A Study on Method of Trend Researches
: Focused on Structures of the Academic World in Science Education

Seok-Hyun Ga*, Eunjin Jang, Chan-Jong Kim, Seung-Urn Choe+
(Seoul National University)

Why is it important to comprehend research trends?

- Trends of research in academic publications represent research interests, approaches, as well as the shared knowledge of researchers (Lin et al., 2014).

- Periodic reports of research trends can help science education researchers understand the recent status in this research field (Lin et al., 2014).

- Revealing research trends contribute to develop a new research topics through identifying previously non-researched areas.

Current Studies of Trends Research in Science Education

Journal	Authors	Target Journal	N	Done by	Focus
IJSE	Tsai et al(2005) Taiwan	IJSE, JRST, SE (2008-2012)	869	Human	Frequency of - country - research type - research topic
IJSE	Lee et al(2009) Taiwan	IJSE, JRST, SE (2003-2007)	802	Human	Frequency of - country - research type - research topic - citing
IJSE	Lin et al(2014) Taiwan	IJSE, JRST, SE (1998-2002)	900	Human	Frequency of - country - research type - research topic - citing - author
JSET	Chang et al(2010) Taiwan	IJSE, JRST, RSE, SE	1401	Computer	Frequency of - research topic - country - author - citing

Limitation of Current Studies

- Only focused on frequency
  - All of researches were focused on “FREQUENCY” of
    - country, research type, research topic, the number of citing, author, and so on.
  - These researches could not provide correlations between numerous keywords.

- Too much labor done by human
  - Most of researches of trends were performed by human.
  - Due to human limit, the number of analyzed articles was small (under 1,000)

The Purpose of Research

- Developing the methodology to calculating the correlations of research keywords between journal articles.
- Applying the methodology to the field of science education, and examining its results

Research Design

Data Mining

- Transforming article information form unstructured to structured format (from HTML to Excel)

Data Analysis

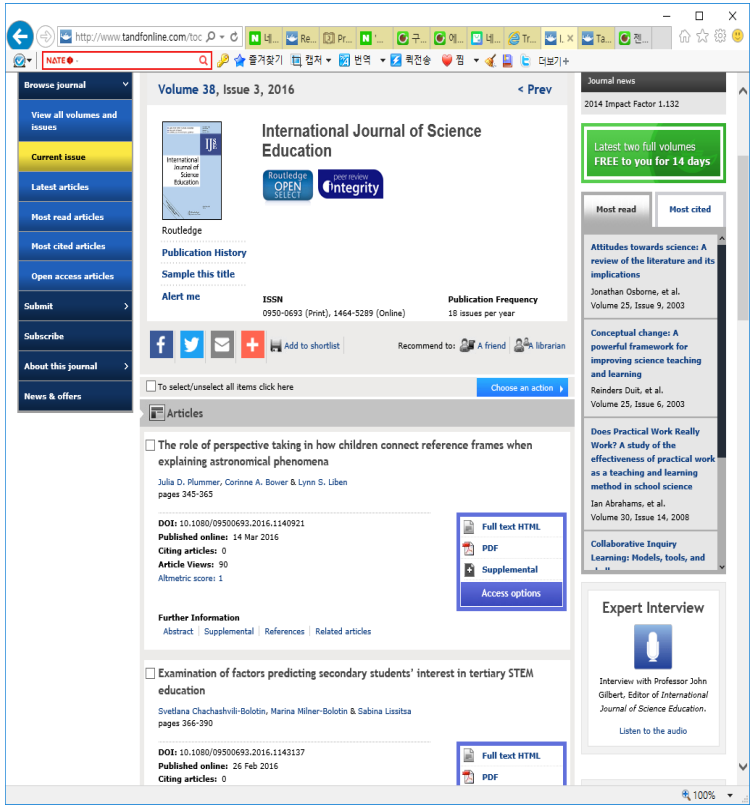
- Determine the unit of analysis
- Develop the automatic analysis program using R

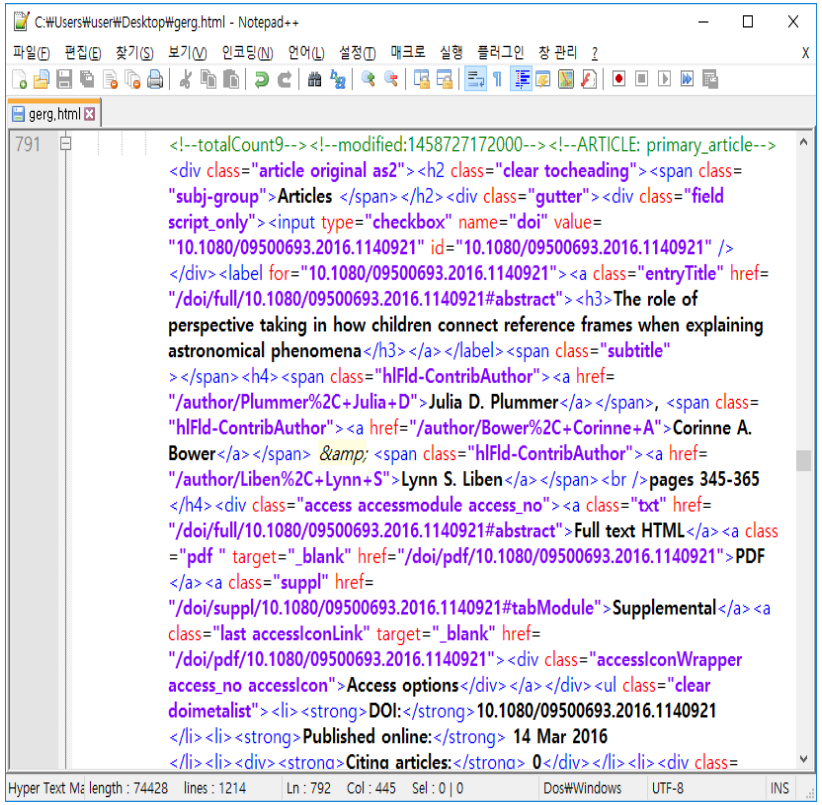
Data Visualization

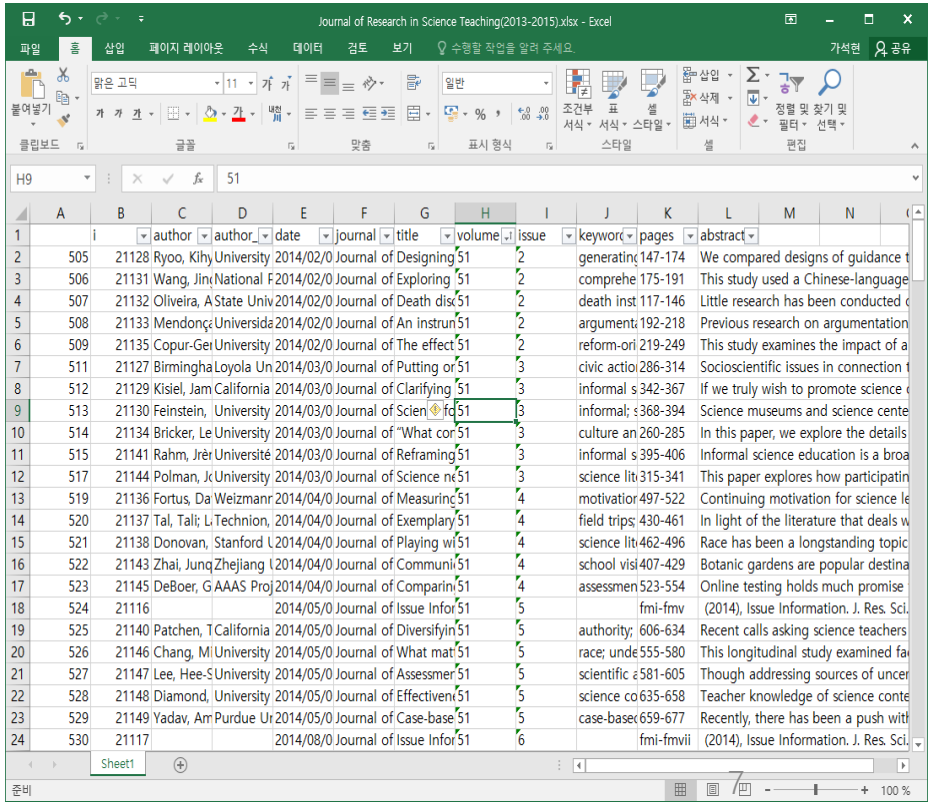
- Determine the format of graph which can visualize effectively
- Visualize information using network graph

Method: (1) Data Mining

- Transforming unstructured data into structured data through HTML Parsing

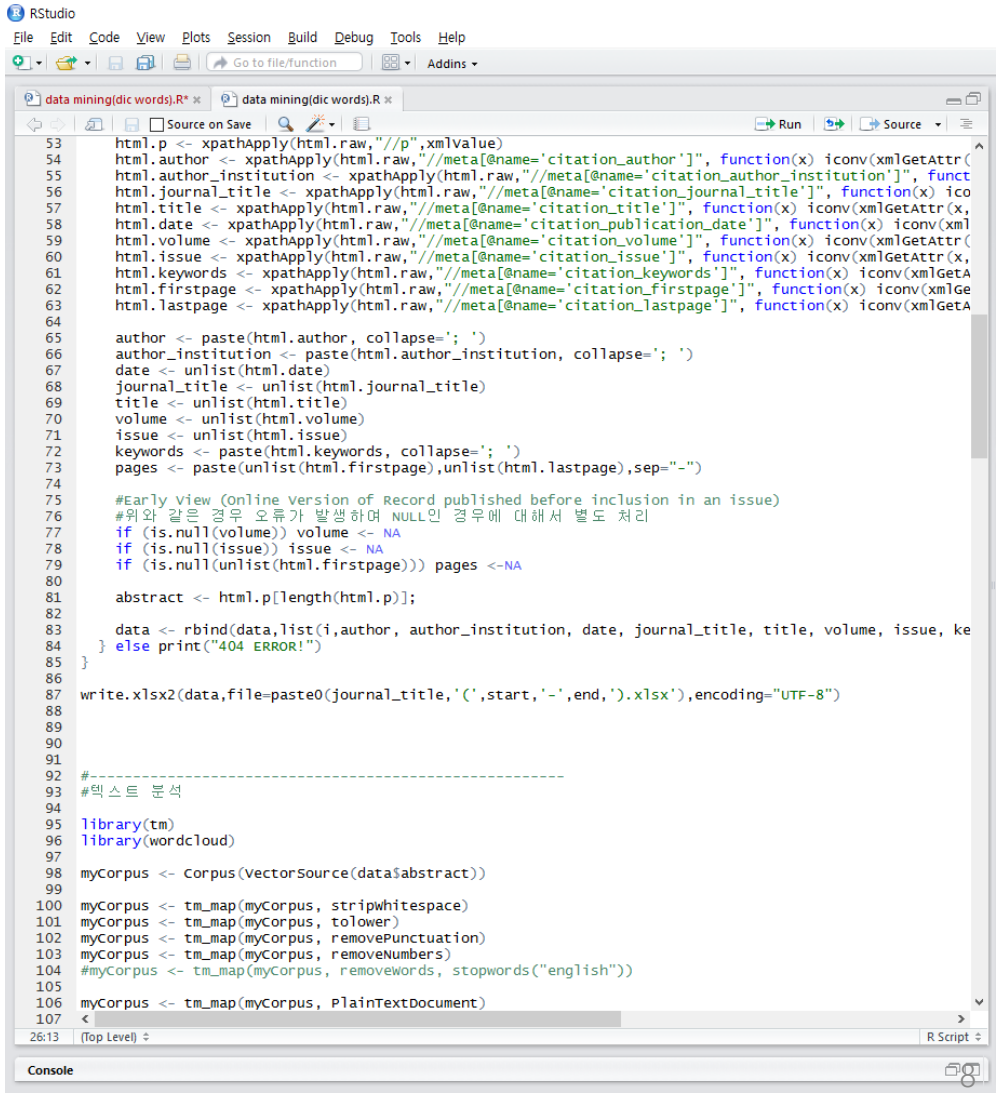






Method: (2) Data Analysis

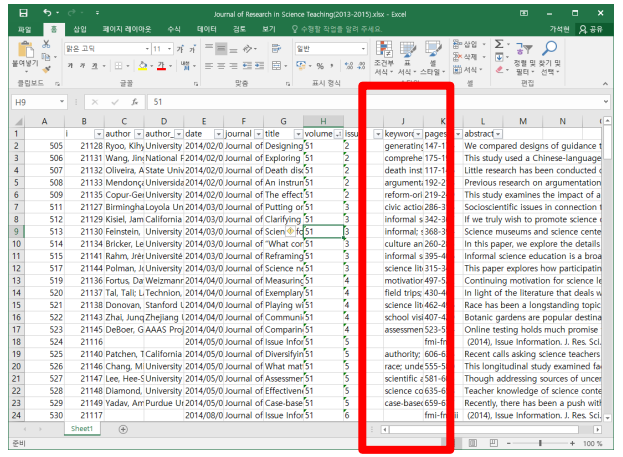
- Designing analysis algorithm
- Developing computer program using R
- Developed computer program process the data which is already mined.



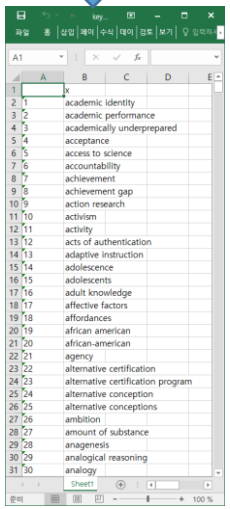

```
    html.p <- xpathApply(html.raw,"//p",xmlValue)
    html.author <- xpathApply(html.raw,"//meta[@name='citation_author']", function(x) iconv(xmlGetAttr(
    html.author_institution <- xpathApply(html.raw,"//meta[@name='citation_author_institution']", funct
    html.journal_title <- xpathApply(html.raw,"//meta[@name='citation_journal_title']", function(x) ico
    html.title <- xpathApply(html.raw,"//meta[@name='citation_title']", function(x) iconv(xmlGetAttr(x,
    html.date <- xpathApply(html.raw,"//meta[@name='citation_publication_date']", function(x) iconv(xml
    html.volume <- xpathApply(html.raw,"//meta[@name='citation_volume']", function(x) iconv(xmlGetAttr(
    html.issue <- xpathApply(html.raw,"//meta[@name='citation_issue']", function(x) iconv(xmlGetAttr(x,
    html.keywords <- xpathApply(html.raw,"//meta[@name='citation_keywords']", function(x) iconv(xmlGetA
    html.firstpage <- xpathApply(html.raw,"//meta[@name='citation_firstpage']", function(x) iconv(xmlGe
    html.lastpage <- xpathApply(html.raw,"//meta[@name='citation_lastpage']", function(x) iconv(xmlGetA

    author <- paste(html.author, collapse='; ')
    author_institution <- paste(html.author_institution, collapse='; ')
    date <- unlist(html.date)
    journal_title <- unlist(html.journal_title)
    title <- unlist(html.title)
    volume <- unlist(html.volume)
    issue <- unlist(html.issue)
    keywords <- paste(html.keywords, collapse='; ')
    pages <- paste(unlist(html.firstpage),unlist(html.lastpage),sep="-")

    #Early View (Online Version of Record published before inclusion in an issue)
    #위와 같은 경우 오류가 발생하여 NULL인 경우에 대해서 별도 처리
    if (is.null(volume)) volume <- NA
    if (is.null(issue)) issue <- NA
    if (is.null(unlist(html.firstpage))) pages <-NA

    abstract <- html.p[length(html.p)];

    data <- rbind(data,list(i,author, author_institution, date, journal_title, title, volume, issue, ke
  } else print("404 ERROR!")
}

write.xlsx2(data,file=paste0(journal_title,'(',start,'-',end,').xlsx'),encoding="UTF-8")



#------------------------------------------------------------
#텍스트 분석

library(tm)
library(wordcloud)

myCorpus <- Corpus(VectorSource(data$abstract))

myCorpus <- tm_map(myCorpus, stripwhitespace)
myCorpus <- tm_map(myCorpus, tolower)
myCorpus <- tm_map(myCorpus, removePunctuation)
myCorpus <- tm_map(myCorpus, removeNumbers)
#myCorpus <- tm_map(myCorpus, removeWords, stopwords("english"))

myCorpus <- tm_map(myCorpus, PlainTextDocument)
```

Method: (2) Data Analysis

- Assumed that keywords in research article has a meaning of research topic.
- Collecting keywords from mined data to use keywords as the unit of analysis
- Eliminating the duplicated keywords from the list.



Eliminating Duplicated keywords



Method: (2) Data Analysis

- Calculating relations between keywords

article 1	article 2	article 3
• action research • authentic inquiry • curriculum • teacher cognition	• constructivism • citizen science • action research • authentic inquiry	• constructivism • feminism • teacher education • authentic inquiry

- If there are two words(action research, authentic inquiry) in both article 1 and article 2, these keywords is added 2-point correlation.
- If there aren't two words(curriculum, constructivism) in one article at the same time, these keywords isn't added correlation point.

Method: (2) Data Analysis

- Calculating relations between keywords

article 1	article 2	article 3
• action research • authentic inquiry • curriculum • teacher cognition	• constructivism • citizen science • action research • authentic inquiry	• constructivism • feminism • teacher education • authentic inquiry

- If there are two words(action research, authentic inquiry) in both article 1 and article 2, these keywords is added 2-point correlation.
- If there aren't two words(curriculum, constructivism) in one article at the same time, these keywords isn't added correlation point.

Method: (2) Data Analysis

- Calculating relations between keywords

article 1	article 2	article 3
• action research • authentic inquiry • curriculum • teacher cognition	• constructivism • citizen science • action research • authentic inquiry	• constructivism • feminism • teacher education • authentic inquiry

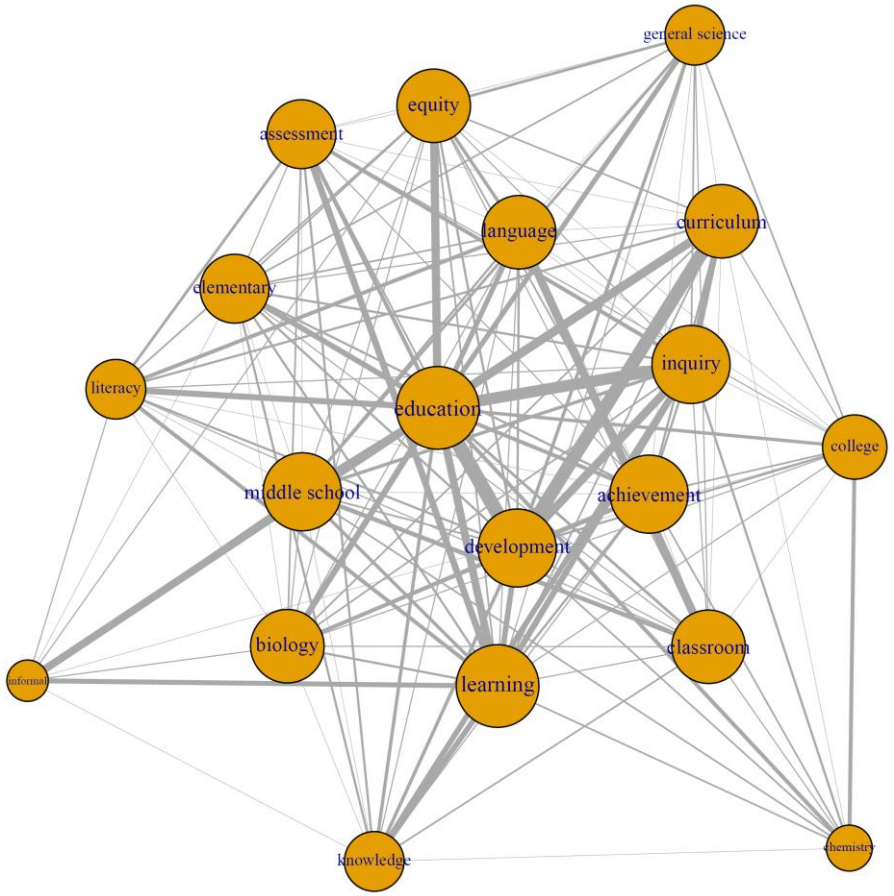
- If there are two words(action research, authentic inquiry) in both article 1 and article 2, these keywords is added 2-point correlation.
- If there aren't two words(curriculum, constructivism) in one article at the same time, these keywords isn't added correlation point.

Method: (3) Data Visualization

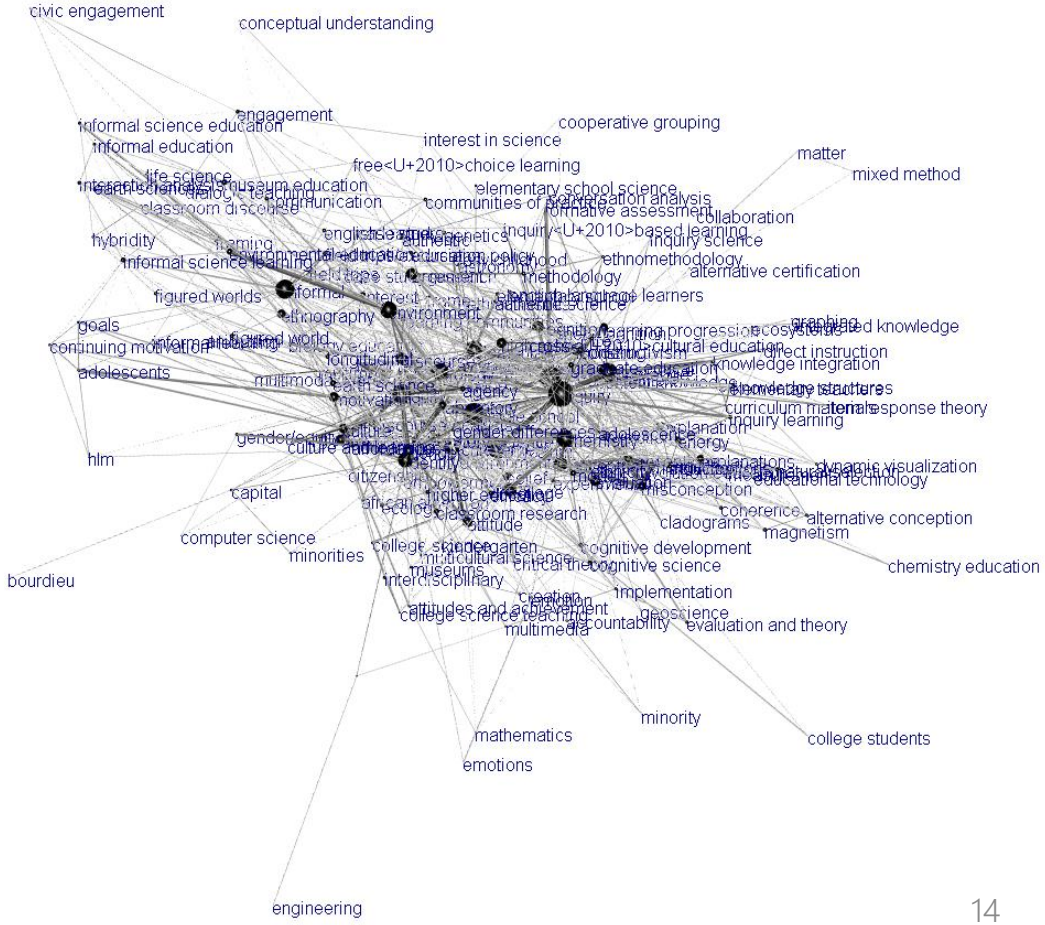
- **Visualizing data through network graph** to show correlation between keywords effectively.
  - A **node** is a keyword which means "topic" or "terms", and the **size of node** is frequency of appearance.
  - A **edge** is correlation between two nodes, and the **width of edge** is degree of correlation.

Method: (3) Data Visualization

- 2D Visualization



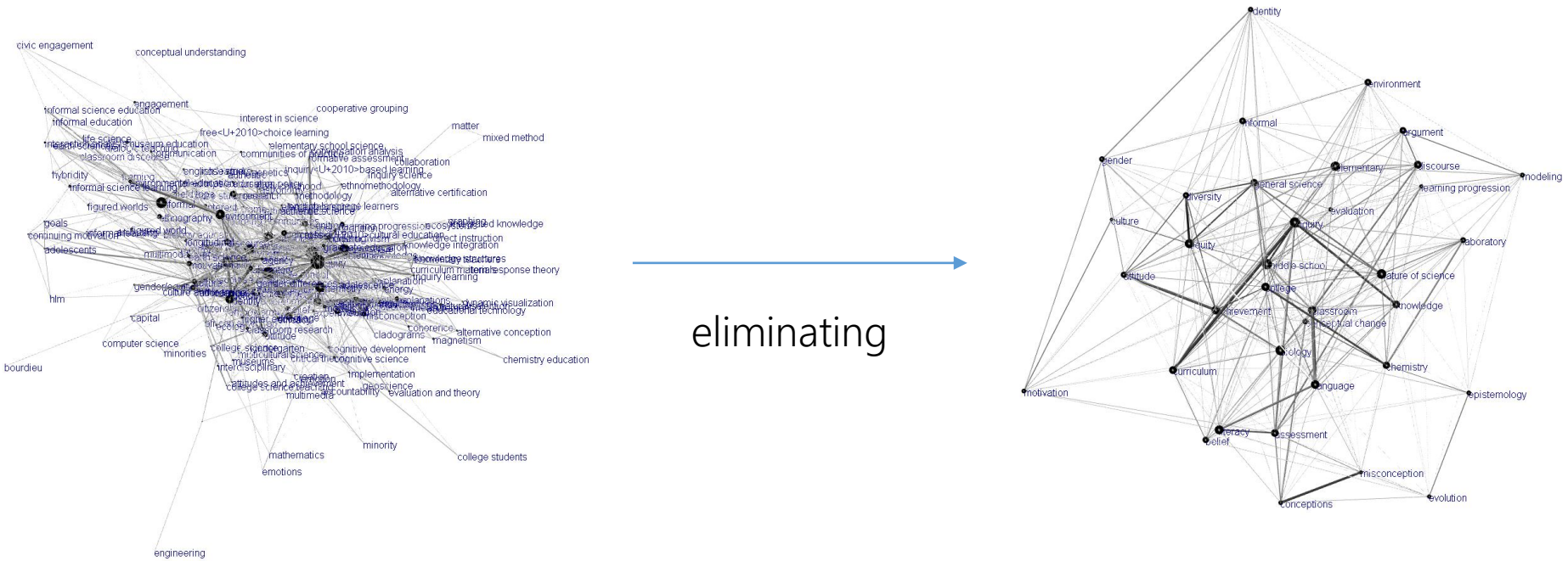
- 3D Visualization



Method: (3) Data Visualization

- Eliminating general keywords(node)

  *General Keyword: the keyword which is commonly used in science education, but doesn't have meaningful information about research topic



Method: (3) Data Visualization

- Criteria of eliminating node
  - Keywords that mostly used or nearly used are not meaningful, so these keywords are eliminated automatically by computer program. (Eliminate the keyword which appear with a probability of 0.02)

$$0.02 < P(K_i)$$

  - Two keywords that has high correlation between them are integrated into one node. (Integrate the keywords which have a 0.90 correlation)

$$P(K_j|K_i) < 0.90$$

  - Limitation: The node is mechanically eliminated or integrated, so it caused inaccuracy.

$$* \ P(K_i): Apperence\ Probability\ of\ i-th\ keyword$$

Method: (3) Data Visualization

- Done by Researcher
  - Most frequently used keywords – i.e. education, science, development – can not provide meaningful information about research topics, so these keyword are selectively eliminated.

    education, learning, development, curriculum development

  - Similar keywords are integrated into one node.

    curriculum, curriculum development / learning progression, learning progressions / Informal, informal science / argument, argumentation / attitude, attitudes / middle school, middle school science

  - Limitation: It depends on subjectivity of researcher.

Results

- Analysis result of IJSE and JRST articles from 2008 to 2015

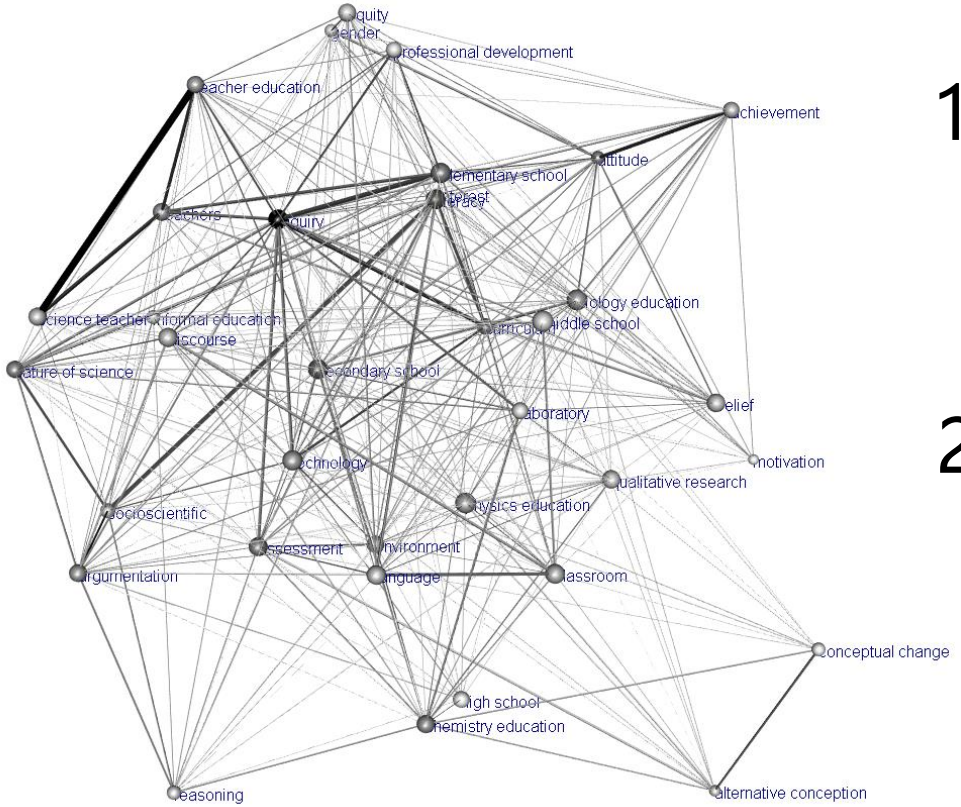


Figure 1

1) **The most prominent keywords**
inquiry, secondary school, middle school, language, classroom

2) **High correlation keywords**
language-classroom
Inquiry-curriculum
Diversity-equity
Inquiry-technology

Results

- Analysis of 4 years
  - from 2008 to 2011

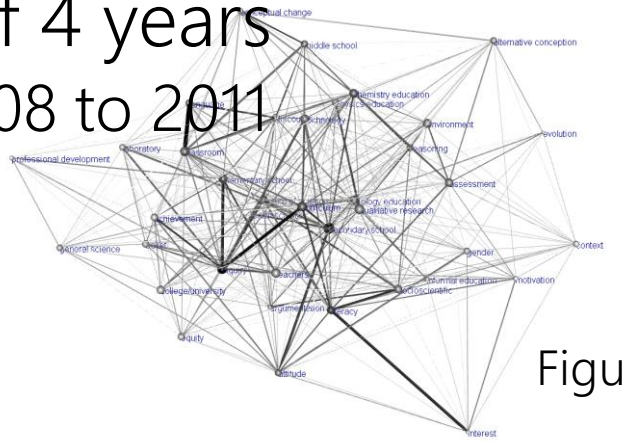


Figure 2

  - from 2012 to 2015

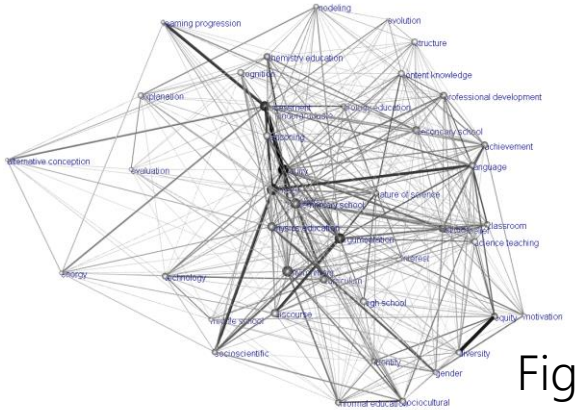


Figure 3

- There is no significant difference between two network graph.
- It may caused a narrow time gap between them.
- If time gap had been broad, there might have been significant differences.

Results

- Analysis result of IJSE or JRST articles from 2008 to 2015

IJSE

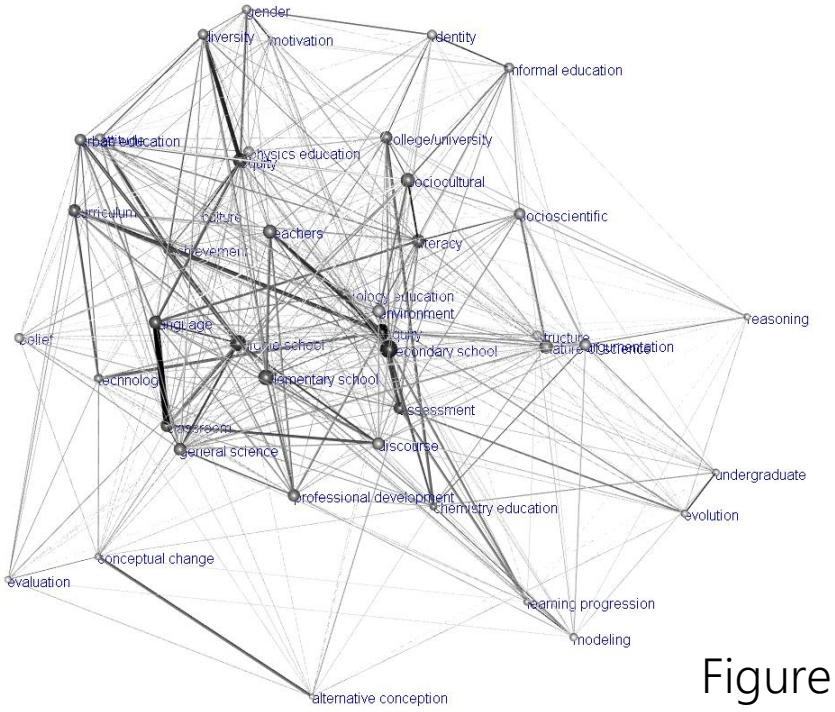


Figure 4

JRST

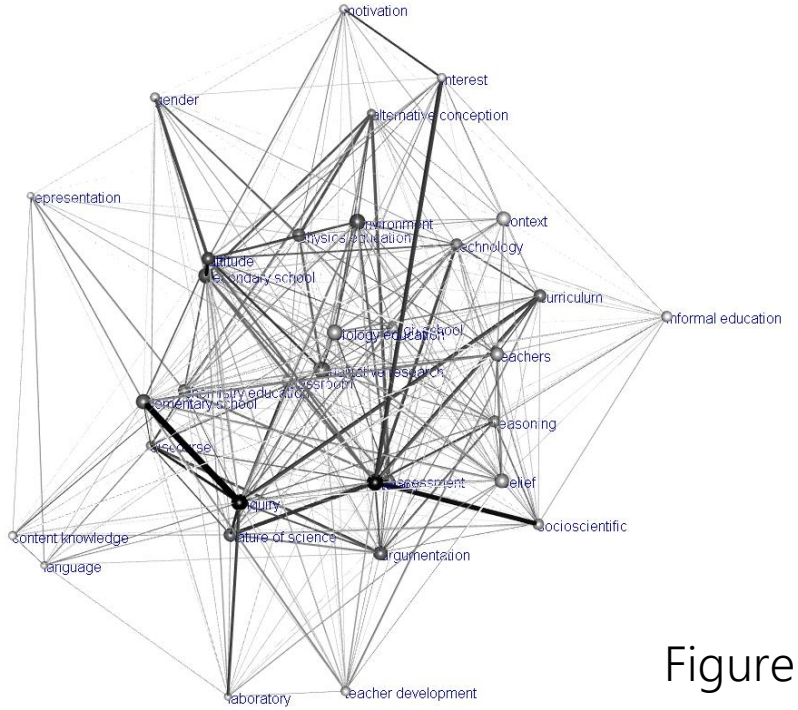


Figure 5

- There are no significant difference between results of JRST and IJSE.

Conclusion

- Analysis of research trends can be processed by computer.
  - So, huge size of articles can be analyzed automatically.

- Using this methodology in science education,

  - **Structure of research field** in science education can be visualized.

  - The **most prominent keywords** and **high correlation keywords** can be revealed.

Limitations of Research

- Some of journals or past journals(before 2007) doesn't have keywords in their articles, so these journals cannot apply this keyword-based methodology.
  - In order to overcome these limitation, new methodology developed that **"authentic keywords" are extracted from full-text of articles**.
  - When it is done, there will be **significant difference between analysis graphs of 4 years**, and **forecasting research trends** can be possible.

Limitations of Research

- This research help researchers who want to know about research trends.

- But it is wondered how this research might contribute to teaching and learning.

Thank you!

If you want to get computer programming source codes of this research, please contact me by email (shga@snu.ac.kr)