Construction and Verification of Knowledge Base of Political & Economy News based on Mixed Algorithm of Subgraph Feature Extraction and RESCAL

Pin Wu, Juanjuan Luo*, Yonghua Zhu, Wenjie Zhang

Department of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

Abstract

With the intelligent development of digital government management services and the advancement of Knowledge Graph study, it is necessary and possible to construct and verify a sound knowledge base of political and economic news to satisfy the users' requirement of learning the information. Due to the high profession and diversity of political and economic news data, the entity link in the initially constructed knowledge base is lacking completeness. Meanwhile, the high frequency of data update leads to the iterative update of knowledge base. To address the problems, this paper builds a comparatively effective system in which we apply the reasoning results to the construction and iterative update of the political and economic news knowledge base. Then, a syncretic reasoning algorithm based on Subgraph Feature Extraction (SFE) and the factorization of a three-way tensor (RESCAL) is proposed to predict a link and accomplish the reasoning. Using the field data of political and economic news as a case of engineering application, the system we built effectively solves the incompleteness of the entity link in the initial knowledge base and the iterative update problem. The function of knowledge reasoning module and iteration module of knowledge base construction and autonomous updating system are verified by designing and implementing knowledge reasoning, as well as updating knowledge iteration. The experimental results demonstrate the effectiveness and feasibility of the functions of the knowledge base construction and autonomous updating system are verified.

Keywords: knowledge base; subgraph feature extraction; RESCAL; political and economy; news

(Submitted on October 29, 2017; Revised on November 12, 2017; Accepted on December 1, 2017)

© 2017 Totem Publisher, Inc. All rights reserved.

1. Introduction

In recent years, with the intelligent development of digital government management services and the rapid growth of political and economic news data, there is an urgent demand for users to acquire and understand such knowledge. The traditional search engines and query methods are unable to meet the growing demand. However, knowledge graph technology has become a hot issue for its extraordinary performance in improving the accuracy and scalability of knowledge query. Meanwhile, the development of knowledge graph technology makes it possible to construct a consolidated Knowledge Base of political and economic news. Besides, the characteristic of political and economic news data (such as historic) and the high frequency of updating data brings about the inconvenience of knowledge base. Therefore, to help relevant organizations make decisions and enterprises improve product quality, it is essential to construct a knowledge base of political and economic news that is complete, efficient and can be iteratively updated.

Researchers have built a variety of knowledge graphs or knowledge bases, such as Yago [27], DBpedia [2], and Freebase [5]. These knowledge bases play an important role in many applications, such as Question Answering System [8] and Recommendation System [29]. However, most of the current knowledge bases are general knowledge-bases, and there are few approaches to construct and represent knowledge bases in specialized field, especially in the political and economic news. Moreover, political and economic news are highly specialized and diverse, which always result in lack of integrity of
the entity link in the initially constructed knowledge base. At the same time, the high frequency of data update will also cause the demand for continuously iterative update of the knowledge base. Therefore, we need a reasoning algorithm to overcome the drawbacks in political and economic news knowledge base. Many scholars devoted themselves to the research and application of knowledge graph, thus bringing about many learning and reasoning algorithms of knowledge graph. Nickel et al. [19,20] proposed the factorization of a three-way tensor RESCAL. It decomposes the tensor value corresponding to each triple in the knowledge graph into entity and relation matrix, which makes the product of the decomposed entity and relation matrix as close to original tensor value, and applies to general large-scale knowledge graph reasoning and learning, extracting unknown relationships and entity resolution. This method achieved great prediction results. Socher et al. [26] proposed a neural network model called Neural Tensor Networks (NTN) by neural tensor networks constructed reasoning algorithm. Chang et al. [7] utilized relational domain knowledge about entity type information in the construction of training model and proposed TRESICAL. Krompaß et al. [14] added the constraint of relationship type in the tensor factorization to reduce the computational complexity during the factorization. Wu Yunbing et al. [3] proposed a knowledge graph reasoning algorithm (PRESCAL) based on path tensor factorization for learning and reasoning. Matt Gardner et al. proposed SFE [10] to do knowledge reasoning through the subgraph features. In the above work, the algorithm that used tensor factorization technique to learn the knowledge graph can only predict the directly linked entities. It doesn’t make full use of the graph structure feature of the knowledge graph. The SFE algorithm makes use of subgraph features to make reasoning, but its prediction effect is relatively unobtrusive. The PRESCAL algorithm makes use of relation path to learn reasoning and obtained better prediction. But it results in lower reasoning efficiency. Despite these emerged lots of knowledge graph learning and reasoning algorithms, it still lacks a reasoning algorithm that is efficient and highly accuracy, as well as a system that can construct and verify a political and economic news knowledge base. We can solve the problems of lack of completeness of the entity link of the initially constructed knowledge base and iterative updating of the knowledge base caused by data updating.

Under the above circumstances, this paper presents a system framework to construct and verify a domain knowledge base for political and economic news. The system applies the reasoning results to the construction and iterative update of domain knowledge base, which effectively addresses the problem of the incompleteness of the entity link of the initial knowledge base and the iterative updating of the knowledge base caused by the data update. First of all, we use political and economic news information as data sources to construct an initial domain knowledge base. Secondly, a fusion reasoning algorithm based on subgraph feature and tensor factorization, which we named as SFRESICAL, is presented to achieve reasoning in knowledge base. Finally, the fusion reasoning algorithm SFRESICAL is exploited to solve the self-improvement and iterative update problem of the knowledge base. As a case of engineering application, the system we build effectively solves the incompleteness of the entity link in the initial knowledge base and the iterative update problem. By putting this system framework into practice, this paper provides a comparatively complete, efficient knowledge base for the relevant institutions of political and economic news as well as experts and scholars, and thus provides a comprehensive and intelligent support for related businesses. The function of knowledge reasoning module and iteration module of knowledge base construction and autonomous updating system are verified.

The rest of this paper is organized as follows. Section 2 introduces the existing knowledge bases and algorithm for knowledge graph learning & reasoning; Section 3 illustrates the fusion reasoning algorithm SFRESICAL based on SFE feature in details. Then a knowledge-based construction and verify system based on the SFRESICAL is put forward in Section 4. Section 5 demonstrates the experiment we conducted to prove the effectiveness of the algorithm and system. Finally, we conclude in Section 6.

2. Related Work

In this section, we will give an overview of the existing knowledge base and then introduce the current algorithms for knowledge graph learning and reasoning.

The research of knowledge graph has gained high popularity after the Google-led internet companies put forward a series of application-oriented knowledge graph. Knowledge graph is actually a semantic network and a graph-based data structure composed of entities and relationships. In other words, a knowledge graph is a network of relationships that connect all the heterogeneous information together. Knowledge graph provides the ability to analyze problems from a "relationship" perspective. Researchers have built a variety of knowledge graphs or knowledge bases like Yago [27], DBpedia [2], Freebase [5], zhishi.me[22]. Google Knowledge Graph [4] was officially released on May 16, 2012 in the purpose of providing users with not only webpage list but other various content and returning the user's query information in the form of a structured list or knowledge card. Information in the Google Knowledge Graph comes from a variety of data sources and is mainly made up of the CIA World Factbook, Freebase, Wikipedia and etc. At the time of its publication, it already contained more than 57 billion objects, 1.8 billion triples, and various links between different objects. DBpedia [2] is a
community aimed at extracting structured multilingual knowledge from different language versions of Wikipedia. Its English version contains more than 3.7 million volumes while its Chinese version contains only about one million entities [24]. These knowledge bases play an important role in many applications, such as Q&A [8] and Recommendation Systems [29].

Although many approaches to construct general knowledge bases were put forward, there are few well-constructed and representative knowledge bases in politics and news. In addition, the high profession and variety of political and economic news result in the incompleteness of entity links in the initially constructed knowledge base, as well as high-frequency data update triggered by the knowledge base constantly iterative update needs. To tackle these problems, there is an urgent need to build a system that can be constructed and verified in political and economic news knowledge base. Based on the research and application of knowledge graph, we found that knowledge graph reasoning algorithm can be utilized to improve the drawbacks of knowledge base mentioned above.

With the research and application of knowledge graph, more knowledge graph learning and reasoning algorithms have appeared. Tensor is the generic term of high-dimensional arrays [13], and tensor factorization represents the process of decomposing high-dimensional arrays into multiple low-dimensional matrices. Some research applies tensor factorization to knowledge graph reasoning. For example, Nickel et al. [19,20] proposed the factorization of a three-way tensor RESCAL. RESCAL decomposes the tensor value of each triple in the knowledge graph into entity and relation matrix representation, and makes the product of the decomposed entity and the relation matrix as close to the original tensor value as possible. Applying the algorithm to general large-scale knowledge-based reasoning & learning for unknown relationships extraction and entity analysis obtains good results. On the basis of RESCAL, Nickel et al [18] proposed a logical probability based tensor factorization, which is an extension version of RESCAL to some degree. A tensor factorization in additional conditions was put forward by Nickel et al. [21] on the basis of the factorization of a three-way tensor and was applied to knowledge graph reasoning & learning. Socher et al [26] introduced Tensor Neural Network (Neural Tensor Networks, NTN) model and put forward a reasoning algorithm based on tensor neural network. Chang et al. [7] introduced TRESCAL, an algorithm that utilizes the constraints of relational semantic types during the construction of training models. Krompaß et al. [14] added the constraint processing of the relationship type, thus reducing computation in the factorization process. The above-mentioned researches mainly take advantage of tensor factorization technique for knowledge graph reasoning & learning. However, generally they can only predict the directly-connected entity relationship and thus poorly utilize the graphical structure features of the knowledge graph. Lao et al. [15] proposed a Path Ranking Algorithm (PRA), which uses random walk to obtain relationships between entities that may exist in the knowledge base. Neelakantan et al. [17] adopted a Recurrent Neural Network (RNN) Model to perform path inference and used PRA to obtain each entity relationship path. Lin et al [16] proposed Path-Based TransE (PTransE) for TransE [1] model, noting that not all relation paths are reliable, and thus designing path constraint allocation algorithm to measure the reliability of the relationship path. Yang et al. [30] took advantage of the neural network embedding model to establish a general reasoning model and perform path inference. Wang et al. [23] made use of incorporating rules into embedding models to conduct path reasoning in knowledge base and complete knowledge base. Gu et al. [11] combined the existing models for path reasoning, automatic Q&A and self-improvement of knowledge base learning tasks. Garcia-Duran et al. [9] considered the relational schema for knowledge graph based on TransE to do answer task and complete task. Matt Gardner et al proposed SFE [10] which conducts reasoning through subgraph features. However, the above-mentioned studies, adopting tensor factorization technique to learn the knowledge graph, can only predict the directly connected entities and thus can’t make full use of the graph structure of the knowledge graph. Although the SFE algorithm makes improvements by using subgraph feature to inference, it performs unsatisfactorily in prediction tasks. Moreover, the PRESCA algorithm obtains better prediction results in the cost of reasoning efficiency. Hence, there is still a demand to find a reasoning algorithm that is not only effective but also accurate.

Under these circumstances, this paper presents a system framework to construct and verify a domain knowledge base for political and economic news. The system applies the reasoning results to the construction and iterative update of domain knowledge base, which effectively addresses the problem of the incompleteness of the entity link of the initial knowledge base and the iterative updating of the knowledge base caused by the data update. First of all, we use political and economic news information as data sources to construct an initial domain knowledge base. Secondly, a fusion inference algorithm SFRESCAL based on subgraph feature and tensor factorization is presented to achieve fast reasoning in knowledge base. Finally, the fusion reasoning algorithm SFRESCAL is exploited to solve the self-improvement and iterative update problem of the knowledge base. The details about fusion algorithm SFRESCAL will be illustrated in Section 3.

3. Algorithm

To remove the inconvenience brought by the iterative update and incompleteness of the political and economic news Knowledge Base, we need to use the reasoning algorithm. Therefore, a syncretic reasoning algorithm SFRESCAL based on
Correlation feature and tensor factorization is proposed in this paper. Section 3.1 gives the definition of SFRESCAL model. While in Section 3.2 we present the description of SFRESCAL algorithm based on the model given in Sec.3.1.

### 3.1. Model Definition

In this Section we present the model definition of SFRESCAL. The initial Knowledge Base of Political & Economy News as Table 1 shown.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number</th>
<th>Detail</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories</td>
<td>9</td>
<td>Southeast Asian entities, East Asian entities and so on.</td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td></td>
<td>China entities, The United States entities and so on.</td>
<td></td>
</tr>
<tr>
<td>City</td>
<td></td>
<td>Beijing entities, New York entities and so on.</td>
<td></td>
</tr>
<tr>
<td>Expert</td>
<td></td>
<td>Martin Wolf entities, Lucy Kellaway entities and so on.</td>
<td></td>
</tr>
<tr>
<td>Politician</td>
<td></td>
<td>Donald John Trump entities, Hillary Diane Rodham Clinton entities and so on.</td>
<td></td>
</tr>
<tr>
<td>Domain</td>
<td></td>
<td>political entities, economic entities and so on.</td>
<td></td>
</tr>
<tr>
<td>Article</td>
<td></td>
<td>&quot;Spotlight: 23 EU members signal intent to launch common defense pact&quot; entities and so on.</td>
<td></td>
</tr>
<tr>
<td>Keyword</td>
<td></td>
<td>Infrastructure entities, Economic Cooperation entities and so on.</td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td></td>
<td>The Belt and Road entity and so on.</td>
<td></td>
</tr>
</tbody>
</table>

| Entities   | 2494685 | Southeast Asia, Philippines, Malaysia etc. |         |

For the convenience of representation, the definitions are given as following.

**Definition 1 (Knowledge Base of Political and Economy News).** Suppose that Knowledge Base of Political and Economy News is $G, T \subseteq G, T$ represents the $i$ triple of $G$; $S$ is triple set embedded to a low dimensional vector space by $T$ through embedding, $S_i = (h_i, r_i, t_i)$. $S_i$ represents the $i^{th}$ element in $S_i$, $h_i$ represents the head of the entity, represents the tail entity $t_i$. $r_i$ represents the relation of $h_i$ and $t_i$. $F_i$ is the relationship feature matrix of the entities $h$ and $t$ obtained through the feature extraction process of SFRESCAL, $f((h_i, r_i, t_i))$ is the model function based on SFE feature and tensor factorization.

**Definition 2 (Sets).** In the Knowledge Base of Political and Economy News $G$, Entity category is called set GC, entity is called set GE, relation is called set GR.

Combined with SFE algorithm [25], we present a method through which we can obtain the relation feature $F_{ij}$ of the entities $e_i$ and $e_j$ in the Knowledge Base of Political & Economy News, $e_i, e_j \in GE$.

Figure 1 shows the feature extraction process in SFRESCAL. We consider political and economic news Knowledge Base as a graph. For one of the entity pairs in the graph, we use the random walk algorithm to get the child graph involving source node and target node. Then, we merge two of these sub graphs into a resulting graph, in which we extract the relationship feature of the entity pair.
The specific process of feature extraction in SFRESCAL is shown as follows. SFE conducts random walk from each source node \( s_i \) (\( s_i \in GE \)) and target node \( t_j \) (\( t_j \in GE \)) of the knowledge base of political and economic news \( G \). Random walk is used to find a comparatively small set of potentially useful path types, for example, from node ExpertA to node ArticleB through path “ExpertPublishArticle”. After the first random walk, construct a binary eigenmatrix, which is 1 if accessible, 0 if inaccessible. For each source node or target node in data, we use k random walk to build a node-centered subgraph. For example, besides node ExpertA itself in the subgraph of node ExpertA, there may be node ArticleB, node DomainC. A random walk from the node \( n \) along the path \( \pi \) end up in some intermediate nodes \( i \). Keep all \((\pi, i)\) as features of subgraph \( G_n \). To build the feature vector of source - target pair \((s_i, t_j)\), we merge \( G_{s_i} \) and \( G_{t_j} \) to the intermediate node. This creates an edge type of feature space that connects the source node to the target node. We construct the eigenvectors simply by taking all the combined path types as binary features.

To sum up, we regard the knowledge base of political and economic news as a graph (regard entities as nodes and regard relations or properties as sides), from \( e_i \) and \( e_j \), execute k random walks on the graph, to construct the relation feature \( F_i \) between political and economic news. For example, take node ExpertA as the source node, the topic node TopicB as the target node, and k random walk is executed on the graph, build relation feature \( F_{ExpertA \ TopicB} \) and we can make judgment utilizing feature \( F_{ExpertA \ TopicB} \) to speculate if there is an “ExpertTalkAboutTopic” relationship between ExpertA and TopicB.

Combining the relation feature \( F \) extracted from the relation feature extracting process, we process the relation feature \( F \) into the factorization step of RESCAL [20] and get the algorithm SFRESCAL in this paper. The traditional RESCAL only takes into account the relation of simple direct nodes, and the SFRESCAL algorithm proposed by us has added more relation features, so it can effectively improve the accuracy of reasoning.

Suppose that in \( G \), the knowledge base of political and economic news, there is \((e_i, r_i, e_j)\). The tail entity of the first triple is the same as the head entity of the second triple, the relation path between \( e_i \) and \( e_j \) is \( r_i r_j \), i.e. the number of relations is 2. So, the tensor factorization function value of relation feature:

\[
f(e_i, r_i, e_j) = x^T_i (R_{ij} + F_{ij}) x_j
\]

Note: \( x_i \in R^d, x_j \in R^d, R_{ij} \in R^{d \times d}, k = 1, 2, \ldots, m \), \( R \) represents embedding entities and relations into low dimensional vector spaces, \( d \) represents dimension of vector spaces; \( x_i \) and \( x_j \) represent the vector matrix of entities \( e_i \) and \( e_j \), \( e_i, e_j \in GE \); \( R_{ij} \) and \( R_{ij} \) represent the matrix of relation of \( r_i \) and \( r_j \); \( r_i, r_j \in GR \); \( F_{ij} \) is the relation feature extracted from the entity pair \((e_i, e_j)\) through the feature extraction process of SFRESCAL.

In order to make the tensor factorization of relation feature more general and extensive, we utilize the relation feature of the entities we get above to extend our model. If there is a relation path \( r_i \cdots r_i \) between entity \( e_i \) and entity \( e_j \), the function value calculation of SFRESCAL can be changed into

\[
f(e_i, r_i, e_j) = x^T_i (R_{ij}^{(1)} R_{ij}^{(2)} \cdots R_{ij}^{(n)} + F_{ij}) x_j
\]

\( x_i \) and \( x_j \) represent vectors in low dimensional vector space of initial entity \( e_i \) and endpoint entity \( e_j \) in this path, and \( R_{ij}^{(i)} (i = 1, 2, \ldots, t, t \leq n) \) represents the matrix of relation of relation \( r_i \). According to the feature extraction process in this paper, the relation feature \( F_{ij} \) of entity \( e_i \) and entity \( e_j \) can be obtained. SFRESCAL demonstrate entity \( e_i \) reaches another entity through relation \( r_i \) by computing the product of entity vector \( x_i \) and \( R_{ij}^{(i)} \), the matrix of relation. The whole reasoning process repeats the above operations till reaching \( e_j \). Finally, the relational features are added and a new matrix is got. So, the upper form can be used for the tensor factorization of relation features.
To avoid the loss of robustness of SFRESCAL due to over fitting of the model in the training process, we need to revise and optimize SFRESCAL. The concrete correction and optimization is as follows:

\[
\min_{\{e_i, r_k, c_j\}} \sum_{i} \sum_{j} \sum_{k} \left| X_{ijk} - x^T_k R^{(F)}_k x_j \right|^2 + \lambda \left( \sum_{i} \left\| x_i \right\|^2 + \sum_{k} \left\| R^{(F)}_k \right\|^2 \right) \tag{3}
\]

\[\left| X_{ijk} - x^T_k R^{(F)}_k x_j \right|^2\] represents the loss function model of the whole tensor in the process of factorization, if \(\exists (e_i, r_k, c_j) \subseteq G \Rightarrow X_{ijk} = 1\), else \(X_{ijk} = 0\); 

\[x^T_k R^{(F)}_k x_j\] represents the factorization of matrix tensor on relation characteristic \(F\); 

\[\lambda \left( \sum_{i} \left\| x_i \right\|^2 + \sum_{k} \left\| R^{(F)}_k \right\|^2 \right)\] represents the modification processes introduced to avoid the model overfitting, \(\lambda\) is the corrected parameter, \(\lambda \geq 0\), the process of normalization of entities and relationships are displayed in bracket.

In the training process, to make the optimization model converge as soon as possible, we need to update the entity matrix \(E\) and the relational matrix \(R^{(F)}_k\). In this paper, alternating least squares are used to update the entity matrix and relational matrix, i.e. fixing \(R^{(F)}_k\), updating \(E\), then fixing \(E\), updating \(R^{(F)}_k\). Update is as follows [20]:

\[
E \leftarrow \left[ \sum_{i=1}^{m} X_i E R^{(F)}_k T + X^T_i E R^{(F)}_k \right] \cdot \left[ \sum_{i=1}^{m} R^{(F)}_k E^T E R^{(F)}_k + R^{(F)}_k E^T E R^{(F)}_k + \lambda I \right]^{-1} \tag{4}
\]

\[
R^{(F)}_k \leftarrow (Z^T Z + \lambda_k I)^{-1} Z^T \text{vec}(X_{ijk}) \tag{5}
\]

Note: \(Z = E \otimes E\), the whole update process continues until \(f(e_i, r_k, c_j) / \|X\|\) converges to a value or reach the maximum number of iterations.

3.2. Algorithm Description

As for the political and economic news knowledge base \(G\), the set of three-tuples is \(T = \{T_1, T_2, \ldots, T_m\}\). The embedding method is embedded into the low-dimensional vector space to obtain the training data set \(S = \{S_1, S_2, \ldots, S_m\}\). In the low-dimensional vector space, we can obtain entity relationship \(F\) through the relationship feature extraction process in 3.1, and then construct SFRESCAL model function \(f(h, r^{(F)}_k, t)\), which is finally used for link prediction.

The flow chart of the SFRESCAL algorithm is shown as Figure 2.
The Input in Figure 2 are the initial entity set $E$, the initial relationship set $R$, the set of three tuples $S$, the boundary adjustment parameters pose, the maximum number of iterations $T$, and the low dimensional space dimension $d$.

The initialized module [15] embeds entities and relationships into a d-dimensional vector space. The feature extraction module represents the relationship feature extraction process for finding the relationship feature $F$ of each entity pair.

Tensor factorization module shows us how use relational features $F$ for tensor factorization.

The update module indicates that the entity matrix and the relation matrix are updated using the alternating least squares method. Output data is the updated entity set $E$ and relation $R$.

The reasoning process in knowledge graph using SFRESCAL is as follows. The first step is to transform the entities and relations into vector matrix by embedding them into the d-dimensional vector space [15]. Then, in the vector space, the relation feature extraction is conducted to find the relational features $F$ of each entity pair and the SFRESCAL model is used to perform the tensor factorization to calculate the loss function value of the model. Finally, in the process of update iteration, we adopt the method of alternating least squares to update the entity matrix and the relation matrix respectively until the update converges to a certain value or reaches the maximum number of iterations. A specific algorithm is described in the following.

**Algorithm 1. SFRESCAL**

**Input:** the initial entity set $E$, the initial relationship set $R$, the set of three tuples $S$, the boundary adjustment parameters pose, the maximum number of iterations $T$, the low dimensional space dimension $d$.

**Output:** the updated entity matrix $E$, relation matrix $R$.

1. Initialize:
   
2. $e_i \leftarrow \text{Uniform}\left((-6)/\sqrt{d},6/\sqrt{d}\right)$ for each $e_i \in E$

3. $r_i \leftarrow \text{Uniform}\left((-6)/\sqrt{d},6/\sqrt{d}\right)$ for each $r_i \in R$

4. for each $e_i \in E$
   
5. Using the relational feature extraction process to calculate the features of each entity pair

6. Using the SFRESCAL model for tensor factorization, calculate the factorization loss function value:

7. $\sum \sum \sum \left\| X_{ij} - x_i^T \tilde{R}_i^{(F)} x_j \right\|^2 + \lambda \left( \sum \| x_i \|^2 + \sum \| \tilde{R}_i^{(F)} \|^2 \right)$

8. end for

9. while the number of iterations< $T$ or $f(e_i, r_i, e_j)/\|E\|^2$Less than a set value

10. update:

11. $E \leftarrow \left[ \sum_{k=1}^{n} X_k \tilde{R}_k^{(F)T} + X^T \tilde{R}_k^{(F)} \right] \cdot \left[ \sum_{k=1}^{n} \tilde{R}_k^{(F)T} \tilde{R}_k^{(F)} + \sum_{k=1}^{n} \tilde{R}_k^{(F)T} \tilde{R}_k^{(F)} + \lambda I \right]^{-1}$

12. $\tilde{R}_k^{(F)} \leftarrow \left( Z^T Z + \lambda \tilde{R}_k I \right)^{-1} Z^T \text{vec}(X_{ik})$

13. end while

The key point of the algorithm lies in taking advantage of relationship feature extraction to compute the relationship feature between each entity pair and the tensor factorization technique to calculate the value of the loss function for each relational feature. In this way, the new entity relationship in the knowledge base can be predicted and the self-improvement and iterative updating of the knowledge base can be realized.
4. Building Framework of Knowledge Base of Political & Economy

In this section, we use the SFRESCAL algorithm to build a political and economic news knowledge base (PENKB) capable of self-improvement in the light of the high profession political and economic news data. In addition, the knowledge reasoning step in the PENKB construction is illustrated in Chapter 3, and then Chapter 4 describes the whole process of PENKB construction.

As for knowledge base construction, it involves the data mining, knowledge extraction, knowledge disambiguation, knowledge reasoning and so on. The overall framework is shown in Figure 3. Among them, data sources mentioned in the data mining part include structured data in existing databases, semi-structured data and unstructured data, as well as general knowledge-based CN-DBpedia [4]. As the Figure 3 shows, the knowledge extraction module includes three sub-modules: entity extraction, relation extraction and attribute extraction. In this module, we mainly use HanLP [12] to process the current data and extract knowledge from data. The relationship extraction module adopts a weakly supervised relationship extraction technique for unstructured data [28]. Given a small number of existing relationship examples, this module acquires all kinds of labels and grammatical analysis results of entity words through grammatical analysis, and gets the entity feature labels from the knowledge base, so as to take advantage of a few examples to complete the labeling of corpus and extraction of the relationship. Then it combines the aggregated knowledge from CN-DBpedia or structured data to conduct disambiguation, where manual intervention and some expert rules are used. Furthermore, the expert rules are adopted for quality assessment and the processed knowledge is put into PENKB (PENKB is stored in graph database Neo4j which uses "graph" to model the data). As for knowledge reasoning, we utilize the SFRESCAL algorithm illustrated in Chapter 3. In the framework process, except the knowledge reasoning in the initial knowledge base, we conduct knowledge reasoning to update and improve the knowledge base according to the time interval.

Before building PENKB, we use knowledge graph related techniques to model political and economic news. We take the entity as the main target to achieve the data from different sources graphing and consolidation. Using attributes to express the descriptions of entity from different data sources forms a comprehensive description of the entity. Moreover, relationships are used to describe the relationships between various types of abstracted data that are modeled as entities, and thus support association analysis. Through the entity link technology, we can accomplish associated storage for various types of data around the entity. Event mechanism is used to describe the dynamic development of the objective world and reflect the connection between events and entities, while timing description is used for the development of the event. The event mentioned is defined as follows:

Definition 3 (Event). In PENKB, the political and economic news information repository, we define the events as 
\[ \text{event} = (\text{Timestamp}, \text{Entities}, \text{Relationship}, \text{Description}) \]

to describe the dynamic development of the objective world. \text{Timestamp} indicates when the events occur. \text{Entities} represents a set of events associated with the entity, \text{Relationship} represents the relation set of event correlation, \text{Description} shows a simple text description of the event.
Specific event data storage flow chart shown in Figure 4. When users publish an article, we can extract the event information from the article. The event information mainly includes the time of occurrence of the event Timestamp, the set of event-related entities Entities, the set of related relations Relationship of the event, the simple text of the event Description. Then, Entities and Relationship are stored in Neo4j's PENKB database, and then Relationship is updated according to the data in Neo4j. The extracted events are then stored in the EVENTS database in the TSDB. TSDB is a specialized database for managing time series data. Different from the traditional relational database, TSDB specially optimizes the storage, query and presentation of time series data to obtain high data compression ability and excellent query performance.

![Figure 4. Event data storage flow chart](image)

Based on the EVENTS database obtained, we can get a timing diagram of an event. For instance, as Figure 5 shows, attacks carried out in New York in recent years. According to the EVENTS database, we can analyze the history-related events of Event-B, the associated persons and the areas affected by Country-A events and provide them to field experts to intelligently support their professional statistics and analysis. Meanwhile, the EVENTS database also lays a solid foundation for the expansion of the event graph in the field of political and economic news.

![Figure 5. An example about timeline of attacks in New York city](image)

In conclusion, this paper describes the overall construction framework and event mechanism of political and economic news knowledge base named PENKB, where the knowledge reasoning module is implemented by the SFRESCAL algorithm stated in Chapter 3. First of all, the knowledge extraction module processes the existing structured and unstructured data and extract the knowledge contained in data. Then, the knowledge from the structured data and general
knowledge-based CN-DBpedia is aggregated and disambiguated. After the data goes into PENKB after quality assessment, we put the data into PENKB to and perform a knowledge reasoning on the original knowledge base to improve its body. Finally, we use the SFRESCAL algorithm to perform knowledge reasoning on PENKB with a specified frequency. The event mechanism of PENKB is based on the knowledge base constructed and stored in Neo4j. The article information is extracted to obtain the event information, and finally saved in EVENTS database of the sequence databases TSDB. The EVENTS database is used to analyze similar historical events, related people and affected areas, which is provided for domain an intelligent support for professional statistics and analysis. At the same time, the EVENTS database also provides a solid foundation for the expansion of the event graph in the political and economic news field.

5. Experiments

5.1. Experiments Dataset

The experiments dataset is based on the Political and Economic News Knowledge Base (PENKB). In order to show the ability of SFRESCAL improving the PENKB, we get PENKB-1 by further reasoning with SFRESCAL on PENKB. Table 2 shows the 72 relationships, 498936 entities, 2264810 number of training set and 180554 number of test set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Relation</th>
<th>#Entities</th>
<th>#Train</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>PENKB</td>
<td>36</td>
<td>249468</td>
<td>997712</td>
<td>78896</td>
</tr>
<tr>
<td>PENKB-1</td>
<td>36</td>
<td>249468</td>
<td>1267098</td>
<td>101658</td>
</tr>
</tbody>
</table>

5.2. Experiments Task and Evaluation Index

There are 2 experimental tasks being set to verify the validity of SFRESCAL: link prediction and path problem answer. Firstly, since most of the existing knowledge graph reasoning algorithm uses the accuracy of link prediction as a benchmark of predicting new triples, we execute link prediction experiments to compare SFRESCAL with the existing algorithm. Furthermore, the path problem answering shows the ability to reach middle entity set given the head entity and relationship path. This article compares the improving effects on PENKB before and after using SFRESCAL according to the path problem answer.

We use Mean Rank and HITS@10 [6] as the evaluation index, which generally adopt on Knowledge Graph Reasoning Algorithm.

Step1: Replacing the head or tail entity of a correct triple A with all other entities in the entire knowledge base in turn, which generates n triples.

Step2: Calculating the energy value of each from the n triples above, from which can we get n energy value corresponding to the n triples.

Step3: Sorting the above n energy values in ascending order.

Step4: Recording the sequence number of the energy values after sorting.

Step5: Repeating the above steps for all the correct triples.

Step6: Obtaining MeanRank by averaging the sequence number of correct triples sorted by energy values.

Step7: Obtaining mean value of the energy values by averaging the sequence number of correct triples sorted by energy values, which we called MeanRank.

HITS@10 calculates the triples score function value in test set and gets the first 10% values of the correct entities. The smaller Mean Rank, the larger HITS@10, the better predictive accuracy of the model.

5.3. Experimental setup

The following parameters are chosen in the experiment. Boundary adjustment parameter $\lambda$, dimension $d$ which embeds entities and relationships into low-dimensional spaces, the size parameter $B$ of training data being sent every batch during each iteration. The values are as follow:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>0.001, 0.01, 0.1, 0.25, 0.5, 1.0</td>
</tr>
<tr>
<td>$d$</td>
<td>20, 50, 80, 100</td>
</tr>
<tr>
<td>$B$</td>
<td>50, 100, 300, 500</td>
</tr>
</tbody>
</table>
Table 4. Optimal parameters for PENKB

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$d$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>20</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 5. Optimal parameters for PENKB-1

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$d$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>50</td>
<td>300</td>
</tr>
</tbody>
</table>

During the experiment, the best parameters of PENKB are shown in Table 4, the number of iteration $T = 300$. The best parameters of PENKB-1 are shown in Table 5, the number of iteration $T = 200$.

For showing the ability of SFRESCAL to entity link predict, the comparative experiment results of RESCAL, TRESCAL and TransE are shown in Figure 6 (a) and Figure 6 (b).

Figure 6 (a) shows the comparisons on MeanRank of SFRESCAL, RESCAL, TRESCAL and TransE on PENKB. SFRESCAL performs best since the value of SFRESCAL is the smallest. Figure 6 (b) shows the comparisons on HITS@10 of SFRESCAL, RESCAL, TRESCAL and TransE on PENKB. SFRESCAL still performs best since the value of SFRESCAL is the biggest.

On the dataset PENKB, the HITS@10 performance of SFRESCAL increases 41.2% , 18.2% and 10.5% comparing with RESCAL, TRESCAL and TransE respectively.

Then the experimental results of Figure 6 (a) and Figure 6 (b) are analyzed. Firstly, the dataset is composed of three tuples, where exists in some subgraph relations. While for RESCAL, TRESCAL and TransE, we only conduct simple inference to predict the three tuples and don’t make full use of the semantic relations between entities. Therefore, the prediction performance is worse than SFRESCAL.

In order to effectively illustrate the performance of SFRESCAL in improving the predictive performance of knowledge graph, we compare the answers to the path problems before and after the improvement of graph. The experimental results are shown in Table 6.

Table 6 shows the comparison of the SFRESCAL algorithm between the indicator Mean Rank and the indicator HITS @ 10 on PENKB and PENKB-1. The HITS @ 10 value of the SFRESCAL algorithm on PENKB-1 is improved by 7.8% compared to the HITS @ 10 value on PENKB. From Table 6, we can see that the PENKB-1 dataset improved by SFRESCAL algorithm performs better in answering path questions. The experiment indicates the algorithm attained by improving PENKB.

Table 6. Value of Mean Rank and HITS@10 on PENKB and PENKB-1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mean Rank</th>
<th>HITS@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PENKB</td>
<td>256</td>
<td>0.819</td>
</tr>
<tr>
<td>PENKB-1</td>
<td>203</td>
<td>0.897</td>
</tr>
</tbody>
</table>
In summary, compared with some existing reasoning algorithms, SFRESCAL performs better in prediction function in entity link task and thus is suitable for enriching and enlarging knowledge graphs. The good results of SFRESCAL showing on path answering tasks prove its effectiveness in enhancing knowledge base called PENKB.

6. Conclusions

The construction technology of knowledge bases is mainly supported by multidisciplinary cross technology including data mining, machine learning, natural language processing and information retrieval. And the combination of qualitative and quantitative analysis, practice and calculation are the basis of their research approaches. Regarding the characteristics of political and economic news data, this paper put forward a novel system framework for political and economic news knowledge base. The system applies inference results to the construction and update of the knowledge base and thus, effectively addresses the problems including incompleteness of entity links and data iterative update. Besides, a mixed algorithm SFRESCAL based on subgraph feature and RESCAL are introduced to achieve the self-improvement of knowledge base. To prove the effectiveness of the system, we constructed a political and economic news knowledge base named PENKB. Using specific data of political and economic news, the system we built effectively solves the incompleteness of the entity link in the initial knowledge base and the iterative update problem. By designing and implementing knowledge reasoning, updating knowledge iteration, the function of knowledge reasoning module and iteration module of knowledge base construction and autonomous updating system are verified. The experimental results show that SFRESCAL performs better than some reasoning algorithms in the task of link prediction in PENKB. Simultaneously, in the knowledge base path question answering, PENKB improved by SFRESCAL. Being the core technology of the next generation search engine, knowledge graph is worthy of theoretical research and practical application. Domain knowledge bases will play an increasingly important role in applications like intelligent linkages of political and economic news, implicit information detection and Q&A. Thus, we hope the paper can provide some reference to political and economic news management in network information system.

Acknowledgements

This paper is supported by Foundation of Science and Technology Commission of Shanghai Municipality under Grant No. 14590500500, the Innovation Project of Next Generation Internet Technology of CERNET under Grant No. NGII20160210, No. NGII20160614 and No.NGII20160325.

References

12. HanLP, https://github.com/hancks/HanLP
15. N. Lao, M. Tom, and W. C. William, “Random walk reasoning and learning in a large scale knowledge base,” in proceedings of
Pin Wu, Juanjuan Luo, Yonghua Zhu, and Wenjie Zhang


17. A. Neelakantan, R. Benjamin, and M.C. Andrew, "Compositional vector space models for knowledge base reasoning," in proceedings of 2015 aaai spring symposium series, pp. 156–166, Austin, Texas, USA, January 2015


