Postprint: Research on a Linked Data-Based Cluster Semantic Discovery Model

Authors: Cui Jiawang, Li Chunwang

Date: 2017-11-08T00:00:00+00:00

Abstract

Purpose: To investigate models and technical approaches for revealing semantic relationships among subject terms within clusters based on linked data. Method: Utilize Google Scholar, Springer, CNKI, and other databases to retrieve literature related to the research topic, investigate and analyze current research on cluster analysis and semantic relationship revelation, construct a cluster semantic relationship revelation model based on linked data, and verify the model' s effectiveness through experiments. Results: Experimental results indicate that utilizing linked data can effectively reveal semantic relationships among subject terms, compensating for the semantic limitations of traditional co-word clustering analysis. Limitations: Constrained by experimental data, the revealed semantic relationships are currently limited to hierarchical relationships, class-instance relationships, associative relationships, and similar types; the impact of linked data quality issues on semantic revelation results has not been considered. Conclusion: The proposed cluster semantic relationship revelation model based on linked data can effectively reveal semantic relationships among subject terms, providing a novel approach for understanding and analyzing co-word clustering results.

Full Text

Research on a Semantic Relation Revealing Model for Clusters Based on Linked Data

Cui Jiawang[1],[2], **Li Chunwang**[1]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract

[**Objective**] This study investigates models and technical approaches for revealing semantic relationships among subject terms within clusters based on linked data. [**Methods**] We retrieved relevant literature through Google Scholar, Springer, CNKI, and other databases, analyzed current research on cluster analysis and semantic relation revelation, constructed a semantic relation revealing model for clusters based on linked data graph structures, and validated the model' s effectiveness through experiments. [**Results**] Experimental results demonstrate that linked data can effectively reveal semantic relationships among subject terms, addressing the semantic limitations of traditional co-word clustering analysis. [**Limitations**] Due to experimental data constraints, the revealed semantic relationships are currently limited to hierarchical relationships, class-instance relationships, and associative relationships. The impact of linked data quality issues on semantic revelation results was not considered. [**Conclusions**] The proposed semantic relation revealing model for clusters based on linked data can effectively reveal semantic relationships among subject terms, providing a novel approach for understanding and analyzing co-word clustering results.

Keywords: Linked Data; Co-word Clustering; Cluster; Semantic Relation Revealing Model

Introduction

Co-word clustering analysis groups subject terms—which inherently lack categories—into clusters representing different research subfields based on the principle that similar objects cluster together, thereby clearly and intuitively revealing the thematic structure and evolution of a discipline [1]. According to clustering principles, clusters aggregate terms with the shortest distances without considering logical relationships between terms, resulting in clusters that are difficult to interpret due to the absence of semantic relationships among subject terms. The publication and application of linked data provide new opportunities for advancing co-word clustering research. In particular, linked data pre-establishes a large number of authoritative and accurate attribute relationships, with each data object containing multiple attributes and characteristics, thereby providing effective support for achieving precise semantic relation revelation across disciplinary domains and data sources.

2 Related Research Overview

Cluster analysis can be divided into two levels: compactness analysis and semantic relation revelation. Compactness analysis primarily measures the tightness of clustering, with relevant research mainly includes cohesion and density metrics, as well as the integration of co-word clustering with auxiliary methods. Semantic relation revelation explores internal semantic relationships within clusters from a knowledge discovery perspective, with relevant research mainly includes:

(1) expert involvement, (2) co-word association analysis, (3) text mining, (4) ontology and thesaurus-based approaches, and (5) linked data-based approaches.

(1) Expert Involvement. Zhang et al. [2] proposed that domain experts should be involved in the co-word clustering process, where experts manually 梳理 semantic relationships within and between clusters to compensate for the reliance on mathematical statistics in co-word clustering.

(2) Co-word Association Analysis. Association rules describe knowledge patterns about items that co-occur within a transaction. Based on this principle, co-word association analysis reveals dependency relationships among subject terms through statistical methods. Zhang et al. [3] used association rule algorithms to analyze the collocation patterns of subject terms and subheadings for four anti-tumor drugs, extracting effective semantic relationship patterns related to these drugs. Zhang et al. [4] extracted knowledge about specific drug-disease relationships from bibliographic databases based on semantic association rules between subject terms/subheadings. Cimino et al. [5] studied collocation rules for subject terms and subheadings, automatically generating semantic relationships between medical concepts through simple pattern-matching rules.

(3) Text Mining. Text mining for semantic relationship discovery primarily involves scanning and automated processing of NLP to identify concepts and semantic relationships between them. Liu [6] explored semantic relationships between concepts in the military aircraft domain by combining text mining with ontology automatic construction methods.

(4) Ontology and Thesaurus-based Approaches. These approaches start from known semantic relationships between concepts. Zhang [7] inferred unknown semantic relationship types in the traditional Chinese medicine domain using association relationship distributions based on probability theory applications in the Traditional Chinese Medicine Language System. Wei [8] used online thesauri as a semantic foundation, introducing an association dictionary mechanism to identify semantic relationships between tags by recognizing relationships between tag sets and concepts in online thesauri.

(5) Linked Data-based Approaches. Research on semantic relationship discovery based on linked data is still in its exploratory stage. Tiddi et al. [9] proposed the Dedalo heuristic linked data traversal mining system, which is representative. Dedalo finds common paths among entities within a cluster through heuristic iterative retrieval of linked data, thereby forming shared semantic relationships among cluster entities. Taheriyan et al. [10] inferred semantic relationships of structured resources using linked data through semantic annotation and relationship building. Additionally, several linked data mining technologies provide important references for this study. In China, Li et al. [11] and Li et al. [12] summarized research on linked data-based data mining and proposed knowledge discovery models based on linked data. Gao et al. [13] proposed a knowledge discovery model based on linked data building upon the pyramid of knowledge discovery processes. Song [14] proposed a tacit knowledge discovery

model based on knowledge maps in linked data environments. Liu [15] proposed a knowledge discovery process model based on linked data. Compared to domestic research, international studies are more abundant. Narasimha et al. [16] proposed the LiDDM linked data mining system, and Paulheim et al. [17] proposed the FeGeLOD feature extractor, both of which transform linked data into formats suitable for traditional data mining algorithms through format conversion or feature extraction. Ramezani et al. [18] proposed SWApriori and Personeni et al. [19] proposed ILP learning methods, which adapt traditional data mining algorithms for application to RDF-formatted data for linked data mining. Jiang et al. [20] proposed frequent subgraph mining methods, and Li et al. [21] proposed deep learning methods that mine structural information such as attribute chains and nodes in linked data.

Each method has certain limitations. Expert involvement is costly and difficult to generalize. Association analysis-based semantic relationship discovery can only identify certain specific types of semantic relationships. Text mining methods suffer from text corpora typically lacking sufficient structured information. Ontologies and thesauri have rigorous structures but insufficient coverage and semantic connectivity, with many being limited in size and scale, making it difficult to cover sufficiently rich concepts and relationships. As an important resource for semantic mining, linked data demonstrates dual advantages in scale and structure. Therefore, although revealing cluster semantic relationships based on linked data represents a new attempt, with the rapid development of LOD data resources and related technologies, this new semantic relationship revelation method may become a future research trend.

3 Semantic Relation Revealing Model for Clusters Based on Linked Data

Subject terms within a cluster correspond to nodes in linked data. Based on the network structure characteristics of linked data, the maximum distance between subject term nodes is 3, and possible association relationships are shown in Figure 1 [Figure 1: see original paper].

In Figure 1, brown ellipses (Ek) represent nodes corresponding to subject terms within a cluster (i.e., subject term nodes), while white boxes representing extra-cluster nodes (E) refer to new nodes discovered through linked data mining. Lines/curves between nodes represent attribute relationships (R). This study limits the research scope to association relationships where the distance between subject terms does not exceed 3 for the following reasons: (1) A maximum distance of 3 ensures sufficient association relationships. LOD is a typical small-world network where, regardless of network size, the maximum search path steps remain relatively stable. Research shows the average shortest path length between nodes in LOD is 2.4 [22]; (2) According to comprehensive path importance evaluation methods, association relationships at greater distances have lower importance and lack semantic revelation value; (3) The search space for linked data graph mining increases exponentially, with longer paths leading

to greater time overhead.

3.1 Model Structure To accurately describe associations within clusters, this paper proposes the following definitions:

(1) **Linked Data Graph:** A directed graph composed of RDF data, where nodes are subjects or objects annotated with URIs, and edges are sets of attributes annotated with URIs.

(2) **Association Path:** This paper defines the collection of attributes R and nodes E traversed from subject term node E_1 to subject term node E_2 as an association path. An association path from E_1 through node E_1 to E_2 can be represented as: $E_1 \rightarrow E_1 \rightarrow E_2$, where E_1 and E_2 represent subject term nodes, E_1 is an extra-cluster node discovered through linked data mining, and R_1 and R_2 represent inter-node attribute relationships. Path length refers to the number of attributes in the path; for example, $E \rightarrow E \rightarrow E$ represents an association path with length equal to 2.

(3) **Path and Attribute Direction:** The direction of an association path from subject term E_1 to subject term E_2 is represented as $E_1 \rightarrow E_2$. Attributes in the association path with the same direction as $E_1 \rightarrow E_2$ are forward attributes, while those with opposite direction are reverse attributes. For example, in $E_1 \leftarrow E \rightarrow E_2$, R_1 is a reverse attribute while R_2 is a forward attribute.

3.2 Association Path Classification Due to the complexity of association paths among multiple subject term nodes, this paper explores semantic relationships among all subject terms within a cluster by starting with pairwise relationships. Taking subject term nodes E_1 and E_2 in Figure 1 as examples, based on different path lengths and attribute directions, association paths between E_1 and E_2 can be classified into four types: direct relation, indirect relation, lowest common ancestor relation, and lowest common descendant relation, with different path types corresponding to different semantic relationships.

(1) **Direct Relation (DR):** Direct relation refers to association paths with length 1 between subject term nodes, where E_1 and E_2 have direct connections $E_1 \rightarrow E_2$ or $E_1 \leftarrow E_2$.

(2) **Indirect Relation (IR):** Indirect relation refers to association paths between subject terms with length ≥ 2 and without reverse attributes. As shown in Figure 2 [Figure 2: see original paper], for length-2 indirect relations between subject terms E_1 and E_2 , there are two types: $E_1 \rightarrow E \rightarrow E_2$ and $E_1 \rightarrow E \rightarrow E_2$.

(3) **Lowest Common Ancestor Relation (LCAR):** The definition of Lowest Common Ancestor (LCA) is: for two nodes u and v in a rooted tree T, LCA(T, u, v) represents a node x that is an ancestor of both u and v with the greatest possible depth. Similar structures exist in linked data,

where association paths with lowest common ancestor nodes are defined as LCAR. As shown in Figure 3 [Figure 3: see original paper], when path length is 2, there is one LCAR: E_{LCA} $\rightarrow$ E_1 and E_{LCA} $\rightarrow$ E_2 . When path length is 3, there are two LCARs: E_1 $\leftarrow$ E_{LCA} $\rightarrow$ E_2 and E_1 $\leftarrow$ E_{LCA} $\rightarrow$ E_2 .

(4) **Lowest Common Descendant Relation (LCDR):** In linked data, nodes have not only lowest common ancestors but also lowest common descendants. In linked data, if two subject term nodes converge to the same node through the shortest attribute chain, this node is called the lowest common descendant (LCD). Association paths with lowest common descendant nodes are defined as LCDR. As shown in Figure 4 [Figure 4: see original paper], when path length is 2, there is one LCDR: E_{LCD} $\leftarrow$ E_1 and E_{LCD} $\leftarrow$ E_2 . When path length is 3, there are two LCDRs: E_1 $\rightarrow$ E_{LCD} $\leftarrow$ E_2 and E_1 $\rightarrow$ E_{LCD} $\leftarrow$ E_2 .

3.3 Association Path Importance Evaluation The number of association paths between subject term nodes is vast and cannot be analyzed one by one, nor are all paths valuable for revelation. Therefore, evaluating the importance of association paths is crucial for revealing semantic relationships among cluster entities, specifically including: entity attribute importance evaluation, entity node importance evaluation, and comprehensive path importance evaluation.

(1) Entity Attribute Importance Evaluation

Common attribute importance evaluation methods for linked data mainly include: - **Information Theory-based:** Meymandpour et al. [23] proposed an information theory-based informativeness measure for linked data. Information theory uses uncertainty to measure information quantity, so the informativeness of a single association attribute P can be expressed as the negative logarithm of its occurrence probability, calculated as $-\log_2 Pr(P)$, where Pr(P) represents the probability of attribute P appearing in the entire dataset. - **Attribute Frequency-based:** Kasneci et al. [24] constructed an informativeness calculation method MING based on attribute frequency. MING provides a weight calculation method for relationship r from node i to node j: W = N(i, r, j) / N(*, r, j), where N(i, r, j) is the number of instances (i, r, j) and N(*, r, j) is the total number of instances reaching node j via relationship r. Balmin et al. [25] combined attribute frequency methods with manual weight assignment, pre-assigning weights to each attribute type based on experience before proportionally distributing weights according to relationship instance counts. Nie et al. [26] used similar approaches for attribute weight calculation in their PopRank object ranking algorithm. - **Associated Node-based:** Ng et al. [27] proposed calculating attribute weights based on associated nodes in the MultiRank algorithm, where attribute importance is computed as the product of importance scores of the two nodes (subject and object) connected by the attribute. - **TF-IDF-based:** Since attribute relationship distributions in linked datasets are often skewed with vastly different frequency magnitudes, this pa-

per proposes a TF-IDF-based attribute weight calculation method. In linked data, an attribute' s high frequency in a subgraph mined from the linked data graph (as shown in Figure 1) indicates strong discriminative ability within that subgraph (TF), while high frequency across the entire linked dataset indicates low discriminability (IDF). The TF-IDF-based attribute weight formula is: TF-IDF(R) = (tf / N) × log(N/n), where tf is the occurrence count of attribute R in the mined subgraph, idf is the inverse frequency of attribute R, N is the total number of associations in the dataset, and n is the total occurrence count of t across the entire linked dataset.

(2) Entity Node Importance Evaluation

Node evaluation methods for linked data mainly include: - **Information Theory Method:** In information theory, if an event consists of several independent sub-events, the information quantity is the sum of these sub-events' information quantities. In linked data, nodes consist of multiple association attributes, and a node' s self-informativeness is the sum of its association attributes' information quantities. - **Network Graph Analysis Method:** Nodes and attributes in linked data networks are analogous to web pages and hyperlinks on the Web, so traditional web analysis algorithms like PageRank and HITS can be adapted for LOD. In linked data, when focusing on a particular node, the node' s overall importance can be formed by comprehensively considering contributions from each adjacent node through their associations, thereby evaluating the core node' s influence. The weight calculation formula for a node j is: W = (1-α) + α × Σ B(j) W , where B(j) is the set of all nodes pointing to j, W is the weight of the association from node i to j, E is all nodes in the linked data network, and α is the damping coefficient (typically 0.85). All nodes start with equal initial importance values. Similarly, the extended HITS algorithm can also be used for node importance evaluation in linked data. Bamba et al. [29] adapted HITS to calculate nodes' authority and hub scores by pre-defining weights for each association relationship. - **Tensor Decomposition Method:** Tensors organize high-dimensional data, and tensor decomposition reduces high-dimensional data like tensors into products of smaller, simpler sub-matrices through methods such as Tucker and Parafac models, where the decomposed matrices describe important characteristics of the original matrix. The rich semantic relationships in linked data networks allow representation as a three-dimensional tensor T, where nodes, neighbors, and connecting relationships can also be represented as three-dimensional tensors. Focusing on a target node, its overall authority can be evaluated by comprehensively considering its authority across topics [30].

(3) Comprehensive Association Path Importance Evaluation

Based on the general assumption that shorter paths between nodes indicate stronger semantic relevance, we can use the extended Katz Centrality Measure from social network analysis to comprehensively calculate a path P' s importance. The fundamental principle [31] is: assuming the effectiveness of a path between two nodes is determined by a known constant probability α, the probability of a path consisting of k nodes is α . This paper introduces attribute

probability into the Katz centrality measure, where the comprehensive importance Pr(P) of an association path with length N can be calculated as: Pr(P) = Π $_{1}$ W(R). Since importance evaluation results for attributes and nodes differ in magnitude, normalization must be performed before calculating comprehensive path importance. Common normalization algorithms include linear function transformation, logarithmic function transformation, arctangent function transformation, and combinations of linear and logarithmic functions.

4 Implementation of Semantic Relation Revelation Based on Linked Data

This study implemented the semantic relation revelation model using Java and Eclipse, leveraging open-source tools including Jena and Virtuoso, and the DBpedia (2016-4) linked dataset.

4.1 Experimental Data Selection Research shows that DBpedia data is more comprehensive and rich compared to other datasets. DBpedia is an innovative knowledge base built on Wikipedia, Semantic Web, and linked data technologies, marking a significant achievement in the transition from document web to data web. The latest DBpedia (2016-4) contains over 9 billion RDF triples, 754 classes, and covers 127 languages. The English DBpedia alone describes over 6 million things (5.2 million resources classified under a unified ontology), including 1.5 million people, 810,000 places, 135,000 musical works, 106,000 films, 275,000 organizations, 301,000 biological species, and over 5,000 diseases, making it one of the largest cross-domain semantic knowledge bases. Given DBpedia' s rich semantic relationships and resource scale, this study adopted DBpedia as the foundation for cluster semantic discovery. To ensure experimental validity and objectivity, the subject terms "Cloning" and "PCR" from the cluster in the paper "Research Focus Analysis and Preliminary Outlook in Veterinary Molecular Biology Based on Co-word Analysis" [32] were selected for semantic revelation.

4.2 Semantic Revelation System Framework To implement semantic revelation based on linked data, we designed the framework shown in Figure 5 [Figure 5: see original paper], which consists of two parts: linked data graph mining and semantic revelation.

4.3 Linked Data Graph Mining Linked data graph mining refers to the process of discovering association paths among subject term nodes as shown in Figure 1 from linked datasets, comprising data preparation and linked data mining.

(1) Data Preparation. We obtained the DBpedia (2016-4) English dataset via Dump download and built a local SPARQL query endpoint based on Virtuoso 7.2.4. After dataset acquisition, we mapped subject term nodes through the keyword search service provided by the semantic browser LodLive,

discovering that the cluster subject terms "Cloning" and "PCR" correspond to the DBpedia nodes "http://dbpedia.org/resource/Cloning" and "http://dbpedia.org/resource/Polymerase_{{{chain}}_{{reaction}}} ".

(2) Linked Data Graph Mining. Building on relevant mining techniques, this paper proposes an iterative SPARQL query-based linked data graph mining method. The fundamental principle is to discover shortest association paths between nodes through iterative SPARQL retrieval, with the search strategy gradually increasing from length-1 paths. Using the subject term nodes "http://dbpedia.org/resource/Cloning" and "http://dbpedia.org/resource/Polymerase_{{{chain}}_{{reaction}}}" as starting nodes and setting the maximum mining path length to 3, we discovered 9,480 association paths in the DBpedia (2016-4) dataset, including 1 path of length 1, 72 paths of length 2, and 9,407 paths of length 3, as shown in Figure 6 [Figure 6: see original paper].

4.4 Semantic Revelation The semantic revelation module applies the cluster semantic relation revealing model to the results of linked data graph mining, consisting of three parts: association path classification, importance index calculation, and semantic relation revelation.

(1) Association Path Classification. After mining association paths, we classified the 9,480 discovered paths according to the model' s definitions of four path types: 1 path (0.01%) was direct relation, 1,076 paths (11.35%) were indirect relation, 3,847 paths (40.58%) were lowest common ancestor relation, and 4,556 paths (48.05%) were lowest common descendant relation.

(2) Path Importance Index Calculation. Based on available data and feasibility, we calculated attribute and node importance indexes using information theory methods and evaluated comprehensive path importance accordingly. The process was: first, dynamically obtaining required evaluation data such as attribute frequencies and node frequencies via SPARQL; then calculating attribute importance and node importance using the methods from Section 3; and finally evaluating comprehensive path importance based on normalized attribute and node importance indexes using an adjusted Min-Max normalization method expressed as x' = x / (Max + Max - Min). Table 1 shows partial comprehensive importance index calculation results, where "<>" *represents nodes and* " →" represents attributes.

(3) Semantic Relation Revelation. Before revealing semantic relationships expressed by association paths, we analyzed the semantic meanings of attribute relationships in linked data. As shown in Table 2 , we retrieved and analyzed high-frequency attributes in DBpedia via SPARQL, defining linked data semantic relationships as: equivalence relationships (including synonyms and near-synonyms), hierarchical relationships (genus-species), part-whole relationships, class-instance relationships, and associative relationships (all other relationships beyond the four types above). Based on these five fundamental semantic rela-

tionships, we analyzed the semantic relationships contained in association paths.

Direct Relation Revelation. Subject term nodes "Cloning"and "PCR"have 1 direct relation: {Cloning → }, where the attribute "http://dbpedia.org/ontology/wikiPageWikiLink" represents associative relationships. It indicates that subject terms "Cloning" and "PCR" have a relevant relationship.

Indirect Relation Revelation. There are 1,076 indirect relations between subject term nodes "Cloning" and "PCR" , with the highest comprehensive importance path being: {Cloning → <Cloning_{vector}> → }, indicating that node "Cloning_{vector}" has associative relationships with both "Cloning" and "PCR" . Additionally, the experiment discovered multiple resources including "DNA", "DNA_{sequencing}", "DNA_{profiling}", and "Molecular_{cloning}" that also have associative relationships with both "Cloning" and "PCR" .

LCAR Revelation. There are 3,847 lowest common ancestor relations between subject term nodes "Cloning"and "PCR", with the highest comprehensive importance being: {Cloning → ← → }, indicating that subject terms "Cloning" and "PCR" both have hierarchical relationships with the class "Biotechnology" , meaning both concepts belong to the biotechnology class.

LCDR Revelation. There are 4,556 lowest common descendant relations between subject term nodes "Cloning" and "PCR" , with the highest comprehensive importance being: {Cloning ← <Category:Molecular_{biology}> → }, indicating that subject terms "Cloning"and "PCR" both have relationships with the class "Molecular_{biology}", meaning both belong to the molecular biology class. Additionally, the path {Cloning ← http://dbpedia.org/dbtax/Technique → } shows that subject terms "Cloning" and "PCR" both have class-instance relationships with the class "http://dbpedia.org/dbtax/Technique" , meaning both are types of techniques.

4.5 Experimental Results Analysis Analysis of the top 300 paths by importance index revealed that 136 paths were worthless due to linked data incompleteness and other quality issues. Among the remaining 164 semantically valuable paths, LCDR accounted for 106 paths (64.6%), LCAR for 54 paths (32.9%), IR for 3 paths (1.8%), and DR for 1 path (0.6%), indicating that LCAR and LCDR are most important for cluster semantic revelation. Analysis of semantic relationship types revealed by these 164 paths showed that associative relationships dominated at 92.7% (152 paths), followed by class-instance relationships at 4.8% (8 paths), and hierarchical relationships at 2.4% (4 paths). The high proportion of associative relationships is primarily because the experimental dataset DBpedia is extracted from Wikipedia, containing numerous attributes related to Wikipedia page information. For example, the attribute "http://dbpedia.org/ontology/wikiPageWikiLink"appears 170 million times, accounting for approximately one-quarter of the dataset' s total attributes (680 million). The abundance of these associative relationship attributes leads to the highest proportion of associative relationships in semantic revelation results.

This experiment effectively revealed multiple semantic relationships between subject terms using linked data, including associative relationships, class-instance relationships, and class-attribute relationships. For example: subject terms "Cloning" and "PCR" both belong to the biotechnology class; both belong to the molecular biology class; and both represent techniques. In the paper "Research Focus Analysis and Preliminary Outlook in Veterinary Molecular Biology Based on Co-word Analysis", experts manually analyzed the cluster containing "Cloning" and "PCR" and named it "Cloning Technology Research", which aligns with our experimental semantic revelation results, demonstrating the feasibility and effectiveness of the proposed model.

The experiment also had some limitations. First, model validation was based solely on a single DBpedia English dataset, limiting revealed semantic relationship types to associative, class-instance, and hierarchical relationships. Additionally, linked data quality issues such as incompleteness, duplication, and inconsistency affected semantic revelation accuracy.

This study proposes using linked data to reveal semantic relationships among subject terms within clusters and validates the model' s effectiveness through empirical experiments, addressing the limitations of traditional cluster analysis in semantic relationship revelation and providing a new approach for cluster semantic relation revelation. Compared to other corpora, linked data offers dual advantages of broad semantic resource coverage and high structural organization, with rapidly developing LOD resources ensuring effective semantic revelation for clusters in most domains. This research has two main limitations: semantic revelation limited to a single dataset, and the impact of linked data quality on revelation results. Future research will investigate cluster semantic revelation based on multiple linked data resources and improve association path importance evaluation metrics to overcome the impact of linked data quality on semantic revelation results.

References

[1] Zhong Weijin, Li Jia. The Research of Co-word Analysis (1)—The Process and Methods of Co-word Analysis [J]. Journal of Intelligence, 2008, 27(5): 70-72.

[2] Zhang Shuliang, Leng Fuhai. Study on the Applicational Development of Literature-based Knowledge Discovery [J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(6): 700-712.

[3] Zhang Han, Ren Zhiguo, Zhang Jian, et al. Study on the Data Mining in Medical Text Database Based on Keywords Association Rules [J]. Journal of Medical Informatics, 2008, 29(1): 32-35.

[4] Zhang Han, Cui Lei. Study of Bioinformatics through Co-word Analysis[J]. Journal of the China Society for Scientific and Technical Information, 2003, 22(5): 613-617.

[5] Cimino J J, Barnett G O. Automatic Knowledge Acquisition from Medline [J]. Methods of Information in Medicine, 1993, 32(2): 120-130.

[6] Liu Mingyan. Research of Text Mining About Semantic Relation Recognition[D]. Nanjing: Nanjing University of Science and Technology, 2010.

[7] Zhang Xiaogang. Traditional Chinese Medical Ontology Based Semantic Relation Discovering and Verification Method[D]. Hangzhou: Zhejiang University, 2010.

[8] Wei Lai. Research of Folksonomy Semantic Association Method Based on Online Thesaurus [J]. Library and Information Service, 2011, 55(5): 104-108.

[9] Tiddi I, D' Aquin M, Motta E. Dedalo: Looking for Clusters Explanations in a Labyrinth of Linked Data [M]. Springer International Publishing, 2014.

[10] Taheriyan M, Knoblock C A, Szekely P, et al. Leveraging Linked Data to Infer Semantic Relations Within Structured Sources[C]// Proceedings of the 6th International Workshop on Consuming Linked Data (COLD). 2015.

[11] Li Nan, Zhang Xuefu. Research on Knowledge Discovery Based on Linked Data [J]. Researches in Library Science, 2013, 1: 73-77.

[12] Li Jun, Huang Chunyi. Knowledge Discovery in Linked Data [J]. Information Science, 2013, 31(3): 79-84.

[13] Gao Jinsong, Li Yingying, Liu Long, et al. Research on Construction of the Knowledge Discovery Model Based on Linked Data [J]. Information Science, 2016, 34(6): 10-13.

[14] Song Lina. Research on Model of Knowledge Discovery Based on Knowledge Map Under the Environment of Linked Data [D]. Wuhan: Central China Normal University, 2014.

[15] Liu Long. Research on Model of Knowledge Discovery Process Based on Linked Data [D]. Wuhan: Central China Normal University, 2014.

[16] Narasimha V, Kappara P, Ichise R, et al. LiDDM: A Data Mining System for Linked Data [C]// Proceedings of the 2011 Linked Data on the Web. 2011.

[17] Paulheim H, Fürnkranz J. Unsupervised Generation of Data Mining Features from Linked Open Data[C]//Proceedings of the International Conference on Web Intelligence, Mining and Semantics. 2012.

[18] Ramezani R, Saraee M, Nematbakhsh M A. Finding Association Rules in Linked Data, A Centralization Approach[C]//Proceedings of the 21st Iranian Conference on Electrical Engineering. 2013.

[19] Personeni G, Daget S, Bonnet C, et al. Mining Linked Open Data: A Case Study with Genes Responsible for Intellectual Disability [M]. Springer International Publishing, 2014.

[20] Jiang X, Zhang X, Gao F, et al. Graph Compression Strategies for Instance-Focused Semantic Mining [C]//Proceedings of the 7th Chinese Semantic Web Symposium on Linked Data and Knowledge Graph. 2013.

[21] Li K, Gao J, Guo S, et al. LRBM: A Restricted Boltzmann Machine Based Approach for Representation Learning on Linked Data[C]// Proceedings of the IEEE International Conference on Data Mining. 2014.

[22] Xia Lixin, Tan Ying. Analysis and Visualization of the LOD Network Structure [J]. New Technology of Library and Information Service, 2016(1): 65-72.

[23] Meymandpour R, Davis J G. Linked Data Informativeness [M]. Springer Berlin Heidelberg, 2013.

[24] Kasneci G, Elbassuoni S, Weikum G. MING: Mining Informative Entity-Relationship Subgraphs [C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009.

[25] Balmin A, Hristidis V, Papakonstantinou Y. Objectrank: Authority-based Keyword Search in Databases[C]// Proceedings of the 30th International Conference on Very Large Data Bases.2004.

[26] Nie Z, Zhang Y, Wen J R, et al. Object-level Ranking: Bringing Order to Web Objects[C]//Proceedings of the 2005 International Conference on World Wide Web. 2005.

[27] Ng M K P, Li X T, Ye Y M. MultiRank: Co-ranking for Objects and Relations in Multi-relational Data [C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011.

[28] Jiang Shiyin, Li Chunwang. Review on the Evaluation of Scientific Research Institution Based on Linked Data [J]. Information Studies: Theory & Application, 2015, 38(2): 136-140.

[29] Bamba B, Mukherjea S. Utilizing Resource Importance for Ranking Semantic Web Query Results[C]//Proceedings of the International Conference on Semantic Web and Databases. 2004.

[30] Franz T, Schultz A, Sizov S, et al. TripleRank: Ranking Semantic Web Data by Tensor Decomposition[C]//Proceedings of the International Semantic Web Conference. 2009.

[31] Hulpus I, Prangnawarat N, Hayes C. Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation[C]// Proceedings of International Semantic Web Conference. 2015.

[32] Yue Yang, Sun Jing, Shi Dayou, et al. Interpretation and Preliminary Outlook of the Research Focus in Veterinary Molecular Biology Based on the Co-word Analysis [J]. Guangdong Journal of Animal and Veterinary Science, 2015, 40(2): 1-4.

Author Contributions: Cui Jiawang: literature collection, program design, paper writing; Li Chunwang: research concept, paper review and revision.

Conflict of Interest Statement: All authors declare no conflict of interest.

Supporting Data: Supporting data is self-archived by the authors, E-mail: cuijiawang@mail.las.ac.cn. [1] Cui Jiawang. Linked Data Mining_{9480}.xls. Experimental dataset.

Received: 2017-02-16 **Revised:** 2017-04-11

Note: Figure translations are in progress. See original paper for figures.