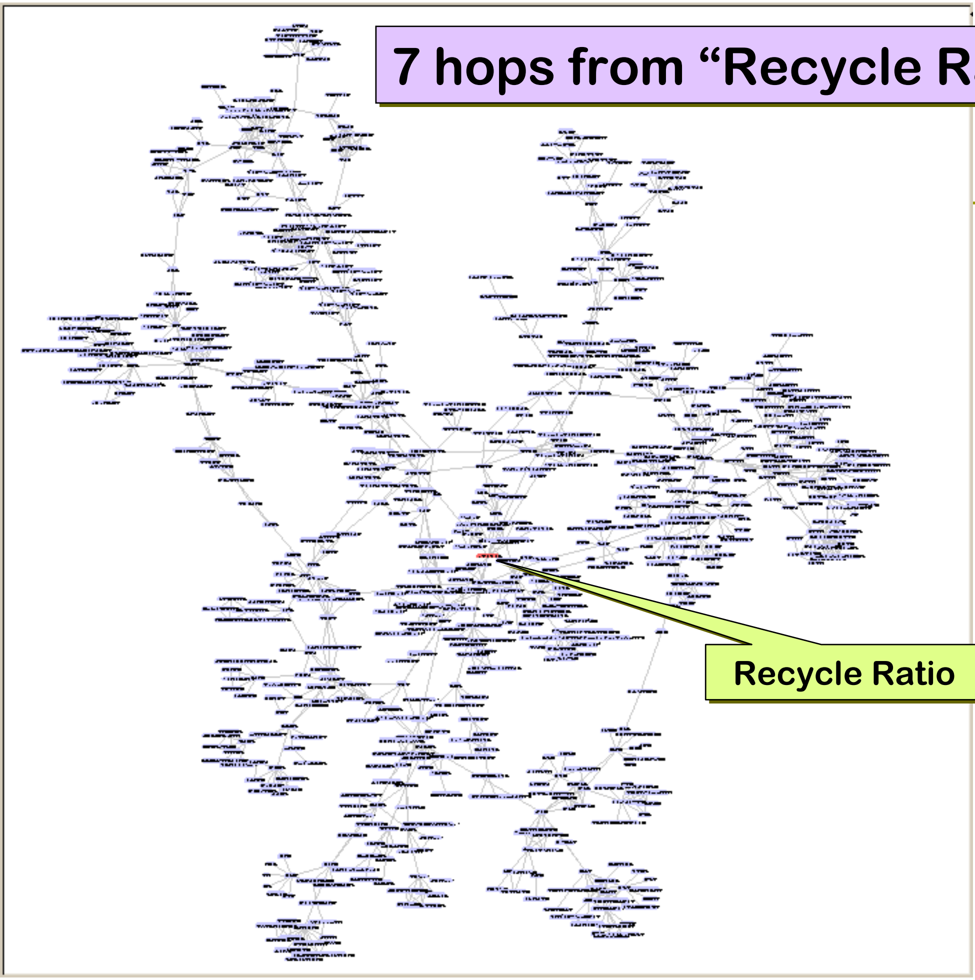
2 hops from "Recycle Ratio"



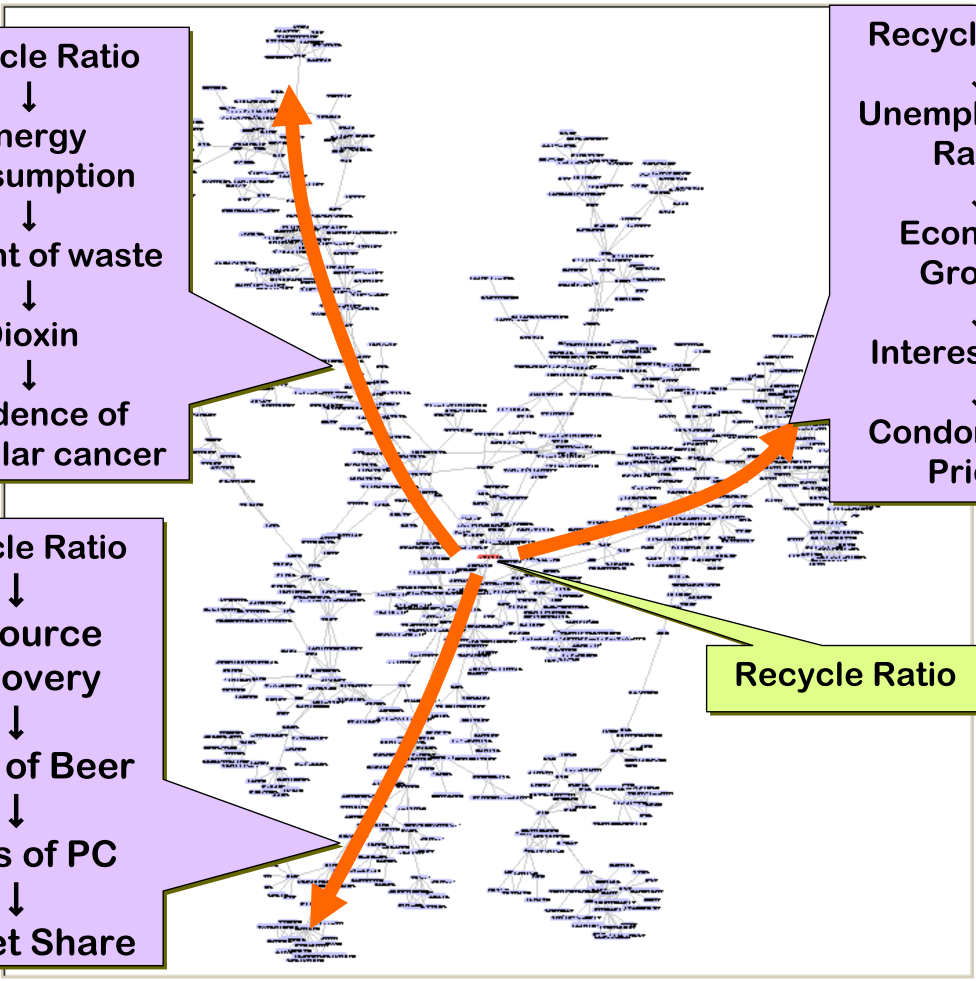
Recycle Ratio

Cliques consist of statistical terms related to a specific topic

7 hops from “Recycle Ratio”



Recycle Ratio



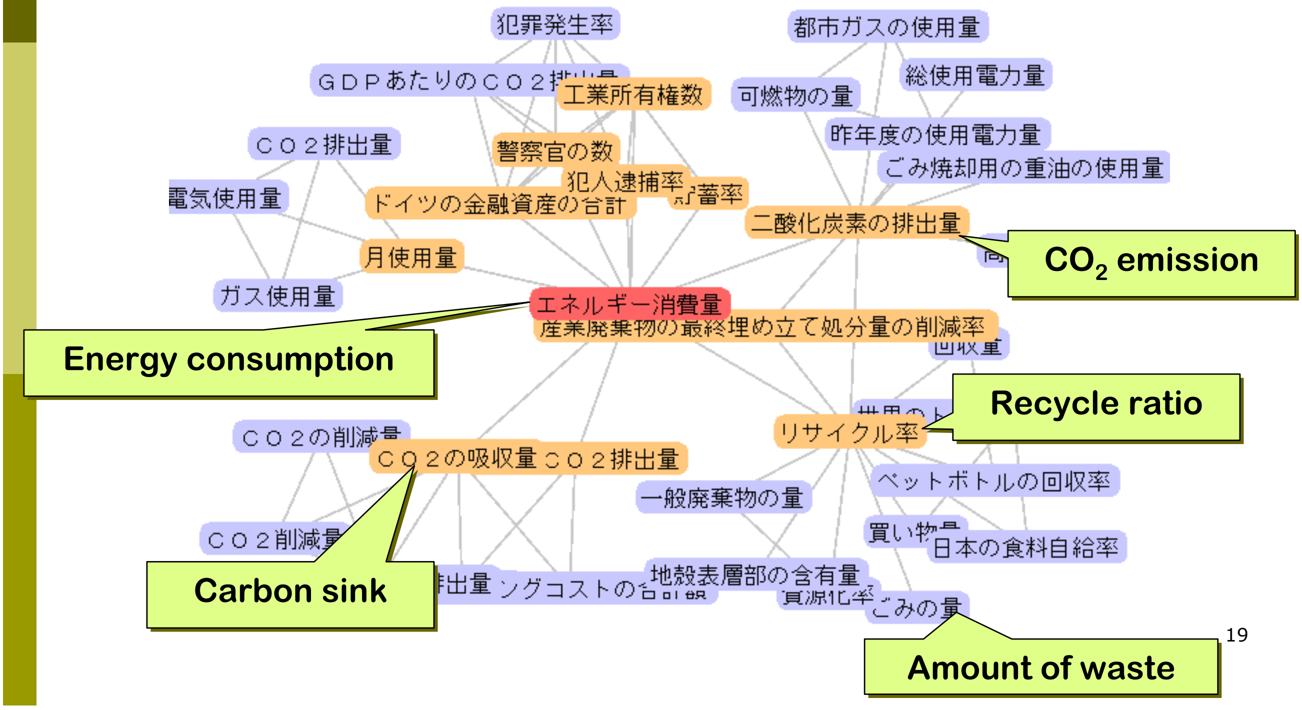
Recycle Ratio
↓
Energy Consumption
↓
Amount of waste
↓
Dioxin
↓
Incidence of testicular cancer

Recycle Ratio
↓
Unemployment Rates
↓
Economic Growth
↓
Interest rates
↓
Condominium Prices

Recycle Ratio
↓
Resource Recovery
↓
Sales of Beer
↓
Sales of PC
↓
Market Share

Recycle Ratio

Example: Energy Consumption



CO$_2$ emission

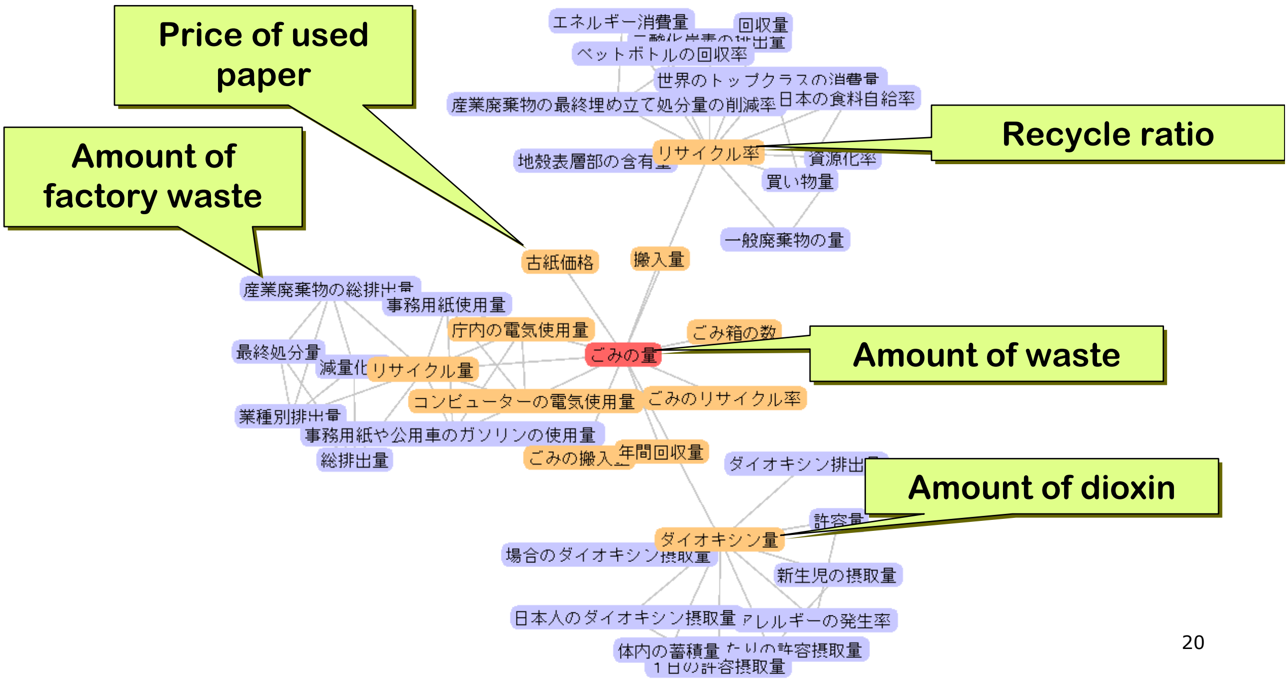
Energy consumption

Recycle ratio

Carbon sink

Amount of waste

Example: Amount of Waste



Price of used paper

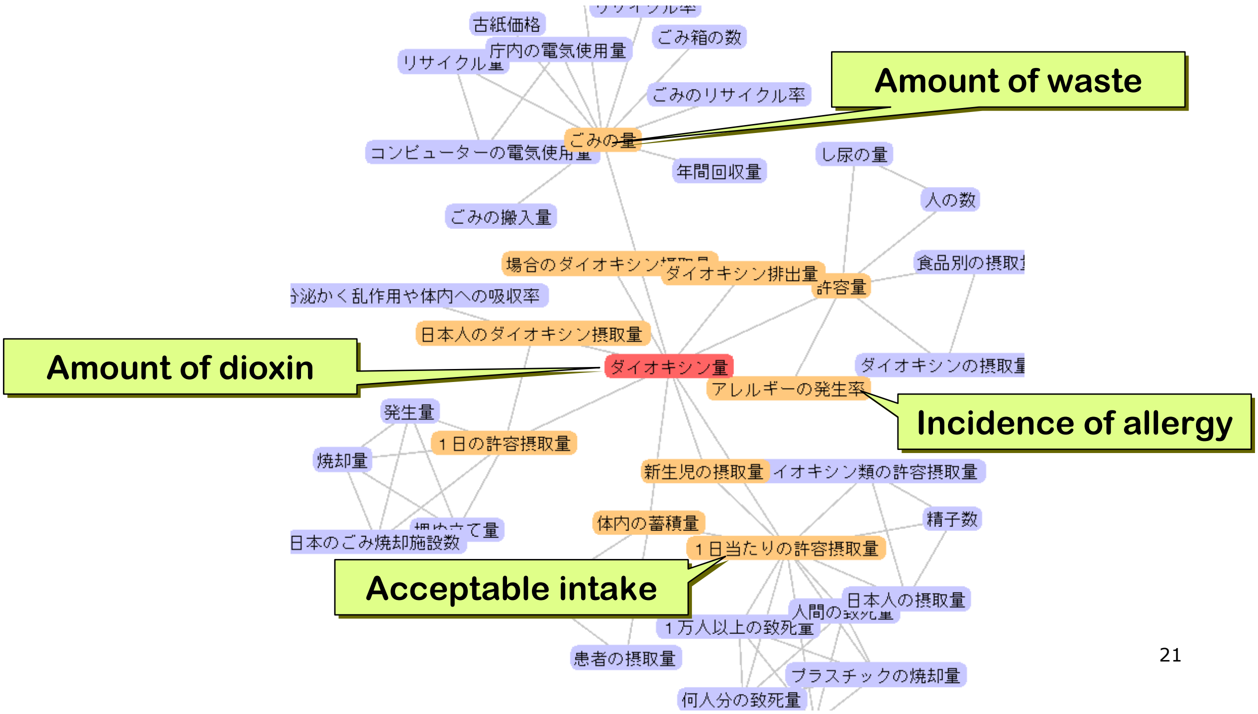
Amount of factory waste

Recycle ratio

Amount of waste

Amount of dioxin

Example: Amount of Dioxin



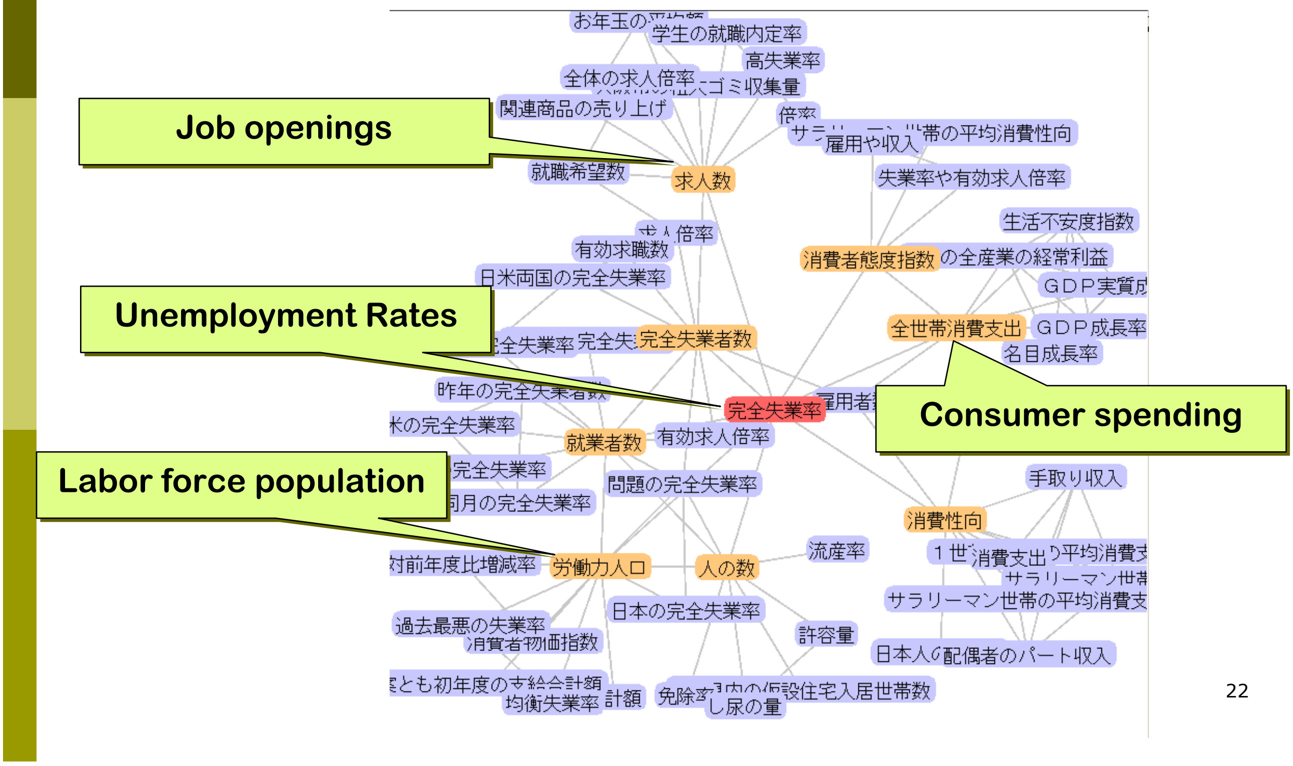
Amount of waste

Amount of dioxin

Incidence of allergy

Acceptable intake

Examples: Unemployment Rates



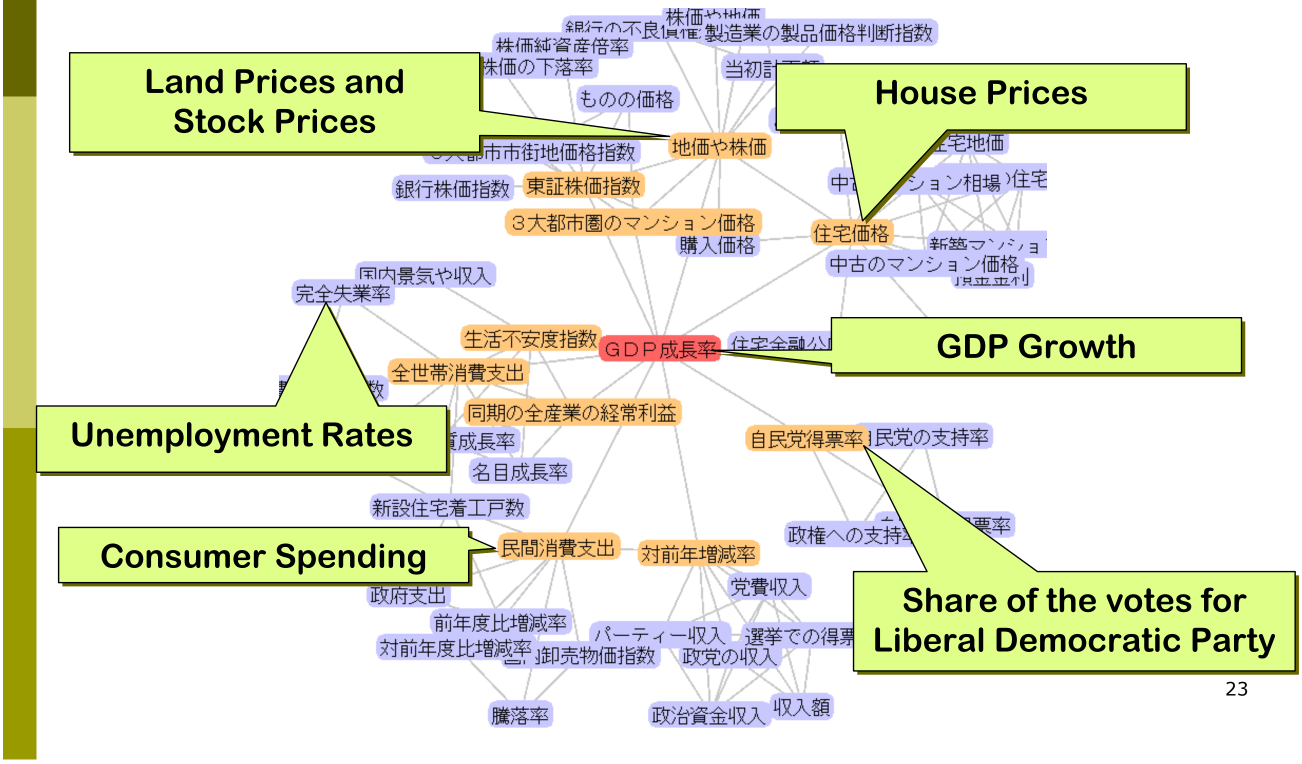
Job openings

Unemployment Rates

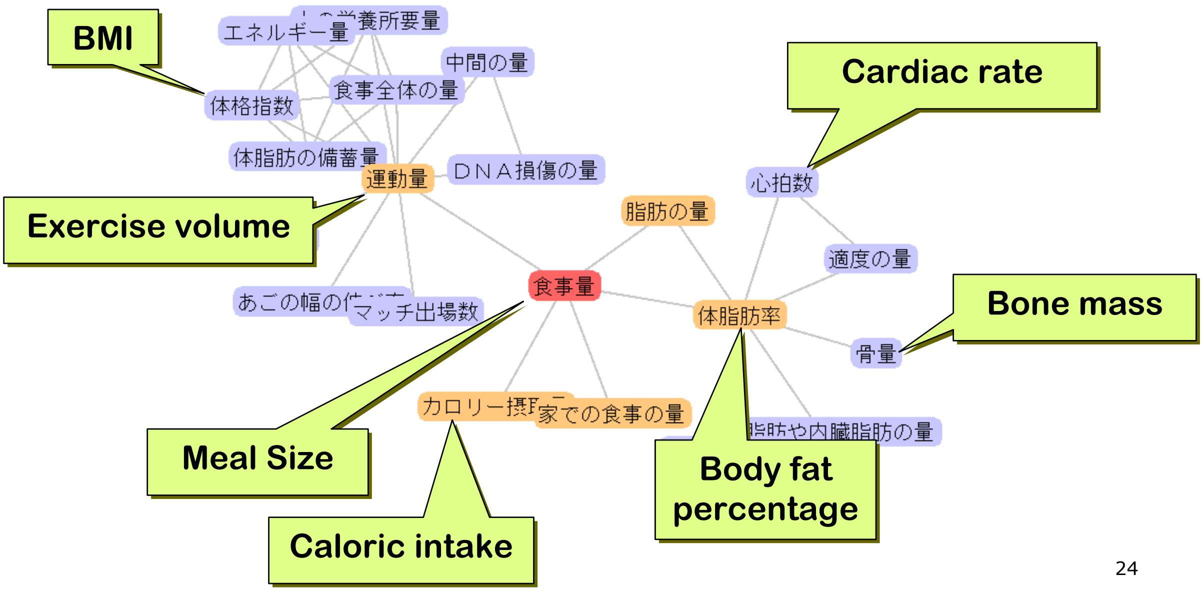
Consumer spending

Labor force population

Examples: GDP Growth



Example: Meal Size



Conclusion

- Visualization for Statistical Term Network
  - Simplification of Network Structure
    - Degree Limit Parameter ω
  - Simplification of Semantic Structure
    - Integrate common base form and co-occurring terms
  - Experiment: Visualize a network consists of 8,600 statistical network
- Future Work
  - Exploit syntactic patterns about causal relation expression and make a direction on the statistical term network
  - Exploit Web corpus