中山大學
SUN YAT-SEN UNIVERSITY

Construction of Classification of Aggregated Units and its Visualization

Speaker:Ma, Cuichang
School of Information Management
Sun Yat-sen University
Guangzhou, China

中山大學
SUN YAT-SEN UNIVERSITY

OUTLINE

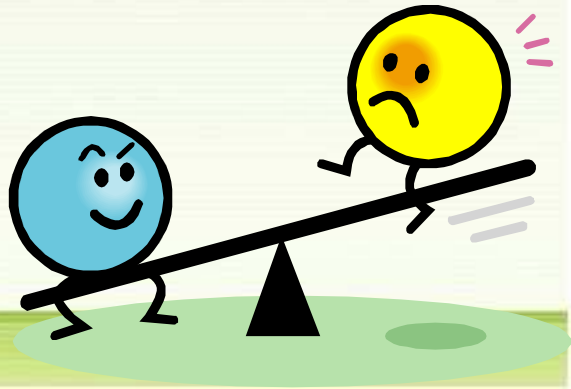
- Introduction
- Theoretical Framework
- Methodology
- Classification and Visulization
- Discussion and Conclusions

Introduction

- Information aggregation has become a hot spot world wide:
  - *ACM SIGIR 2008*
  - *National Science Fund or National Social Science Fund in China*
- Information units of digital resources has been utilized successfully:
  - *NSF/DARPA/NASA Digital Libraries Initiative (DLI) project at the University of Illinois*
  - *China National Knowledge Infrastructure*
  - *Studies on functional units from genre theories and knowledge units*

Introduction

- What about utilize information units on the web?
- Research Questions:
- RQ1:
  - *Do OA (Open Acess) papers of empirical study, microblog or weblog posts have the same kind of aggregation units*
- RQ2:
  - *How to integrate and visualize the ontology of domain and aggregation units?*



Theoretical Framework

- Genre type is chosen as the first level of classification of Internet resources:
  - *Studies on organization of Internet resources pay more attention to the automatic classification of information by genre type.*
  - *Studies of user behaviors indicate that genres of information have significant influence on users' perceived information usefulness.*

Theoretical Framework

- Taxonomy of functional units of journal articles is chosen as the base to develop a classification of aggregation units for resources on the web:
  - *classification of genre or even functional units of them have been studied widely in the field of linguistic and LIS*
  - *OA papers are always important resources of the academic information on the web in most disciplines*

Table 1. Granularity of Aggregation

Theoretical Framework

- For granularity of information, with the help of knowledge organization systems, have influence on the efficiency of information retrieval and utilization
- We define the aggregation level of information resources from two aspects, as shown in Table 1:
- *granularity of domain concepts in knowledge organization system*
- *granularity of information as control units.*

	Granularity of Concept		Granularity of Control Unit
K1	Main concept	C1	Papers or Posts as a whole
K2	Sub-concept	C2	Chapters or Sections
Kn	…….	C3	Sentences or Paragraphs

Methodology

Sources of Data

- *Genre types: OA papers, blogs and microblogs*
- *Keyword:"Citation analysis"*
- *Retrieval time was June, 2014*

Methodology

Sources of Data

- *We retrieved 13 OA papers from the website of Journal of Library Science in China and 11 of which can be downloaded.*
- *We retrieved 123 blogs from Sciencenet.cn, and download 15 blogs.*
- *We retrieved 870 microblogs from Xinlang weibo(not exactly match) and selected 15 relevant microblogs.*

Methodology

Sources of Data

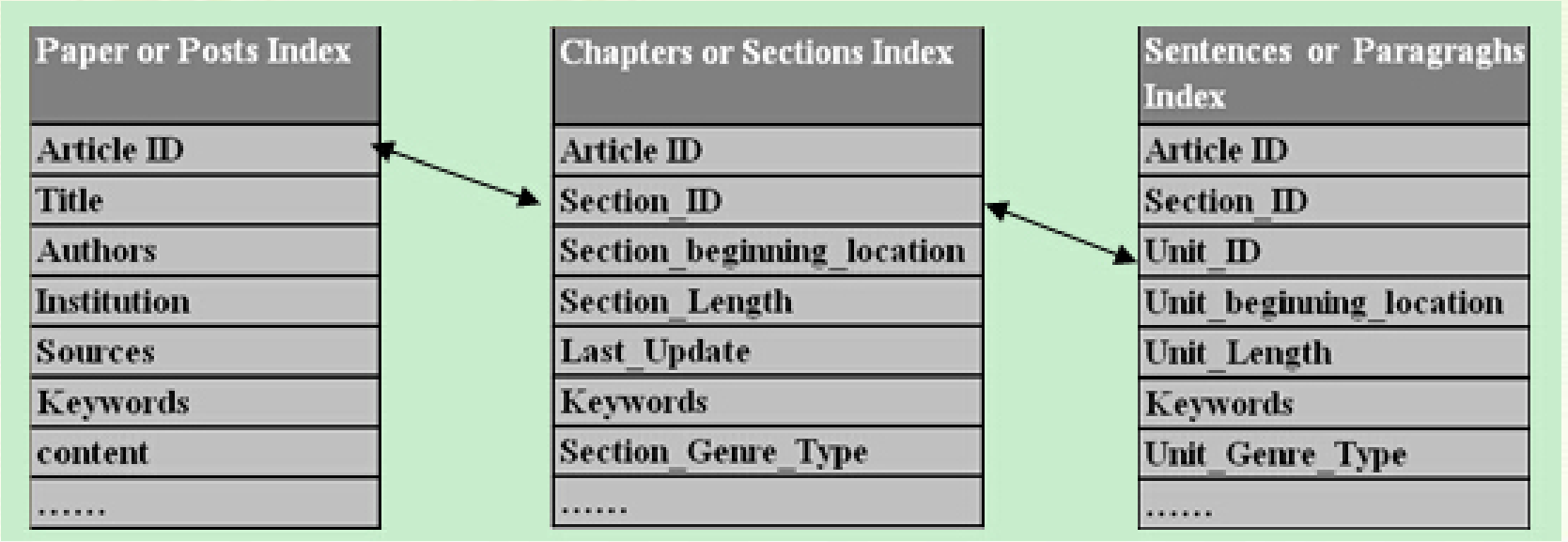
	OA papers	weblog post	microblog
Total	11	15	15
Including	6 empirical studies 1 Literature Review 3 theoretical studies 1 report	5 commentaries 4 introduction with links 2 tutorials 1 news	3 introduction 7 introduction with links 1 suggestion 3 commentary 1 news

- *We chose 1 empirical study, 1 commentary weblog post and 4 introductory microblog post as a small data set to develop the classification of aggregation.*

Methodology

Database Structure

- *Use Excel 2007 as database to store and index data*
- *The relationship among three tables above makes the jump among different level of units become possible as we can see as follows.*



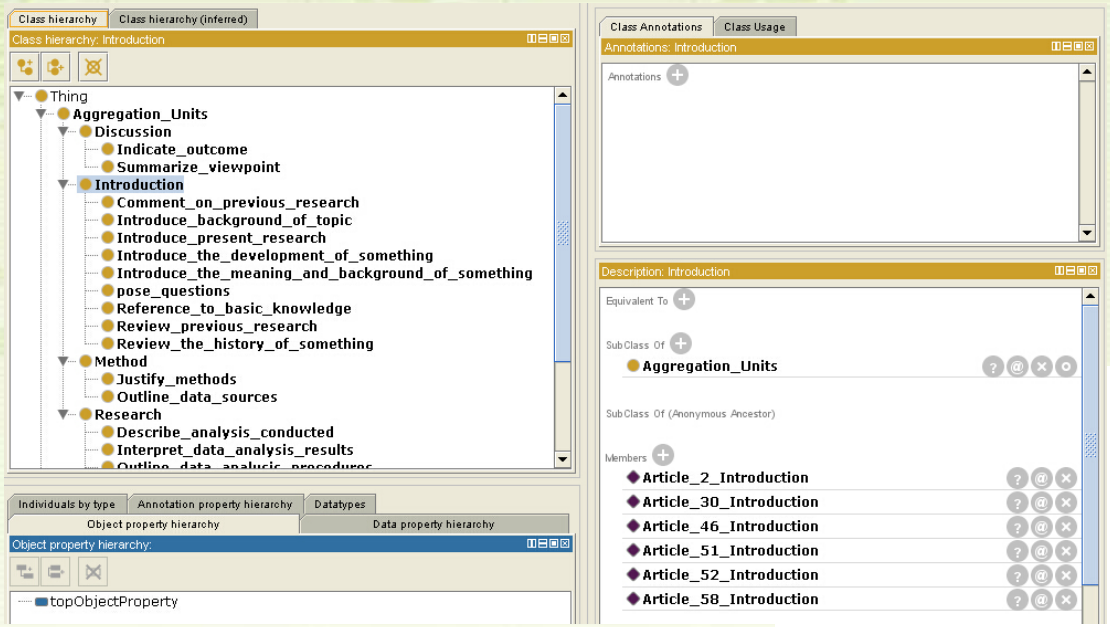
Classification and its Visualization

- **Construction of ontology for a domain and its visualization**
  - *Base on the knowledge structure of "Citation analysis" in Baidu Encyclopedia (http://baike.baidu.com/), we construct the knowledge organization system of "Citation analysis" as follows.*

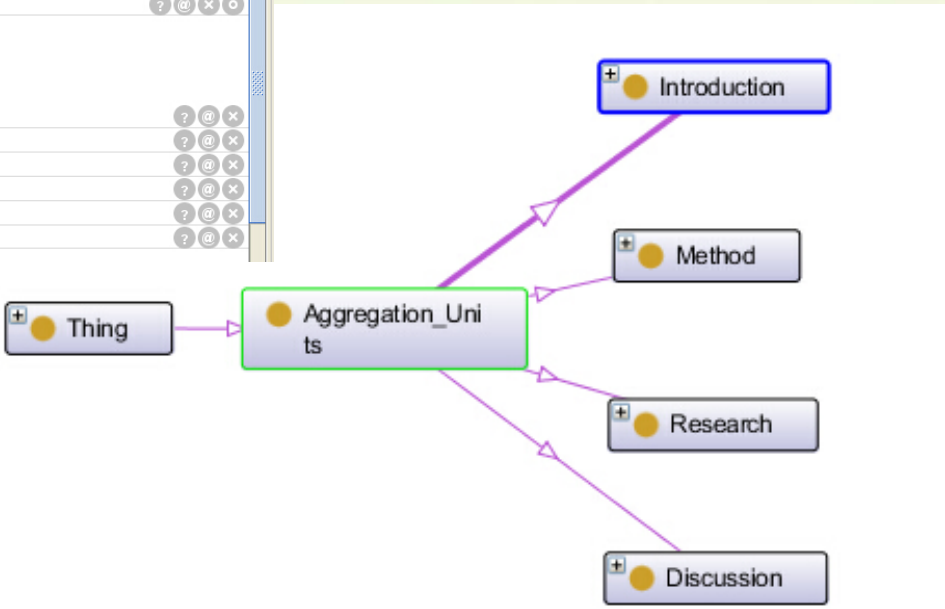
Classification and its Visualization

- **Construction of ontology for a domain and its visualization**
- *Through word segmentation and word frequency analysis with ROST CM(a Chinese natural language processing software), we added meaningful new words to the knowledge organization system of "Citation analysis" so as to improve its coverage.*
- *With the knowledge organization system, we built ontology of "Citation analysis" with Protégé.*

Classification and its Visualization



an example of its entities and classes



a snapshot of the graph of ontology

Classification and its Visualization

- **Development of classification of aggregation units.**
  - *Two doctoral students in discipline of LIS were asked to examine the sections and units according to Zhang's model.*
  - *With the taxonomy of Zhang's functional units, they examined the empirical study in dataset first and then extended it to weblog and microblog.*

The classification was justify after the two students agree on the kinds and scopes of components.

Classification and its Visualization

- **Development of classification of aggregation units.**
  - *However, even though the IMRD sections was fit for our resources, the units within them were not so much the same for they were highly discipline and genre-specialized.*
  - *Therefore the set of aggregation units within IMRD were mostly developed according to the content of the resource.*

Classification containing 4 sections and 18 units were developed as follow

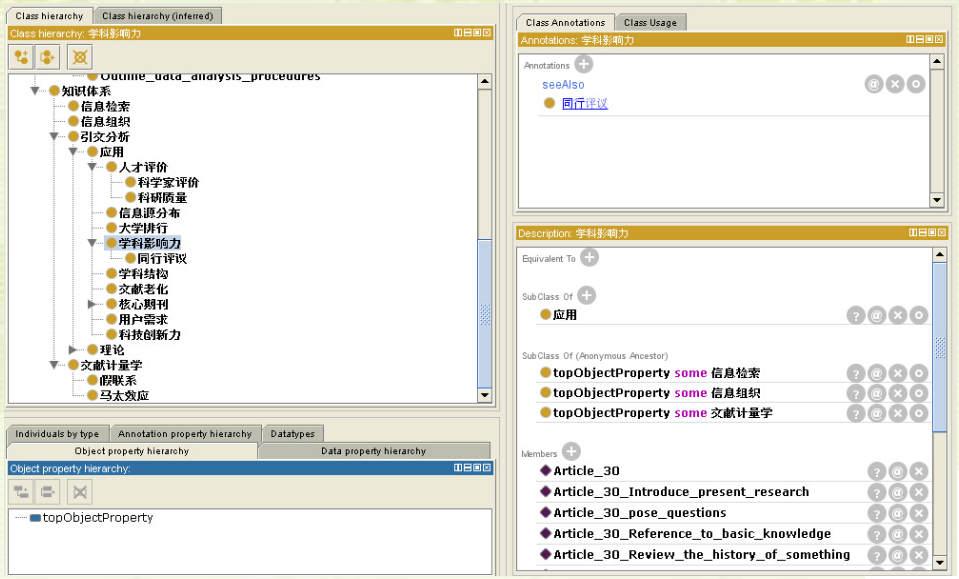
Classification and its Visualization

C2	C3
Introduction	Introduce background of topic
	Review previous research
	Comment on previous research
	Introduce present research
	Pose questions
	Introduce the meaning and background of something
	Reference to basic knowledge
	Review the history of something
	Introduce the development of something
Method	Outline data sources
	Justify methods
Research	Outline data analysis procedures
	Describe analysis conducted
	Interpret data analysis results
Discussion	Indicate outcome
	Summarize viewpoint

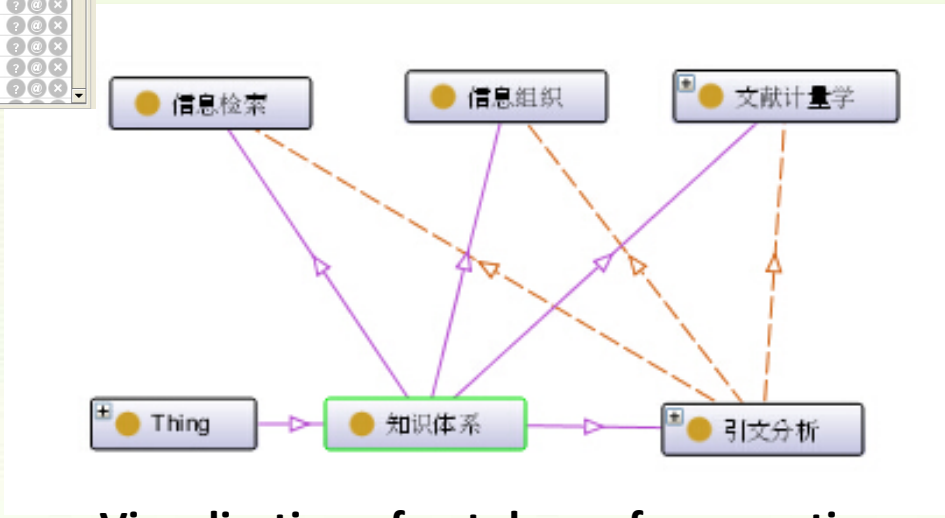
Classification and its Visualization

- **Construction of ontology of aggregation units and its visualization**
- With the classification of aggregation units the ontology of aggregation units was built using protégé. The entities and classes are as follows:

Classification and its Visualization



Entities and Classes of ontology of aggregation units



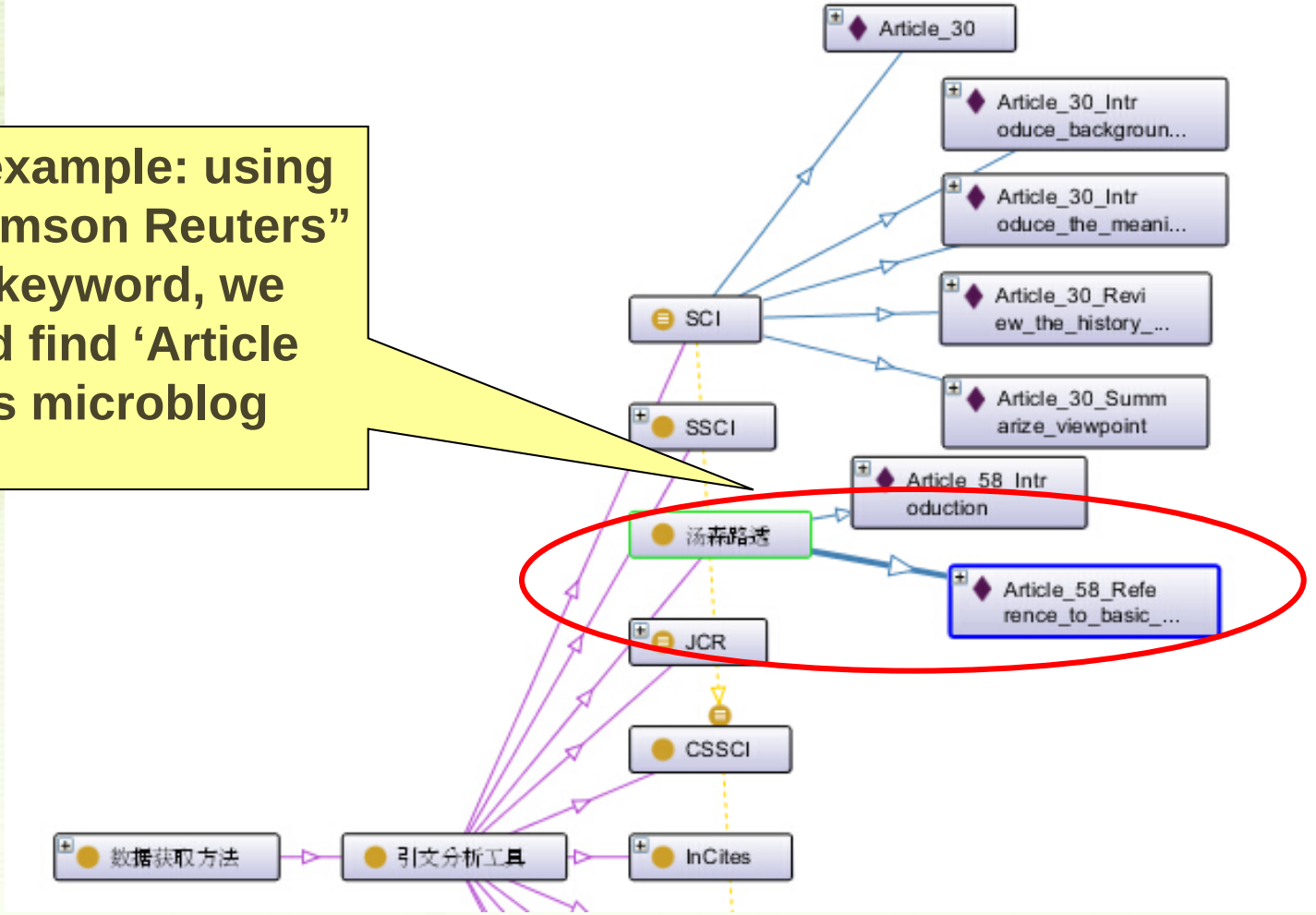
Visualization of ontology of aggregation units

Classification and its Visualization

- **Fusion of ontology of aggregation units , citation analysis and its application**
  - *Resources in the database were annotated with both knowledge organization system and classification of aggregation units and were added to the ontology as individuals.*
  - *So two ontologies were fused which could facilitate the expansion of knowledge retrieval and location of knowledge units.*

Classification and its Visualization

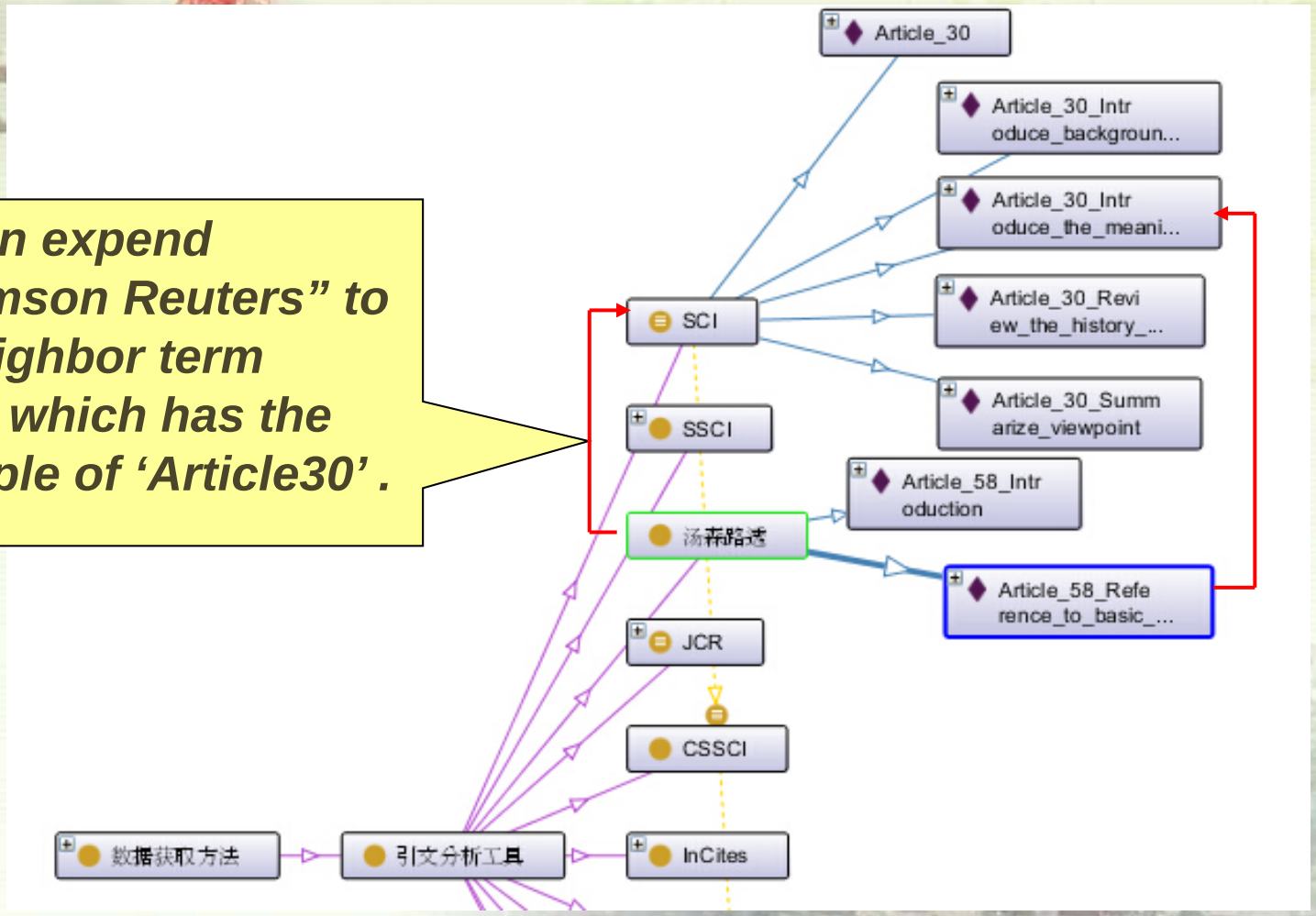
For example: using “Thomson Reuters” as a keyword, we could find ‘Article 58’ as microblog post.



Visualization of the expansion and location of resources

Classification and its Visualization

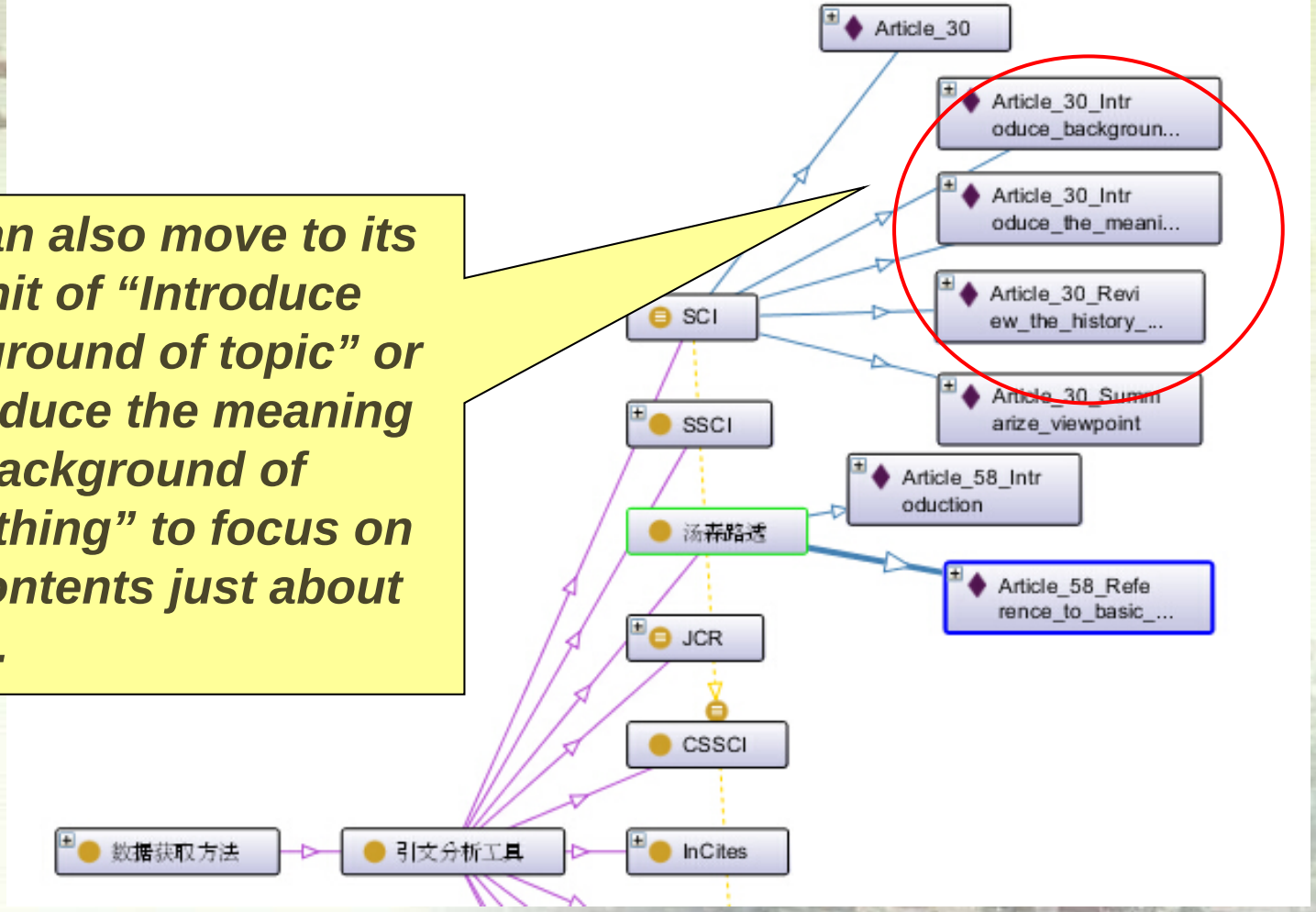
we can expend "Thomson Reuters" to its neighbor term "SCI" which has the example of 'Article30' .



Visualization of the expansion and location of resources

Classification and its Visualization

We can also move to its subunit of "Introduce background of topic" or "Introduce the meaning and background of something" to focus on the contents just about "SCI".



Visualization of the expansion and location of resources

Discussion and Conclusions

Aggregated units

4 section and 16 units typically exists within web resources of empirical study, weblog posts and microblog posts in LIS

Visualization

ontology of “citation analysis” and aggregation units can be fused and visualization would be realized with Protégé

Application

This study can be used to inform aggregation system design with the aim of enhancing use of various kinds of academic internet resources

Thank you !

Question?