Scientific Research Publishing

Modern Economy, 2020, 11, 686-700
https://www.scirp.org/journal/me
ISSN Online: 2152-7261
ISSN Print: 2152-7245

Semantic Primitives Extraction for XBRL Domain Ontology

Di Ye, Ding Pan

School of Management, Jinan University, Guangzhou, China
Email: ye950608@foxmail.com

How to cite this paper: Ye, D., & Pan, D. (2020). Semantic Primitives Extraction for XBRL Domain Ontology. *Modern Economy, 11,* 686-700.
https://doi.org/10.4236/me.2020.113051

Received: February 9, 2020
Accepted: March 20, 2020
Published: March 23, 2020

Abstract

At present, XBRL (Extensible Business Reporting Language) has been used in more and more countries and organizations. Although XBRL has achieved a series of remarkable achievements, the current development of XBRL has also encountered bottlenecks. On the one hand, the XBRL field has not been unified. On the other hand, the semantics of concepts in the domain ontology of XBRL is weak, resulting in the slow application and promotion of XBRL. This paper regulates the extraction of semantic primitives in the ontology of XBRL domain from the perspective of semantics and solves the problem of "how to extract semantic primitives". The solution to this problem can promote computers to better understand XBRL financial reports and reduce the technical difficulty of XBRL. This paper comprehensively uses the theories of semantic primitives, graph theory, and domain ontology. First, it analyzes the research status and shortcomings of semantic primitive extraction methods. Second, it constructs a graph of accounting term relationship network from the perspective of graph theory. The extraction of semantic primitives. Finally, the validity of semantic primitive extraction is analyzed and verified.

Keywords

XBRL Domain Ontology, Semantic Primitives, List of Elements, PageRank

1. Introduction

Currently in the era of big data, information technology continues to develop, XBRL is used as a digital standard for business reporting. With its enhanced semantic expression mechanism, it has become an effective method to break the silo of information. XBRL has been widely used by many companies around the world. As of October 2019, more than 150 important regulatory agencies in various countries have required the standard to be used to submit business reports for more than 20 million companies (XBRL Conference, 2019). XBRL has become the mainstream digital financial report in global capital markets, financial regulation and other areas.

At present, China and the United States have implemented their own XBRL classification standards. Because different countries have different accounting standards, the XBRL classification standards are also different, which makes it difficult to convert information between classification standards. In order to realize the conversion of XBRL financial information, it is usually necessary to establish a mapping relationship between different classification standards. If there are N different classification criteria, N^2 data conversion templates need to be established. When adding classification criteria, you also need to add mapping relationships. The value of the conversion N gradually increases, and the cost of creating a data conversion template also increases significantly. There are many related solutions for the interoperability research of different data. The ontology method is one of the more effective methods.

Domain ontology is a specification of the concepts in the domain. By representing concepts and relationships to reflect the structure of knowledge, it helps to enhance human-computer interaction and information exchange between machines. For the field of financial reporting, the XBRL domain ontology is the formal concept and example of financial reporting. Relying on the classification standards that can be automatically derived by the XBRL domain ontology, it can support the inference verification and semantic control of financial data, so the research on the XBRL domain ontology is very necessary. However, the system ontology has not yet been constructed in the field of financial reporting, and the research on financial reporting based on ontology has mostly focused on the simple verification of the theory. The main reason is that the standards for the description of financial reporting domain knowledge are not uniform. In addition, the specificity and scalability of concepts also make the construction process of domain ontology difficult.

In general, the reference to a concept is a term. Domain ontology is a logical theory of terminology. This term needs to be expressed with the help of a "core language". Therefore, how to construct this "core language" is the basis for establishing the XBRL domain ontology. Semantic primitives are carriers of the conceptual and attribute meanings of terms and are the controlling factors of semantic representation. They must have the following characteristics: 1) They are the basic components of semantics and the structural materials of semantics; 2) They are the smallest particles of semantics, that is, they can no longer be refined or decomposed into smaller particles; 3) They can be passed from the superordinate to the subordinate. Semantic primitives, as the basic material for describing the semantic microstructure, are the basic carrier of semantic information. They have the advantages of simplicity, exhaustion, and systematisms, which help to enhance the semantic interpretation of accounting terms Ability and formal expression.

In order to be able to construct the XBRL domain ontology more effectively, this paper applies semantic primitives to the field of financial reporting and builds a "core language" based on XBRL domain knowledge, which is also an important foundation for the construction of domain ontology. Therefore, how to extract semantic primitives to realize the construction of XBRL domain ontology is a scientific issue worthy of research. Based on the expression needs of the ontology in the XBRL domain, this paper uses the relevant corpus of the accounting field to construct a directed graph of accounting terms. To address the shortcomings of the current primitive extraction method, this paper uses the PageRank algorithm to extract semantic primitives that can describe the knowledge of the accounting domain. The expression of element list is realized to verify the validity of semantic primitive extraction.

(Ochoa et al., 2013) proposed a domain ontology construction method based on semantic role labeling. (Obitko et al., 2004) used Stanford University's natural language processing tools to preprocess the input text first and obtained the PCFGs syntactic analysis results and dependent syntactic analysis results of each sentence. (Jiang & Tan, 2010) first used some natural language processing tools to perform full-text analysis on input documents, including part-of-speech tagging, syntactic analysis, and word sense disambiguation. Finally, the author uses WordNet as a reference to classify and integrate the extracted concepts and corresponding relationships into the final domain ontology. (Hou et al., 2011) proposed a method for automatic construction of domain ontology based on graph theory. (Shih et al., 2011) proposed a domain ontology construction method based on a crystallizing model. (Lee et al., 2007) adopted a self-organizing mapping clustering algorithm for concept clustering and defined the hierarchical relationship between concepts based on the clustering results. (Salton et al., 1975) first introduced machine learning technology to automatic keyword extraction. Then (Frank et al., 1998) used Naive Bayesian model for keyword extraction. (Zhang Kuo, 2006) transformed the keyword extraction problem into a classification problem and classified the words/phrases in the document into three categories: good keywords, neutral keywords, and poor keywords, and then used the SVM model to classify the document words. Get keywords. (Wu et al., 2010) extended the graph representation of text to a semantic-based graph structure representation on the basis of Schenker's research, so that the semantic information of text can be used to improve the performance of text processing.

Inadequate research:

1) The current research only stays at the lexical level and does not go deep into the semantic level.

2) The degree of fit with domain knowledge is not strong.

3) The research perspective is single and there are few cross-domain research results.

So, the innovation of this article lies in:

1) Select an accounting dictionary that is more in line with domain knowledge as the data source;

2) Semantic primitives are linguistic concepts, which realize the expression of knowledge in the field of accounting with semantic primitives and realize cross-domain research.

2. Financial Report Vocabulary Characteristics

In order to make the extraction of semantic primitives more practical and pertinent, this chapter analyzes the stylistic characteristics of the notes to the financial statements. Through a combination of qualitative and quantitative methods, the characteristics of the notes to the financial statements at both the structure and vocabulary level are explained as a guide and basis for the subsequent extraction of semantic primitives.

At the same time, in order to clarify whether the stylistic characteristics of the notes to the financial statements have industry differences, this article is based on the 2012 industry classification standard of the Securities Regulatory Commission. This standard divides the industry into 19 categories. Each industry selects 10 listed companies. The Center selected the notes of the financial statements of 190 listed companies on December 31, 2018 for text analysis.

This section mainly conducts quantitative analysis from three aspects: part-of-speech type, compound words and term polysemy. The purpose is to clarify the wording characteristics of financial reports and provide guidance and basis for the subsequent extraction of semantic primitives.

2.1. Part of Speech

This text uses python3.7 to process the text of the notes of the 190 listed companies. It mainly includes four parts: Jieba cut words, stop words, part of speech tagging and part of speech statistics. As can be seen from Table 1.

Then the same treatment was performed on the 190 notes of the financial statements of listed companies, and the frequency of occurrence of each part of speech was counted.

As can be seen from Figure 1, nouns are the main parts of speech in financial reports. Therefore, when extracting semantic primitives, we prefer to select words that are mainly nouns and have high PR values, which is consistent with the part of speech distribution in financial reports.

In order to investigate whether there are large differences in the frequency of part-of-speech between different industries, this industry chooses 10 points of the financial statement notes for statistics. The results are shown in Figure 2.

As can be seen from Figure 2, under different part-of-speech, the frequency difference between industries is small. This is due to the fact that financial reporting is a rigorous and standardized document. Although there are differences in the types of accounting affairs between different industries, they remain consistent at the part-of-speech level.

The purpose of counting the differences in the part-of-speech frequency of different industries is to confirm whether there is a large difference in word characteristics between different industries. Under the condition that the parts-of-speech

Table 1. Part of speech tag code.

Abbreviation	Part of Speech
n	Common noun
p	Preposition
nz	Other proper nouns
a	Adjective
ns	Position noun
d	Adverb
m	Time noun
r	Pronoun
v	Verb
c	Conjunction

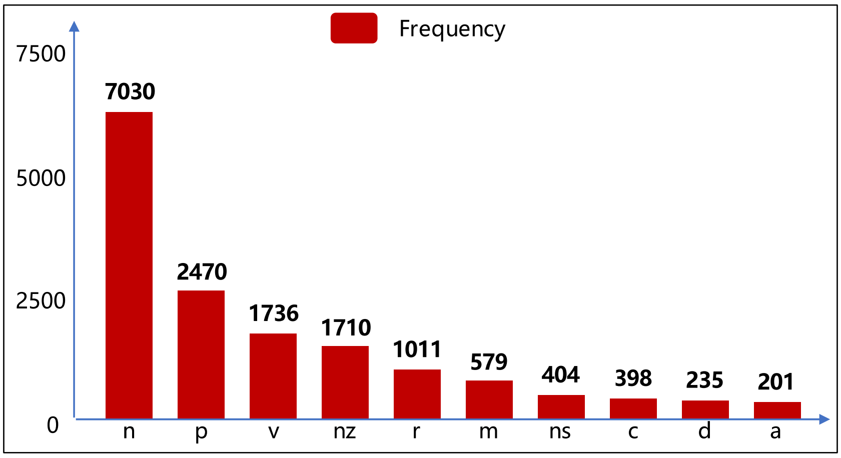

Figure 1. Frequency of parts of speech.

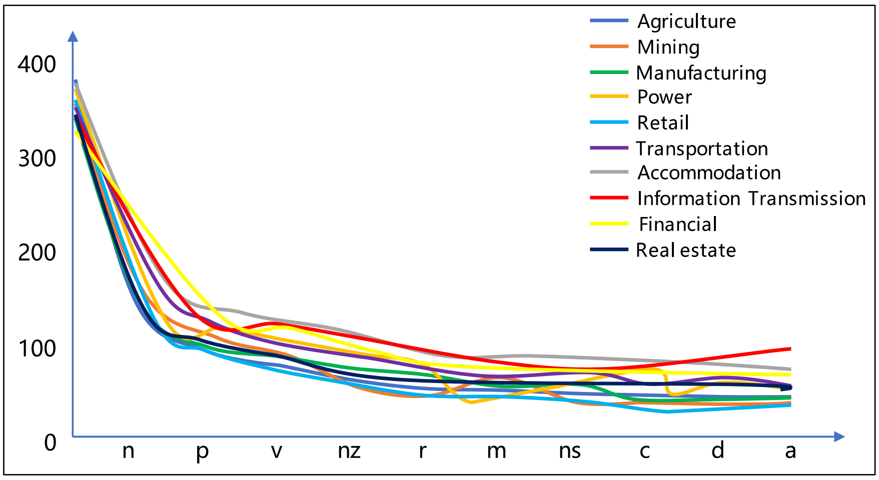

Figure 2. Frequency of each part of speech type in different industries.

frequency of each industry is roughly balanced, the financial report texts of different companies in different industries can be used when extracting the semantic meanings Seen as a whole text library for processing.

2.2. Compound Word

Due to the continuous updating of accounting affairs, more vocabulary is needed to describe the affairs. At the same time, in order to ensure the consistency of the accounting knowledge system, new words based on basic vocabulary appear in accounting, such as "assets" versus "long-term assets", "Current assets", etc. Such words are called compound words in this article. The emergence of compound words can enrich the field vocabulary to a large extent. Therefore, the words describing accounting affairs are mainly composed of two types: basic words and compound words.

The basic vocabulary occupies a core position in the accounting vocabulary system. It is quite stable and rarely changes but has the ability to form new words. There are two basic vocabularies of accounting, which are used to express data information. One is to express economic information at the time (assets, liabilities, owner's equity), and the other is to express financial information for the period (income, expenses, profit). Compound words are an in-depth study of basic words. According to the summary of the accounting elements, assets generally include current assets, long-term investments, fixed assets, etc. Liabilities generally include current liabilities, long-term liabilities, etc., while current assets include cash, receivables, and long-term liabilities including long-term loans and long-term payables. Compound words are more variable than basic words, including the constant addition of new words and the withdrawal or revision of meaning of some old words. For example, with the development of financial instruments, derivative financial assets have begun to enter financial reporting; subjects such as fixed asset maintenance funds have now withdrawn from financial reporting with changes in national economic management policies.

The frequency of occurrence based on basic words and compound words is shown in Figure 3.

It can be seen from the figure that the number of occurrences of compound words is greater than that of basic words. This is because the basic words are more stable, and compound words are constantly adding new words with the development of accounting affairs. Therefore, the relationship between compound words and accounting affairs is closer. However, it should be noted that, since the text is first segmented when extracting semantic primitives, in order to ensure the integrity of the compound word, a custom compound word dictionary is introduced when performing the segmentation, so that the resulting segmentation result is completed and more accurate.

It can be seen from Figure 4 that although there are differences in basic words and compound words between different industries, the frequency of compound words is higher than basic words in the same industry. This is because compound words are closer to accounting practice. Therefore, in the subsequent extraction of semantic primitives, the financial reports of different industries can be regarded as a whole and the integrity of the compound vocabulary after text processing needs to be paid attention to.

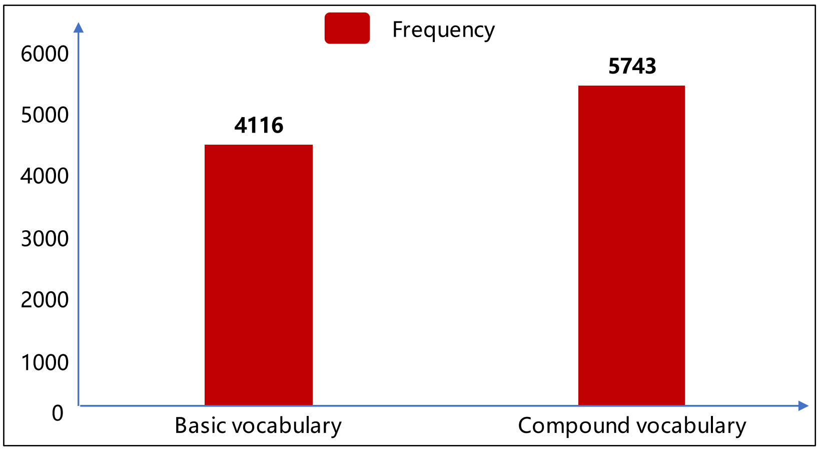

Figure 3. Frequency comparison of basic words and compound words.

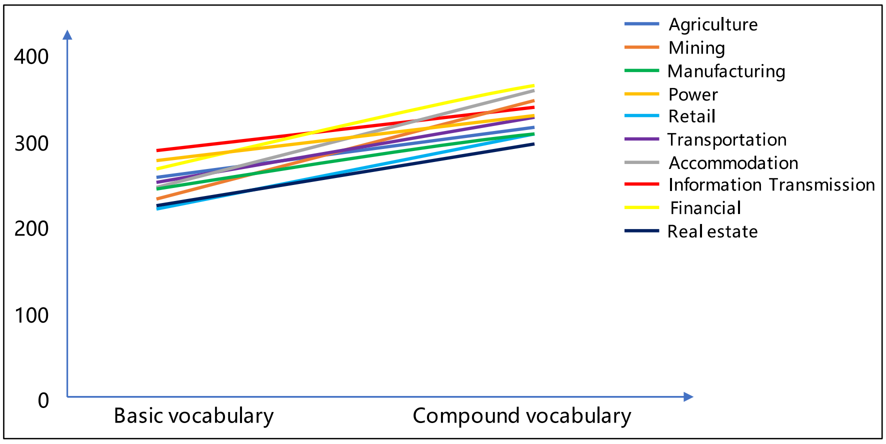

Figure 4. Frequency differences of basic words and compound words in different industries.

From the above analysis, it can be seen that the financial report is a rigorous and standardized text format. The frequency of part-of-speech tagging and compound words between different industries is roughly the same as the basic word frequency. And in order to better match accounting transactions, a large number of compound vocabularies in the field of accounting are used. Therefore, in the subsequent extraction of semantic primitives, financial reports of different industries are considered as a whole, and the importance of compound words in text processing needs to be paid attention to because Compound words are closer to specific accounting practices, so in the subsequent extraction of semantic primitives, the integrity of the compound words after word segmentation needs to be guaranteed. The above analysis provides guidance and basis for the extraction of semantic primitives in the following.

3. Semantic Primitive Extraction Model Oriented to XBRL Domain Ontology

3.1. Problem Description

The accounting dictionary is a set of word strings X composed of accounting terms and their corresponding definition texts. The problem of extracting semantic primitives from an accounting dictionary can be described as:

There is a hypothetical set p and $p \subseteq X$ in the dictionary; if the word $w \in p$, then w is defined by the hypothetical set p in the 0th round; if each word in the definition of the word w is assumed by the hypothetical set p in $r(0 \le r \le k-1)$ round definition, then it is said that w is defined by the hypothetical set p in the round k; the set consisting of words defined by the hypothetical set p in the $\le k(k \ge 0)$ rounds is called the k round definition of p, set $D(p,k)$; if $D(p,k) = D(p,k+1)$, then $D(p,k)$ is called the largest defined set of hypothetical set p, and $\max(p)$; If $\max(p) = X$, p is called a set of semantic primitives in the dictionary.

In this way, the problem of obtaining a minimum set of semantic primitives in a dictionary is to give a dictionary, find the set of semantic primitives p in this dictionary, and for any set of semantic primitives p' in the dictionary, $|p| \le |p'|$ (refers to the number of elements).

The existence of the semantic primitive set is further explained below. Considering the simplest case, all the words in the dictionary are included in the set P. According to the definition, $D(p,0) = P = X$, that is, P is a set of semantic primitives of the dictionary. Therefore, a set of semantic primitives exists.

According to this formal description, a dictionary can be transformed into a directed graph, and the problem of obtaining primitives from the dictionary can be transformed into a graph theory problem. A directed graph can be represented by $G = (N, E)$, where N represents the set of nodes in the graph and E represents the set of directed edges in the graph. A dictionary can uniquely correspond to a directed graph in the following ways:

For each word w in the word set, there is only one node n corresponding to it in the graph; and for each node n in the word set, there is only one word w corresponding to it;

For any two words w_1, w_2 (where w_1 corresponds to n_1, and w_2 corresponds to n_2), if w_2 appears in the definition string of w_1, there is a directed edge from n_1 to n_2 in the figure; otherwise, the graph There are no directed edges in n_1 to n_2.

This directed graph has the following properties:

1) No self-loop exists. When a defined word appears in a word string defined by itself, it is not considered. In this way, a directed graph does not have a node to its directed edge, that is, there is no self-loop.

2) The output degree of each node is greater than or equal to 1, because each word in the dictionary has a definition, and it is impossible to define itself only by the word itself.

In the following, the problem of obtaining primitives from a dictionary corresponds to a problem in graph theory.

A hypothetical set p in graph $G = (N, E)$ is a subset of N; if node n is an element in hypothetical set p, then n is defined by the hypothesis set p in the 0th round; for node n, If every node n' in the graph directly reached from n through a directed edge (that is, from n to n', there is a directed edge in the graph), all the imaginary sets p are in $r\left(0 \le r \le k-1\right)$ Round definition, it is said that n is defined by the hypothetical set p in the round k; the set of nodes defined by the hypothetical set p in $\le k\left(k \ge 0\right)$ rounds is called the round k definition set of p, and $D\left(p,k\right)$ Representation; if $D\left(p,k\right)=D\left(p,k+1\right)$, then $D\left(p,k\right)$ is called the largest defined set of hypothetical set p, expressed by $\max\left(p\right)$; if $\max\left(p\right)=N$, we call p a set of semantic primitives in the graph.

It can be seen that the problem of obtaining the smallest set of semantic primitives can be transformed into the following graph theory problem: find a semantic primitive set p in the graph, so that any semantic primitive set p' in the graph has $\left|p\right| \le \left|p'\right|$.

3.2. PageRank

The PageRank algorithm is a link analysis algorithm proposed by Google founders Larry Page and Sergey Brin in 1998 to rank Internet pages. In the WWW network map model, for a web page S, its PageRank value is based on two assumptions: the first is a quantitative assumption, that is, if the number of incoming links on the web page S is greater, the quality of the web page S is higher, the second is the quality assumption. If a web page receives links from high-quality pages to it, the better the quality of this web page, we know that the quality of the web page pointing to the S page is uneven, so the higher the quality of the web page pointing to S, The higher the quality of S pages, the PageRank algorithm is based on this principle to rank the pages. Here we abstract the Internet into a directed graph model. Assuming the number of web pages is n, the model is represented by graph $G=\left(V,E\right)$, where V represents vertices, E represents edges, and the number of vertices is n. By linking the graph, an adjacency matrix H can be established, take any $h_{ij} \in H$, if there is a link from webpage i to j, then $h_{ij}=1$, otherwise it is 0. Then define the output degree of node i as $O_i=\sum_{j=1}^{n} h_{ij},\ i=1,2,3,\cdots n$. The PageRank value of node j is thus shown in Equation (1):

$$PR\left(j\right)=\sum_{i=1}^{n}\frac{PR\left(i\right)}{O_i} \tag{1}$$

That is, the PageRank value of node j is not only affected by the PageRank value of node i, but also by the output degree of node i. Formula (1) is the original model of PageRank algorithm.

We divide each element in the above adjacency matrix H by each row sum to get the normalized adjacency matrix A, that is, $A_{ij}=\begin{cases}\frac{h_{ij}}{O_i}, \text{if } O_i>0\\ 0, \text{if } O_i=0\end{cases}$. We define the transition probability matrix of graph G as M. After transposing the adjacency matrix A, we obtain matrix M. Any element M_{ij} in the matrix M represents the conditional transition probability $P\left(i\middle|j\right)$ from web page j to

i. The PageRank value of a node can be abbreviated as $P = MP$. Obviously from the formula $P = MP$, it can be seen that this is the eigenvector P of the required matrix M, and the eigenvalue corresponding to this eigenvector is 1, which can be achieved by the iterative method. However, due to the randomness of the link relationship, there may be "dangling nodes" in the network links. These nodes have no links to other nodes, that is, its outdegree is 0, then there will be a column in the transition probability matrix that is all 0, Causing the matrix to fail to converge. To use the iterative method to calculate feature vectors, three conditions must be met. Before talking about the three conditions, let's take a look at the model nature of the PageRank algorithm.

The link relationship of the web pages here is actually a Markov model. Each web page is equivalent to the "state" of the random surfer, and the PageRank value is equivalent to the probability of the random surfer in each state. In the original PageRank, Brin and Page described the random surfing model like this: "Suppose there is a random surfer on the Internet, who starts browsing from a randomly selected page. If the output of the current page is greater than zero, then the probability λ (0 < λ < 1) Randomly click a hypertext link on the current page to enter the next page, with a probability of 1 − λ completely randomly selecting a page on the entire WWW as the next page to be browsed; if the output of the current page is 0, choose a page completely randomly as the next page to be viewed. "With the introduction of this model, the problem of overhang and convergence of the operation are solved. There is actually a deep mathematical principle behind this; if the PageRank algorithm is to be successfully solved, for the random matrix A (that is, the normalized adjacency matrix), three conditions should be satisfied: 1) A is a random process matrix. That is, each element is positive and the sum of the elements in each row adds up to 1. For the dangling point, that is, the point that has no degree, it is obviously not satisfied. All the elements in the corresponding row are 0 (and must also be 0), so Brin and Page cleverly replace these 0 with $1/N$ to ensure that A is a random process matrix; 2) A is not simplifiable. In other words, A is a strongly connected graph, that is, all points in this directed network can reach each other; 3) A is aperiodic. In a Markov chain, there is more than one path from state i to this state, and the lengths of the paths are also different. If the length of these paths has a minimum common divisor k and $k > 1$, then state i is said to be periodic. The implementation of (2) and (3) is also realized by $1/N$, that is, all points are added to $1/N$, whether it is a dangling point or not, so that matrix A becomes a strongly connected graph and is obviously aperiodic. In this way, the revised PageRank algorithm of the web page is shown in format (2):

$$PR(j) = \frac{1-\lambda}{N} + \lambda \sum_i PR(i) \times \frac{1}{O_i} \tag{2}$$

where λ is the damping factor, and generally takes 0.85. Simplify Equation (2) as $P = \left(\frac{(1-\lambda)E}{N} + \lambda M \right) P$, Where E is a matrix of all 1s of $n \times n$, where the new transition probability matrix is essentially $B=\frac{(1-\lambda)E}{N}+\lambda M$. At the beginning, each page is given an initial PageRank value. The size of this initial PageRank value does not matter. Generally, it is set to $1/N$. As the number of iteration cycles increases, the PR value of each page will eventually converge to *A* stable value, that is, the final steady-state distribution column vector PR, that is, the eigenvector of the transition probability matrix *B*, the revised formula is the most widely used.

4. Experiment

In this article, the "English-Chinese Modern Finance and Accounting Dictionary" edited by China Finance and Economics Publishing House Chen Jinchi in 2009 was used as experimental data. From this, 4289 accounting terms and 32086 terms were sorted out, which were used as experimental texts in the field of accounting.

Text Processing

The main programs and software used for data processing here are: Excel2016, Python3.7, MATLAB R2016a, etc. Among them, Excel is used for the structured organization of accounting dictionaries, and the jieba package of Python is used to cut word definitions and draw based on MATLAB Directed loop graph and calculate PageRank value. The specific work is as follows:

1) Manually extract and organize definition texts of accounting terms.

According to the text analysis of the accounting dictionary above, in the dictionary, there is not only a descriptive description for the interpretation of an accounting term, but also non-definitive descriptions such as examples and calculation formulas, and the extraction of semantic primitives in this part It is a redundant part, so this article manually extracts and organizes the definition text of accounting terms, and summarizes it in Excel.

2) Text cutting, stop words, and duplicate processing.

Then use the jieba package that comes with Python to cut words. It is worth noting that in order to ensure the completeness of accounting terms, it is necessary to import 4289 accounting terms in the accounting dictionary into a custom dictionary, and then establish a stop-word list, and the terms in the definition text of a term are deduplicated.

3) Construct a directed network diagram of accounting terms.

According to the above word segmentation results, a directed loop graph can be constructed for these texts, as shown in Figure 5. The specific construction idea is to use the vocabulary and the definition text after the word as the node. There is a directed edge between the vocabulary and the definition text. Specifically, the vocabulary points to several definition text words. If there is another word *B* (such as lease) in the definition text, then there is a directed edge between *A* and *B*, specifically a directed edge where *A* points to *B*. Graphically describe the above relationship.

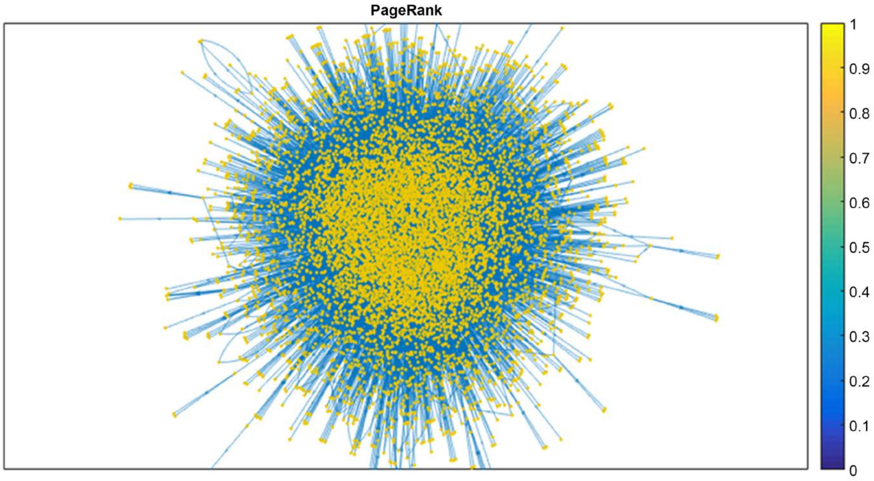

Figure 5. Example of loop diagram and PageRank value distribution.

4) Calculate the PageRank value.

After constructing the network graph based on the accounting dictionary, the PageRank value of each node was calculated using MATLAB R2016a as the basis for the extraction of semantic primitives.

The selection of semantic primitives in this paper follows two principles: 1) Leaf nodes must be semantic primitives. Because leaf nodes represent interpreted vocabulary, which conforms to the definition of semantic primitives, it is a vocabulary used to explain domain knowledge. In the path where leaf nodes are located, the PageRank algorithm is invalid. The calculation of the PR value is performed at a point. 2) The node in the loop selects the node with the higher PR value. Based on the characteristics of PageRank's algorithm, it can be concluded that the point with the higher PR value is the semantic primitive.

As can be seen from **Figure 5**, the PR value of leaf nodes is generally high, and because semantic primitives are used to explain other words and cannot be interpreted by themselves, the extraction of semantic primitives should be at the leaf nodes. This is consistent with the analysis above. For the nodes on the loop, this article will choose the point with the highest PR value as the semantic primitive.

5. Conclusion

The extraction of semantic primitives is an important part of constructing the ontology in the field of XBRL, and its research has received widespread attention from scholars. Based on previous studies, the text first combed the current extraction methods of semantic primitives, including linguistic-based extraction methods, statistical-based extraction methods, machine-learning-based extraction methods, and graph-theory-based extraction methods. The specific methods are introduced separately; the shortcomings are pointed out; and according to the characteristics of the XBRL field, graph theory and PageRank algorithm are proposed to extract semantic primitives. Then, starting from the XBRL domain ontology, this paper analyzes the lexical features of the element list, which is an important corpus for constructing the XBRL domain ontology, and provides ideas and basis for the extraction of semantic primitives later. Then it introduces the construction of accounting term relation network and introduces the basic principle and processing process of PageRank algorithm in detail and explains the applicability of this algorithm to XBRL field. In order to test the validity of the extracted semantic primitives, a structural expression of the list of elements was implemented using semantic primitives.

Through the discussion, this article mainly draws the following conclusions:

1) This paper is oriented to the domain ontology of XBRL, taking into account the needs of ontology expression and domain characteristics, to achieve the extraction of semantic primitives.

This study is different from the extraction of entities in the construction of ontology in the past. In the previous studies, the extraction of entities stayed only at the lexical level, that is, the simple text processing of the corpus. expression ability. And this article studies and analyzes the lexical features of the element list based on the important corpus of constructing the ontology in the XBRL field, and manually extracts the structural collocation of terms in the element list to provide guidance and basis for the extraction of semantic primitives in the following. Provide a verification path for verifying the validity of semantic primitive extraction later.

2) Use the PR value and the degree of nodes to achieve the double index extraction of semantic primitives.

According to the nature of semantic primitives, it can be known that semantic primitives are a small set of languages used to explain knowledge in other fields. Based on the rules for the construction of directed graphs, we can conclude that there are two forms of directed graphs constructed by accounting dictionaries, one is not looped, and the other is the situation on a loop. For the first non-looping case, we choose the leaf node as the semantic primitive, because the leaf node shows that the vocabulary can explain the vocabulary of the parent node and cannot be interpreted by itself, which meets the definition of the semantic primitive, but this case The scale of the semantic primitives extracted is large, and the scale effect of semantic primitive extraction needs to be further considered. For the second case on the loop, we extract the point with the largest PR value on the loop based on the PageRank algorithm as the semantics. The semantic primitives thus obtained can guarantee the comprehensiveness and scientific of the extraction.

The related conclusions of this study are similar to those of previous studies, and there are obvious differences. The level of extraction in this study is more comprehensive and the verification of the extraction effect is different from previous studies. The extracted semantic primitives are more intimate with actual needs, the research angle is more novel, and the method is more efficient.

Aiming at the above research conclusions, this paper summarizes the research suggestions for semantic primitive extraction for ontology in XBRL domain:

1) Starting from the term's interpretation strength, the term's expression of domain knowledge should be enhanced, and the extracted semantic primitives should meet the premise of effective disclosure of actual financial information.

2) Starting from the term structure, perfect the dimensional modeling and non-dimensional modeling of the terms, and by extracting the regularity of the internal structure of the terms, improve the expression of semantic primitives.

3) Consider the scale effect of semantic primitive extraction and pay attention to controlling the number of semantic primitive extraction to the minimum scale that satisfies the knowledge expression in the financial reporting field.

Extracting semantic primitives is one of the key tasks in constructing the ontology in the XBRL field. Based on the complexity of the research problem, this study also has the following deficiencies:

1) The sample size is too small: For the accounting field, there is a large amount of unstructured text information. In addition to the accounting dictionaries and element lists mentioned in this article, it also includes a large amount of available information such as corporate annual reports and financial news of listed companies. Combining accounting information at the structured, semi-structured, and unstructured levels can be used to sort out the domain characteristics of accounting information and expand the boundaries of accounting field knowledge.

2) Undivided industry areas: This article does not subdivide the fields between different industries but identifies the basic terms of the entire accounting field based on the overall perspective. In the future, the accounting field can be subdivided based on industry perspectives. Select 3 - 5 industries for research, such as education, construction, agriculture, forestry, animal husbandry, and fishery. Collect data from these industries and apply new algorithms for core semantic primitives. Discover and verify the scientific of the new algorithm.

Acknowledgements

The paper is supported by Natural Science Foundation of China (71771104): On XBRL domain ontology automatic construction.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

XBRL Conference (2019). *Extensible Business Reporting Language (XBRL) International Conference*. http://www.mengyinnews.com/html/flea/47993.html

Ochoa, J. L., Valencia-Garcia, R., Perez-Soltero, A., & Barcelo-Valenzuela, M. (2013). A Semantic Role Labelling-Based Framework for Learning Ontologies from Spanish Documents. *Experts Systems with Applications, 40,* 2058-2068. https://doi.org/10.1016/j.eswa.2012.10.017

Obitko, M., Snasel, V., & Smid, J. (2004). Ontology Design with Formal Concept Analysis. In *Proceedings of the 2nd International Workshop on Concept Lattices and Their*

Applications (pp. 111-119). Athens: CSREA Press.

Jiang, X., & Tan, A.-H. (2010). CRCTOL: A Semantic-Based Domain Ontology Learning System. *Journal of the American Society for Information Science and Technology, 61,* 150-168. https://doi.org/10.1002/asi.21231

Hou, X., Ong, S. K., Nee, A. Y. C. et al. (2011). GRAONTO: A Graph-Based Approach for Automatic Construction of Domain Ontology. *Expert Systems with Applications, 38,* 11958-11975. https://doi.org/10.1016/j.eswa.2011.03.090

Shih, C.-W., Chen, M.-Y., Chu, H.-C., & Chen, Y.-M. (2011). Enhancement of Domain Ontology Construction Using a Crystallizing Approach. *Experts Systems with Applications, 38,* 7544-7557. https://doi.org/10.1016/j.eswa.2010.12.112

Lee, C.-S., Kao, Y.-F., Kuo, Y.-H., & Wang, M.-H. (2007). Automated Ontology Construction for Unstructured Text Documents. *Data & Knowledge Engineering, 60,* 547-566. https://doi.org/10.1016/j.datak.2006.04.001

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of ACM, 18,* 613-620. https://doi.org/10.1145/361219.361220

Frank, E., Paynter, G. W., & Witten, I. H. (1998). Domain-Specific Key Phrase Extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence* (pp. 517-523). San Francisco: Morgan Kaufmann Publishers Inc.

Zhang, K., Xu, H., & Tang, J. (2006). Keyword Extraction Using Support Vector Machine. In *Proceedings of the Seventh International Conference on Web-Age Information Management* (pp. 85-96). Berlin: Springer. https://doi.org/10.1007/11775300_8

Wu, J. N. et al. (2010). Textual Knowledge Representation through the Semantic-Based Graph Structure in Clustering Applications. In *Proceedings of the 43rd Hawaii International Conference on System Sciences* (pp. 1-8). Piscataway, NJ: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HICSS.2010.366