Article

Construction and Recommendation of a Water Affair Knowledge Graph

Jianzhao Yan *, Tiantian Lv and Yongchuan Yu

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
1543368448@emails.bjut.edu.cn (T.L.); yuyongchuan@bjut.edu.cn (Y.Y.)

* Correspondence: yanjianzhao@bjut.edu.cn; Tel.: +86-188-1131-1569

Received: 21 July 2018; Accepted: 19 September 2018; Published: 26 September 2018

Abstract: Water affair data mainly consists of structured data and unstructured data, and the storage methods of data are diverse and heterogeneous. To meet the needs of water affair information integration, a method of constructing a knowledge graph using a combination of water affair structured and unstructured data is proposed. To meet the needs of a water information search, an information recommendation system for constructing a water affair knowledge graph is proposed. In this paper, the edit distance algorithm and latent Dirichlet allocation (LDA) algorithm are used to construct a water affair structured data and unstructured data combination knowledge graph, and this graph is validated based on the semantic distance algorithm. Finally, this paper uses the recall rate, accuracy rate, and F comprehensive results to compare the algorithms. The evaluation results of the edit distance algorithm and the LDA algorithm exceed 90%, which is greater than the comparison algorithm, thus confirming the validity and accuracy of the construction of a water affair knowledge graph. Furthermore, a set of water affair verification sets is used to verify the recommendation method, which proves the effectiveness of the recommended method.

Keywords: structured data; unstructured data; water affair; knowledge graph construction; information recommendation

1. Introduction

With the development of water affair information, water affair data have the problems of multisource heterogeneity and a large quantity. There is a large amount of structured data in the database and a variety of unstructured data, such as text, images, and video, and the data storage locations are diverse. However, successful knowledge management relies on the construction and application of an effective knowledge graph for the formation and description of knowledge in specific areas, enabling the water affair industry to achieve an effective management of domain knowledge.

With the development of the Internet, network data content has undergone an explosive growth trend. Due to its large scale, substantial heterogeneity, and the loose organizational structure of Internet content, people find it challenging to meet the needs of accurate and efficient access to information and knowledge. The knowledge graph [1] came into being and, with its strong semantic processing and open organization, laid a suitable foundation for knowledge organization and an intelligent application of the Internet era. A knowledge graph is essentially a semantic network; it is a graph-based data structure that consists of nodes and edges. Each node is an objective “entity” that exists in the world, and each edge marks a “relationship” between two entities. The knowledge graph is a large relational network that connects all kinds of different information and integrates them. In contrast to the previous ontology [2], which focused on the upper and lower relations, the knowledge graph focuses on the semantic relationship. To put it simply, when the semantic relationship is merged

into the ontology, a knowledge graph is formed. Thus, ontology is the conceptual model and logical basis of the knowledge graph [3]. The hierarchy of a knowledge graph can be divided into two levels: a data layer and a pattern layer. The pattern layer of the knowledge graph is similar to the relationship and structure between concepts in the ontology. In the data layer of the knowledge graph, the storage is in various forms, and there is a graph database, which is stored in facts, as with Google's Graphd and Trinity, which are typical graph databases. Data are also stored according to the resource description framework (RDF) triple, which is the basic expression of the fact according to the format of the "entity-relationship-entity" or "entity-attribute-attribute value" triplet. The entity relationship network formed by putting all of the data of the data layer and the structural relationship of the pattern layer together is the knowledge graph.

Traditional methods of knowledge graph construction include the skeleton method, the Toronto Virtual Enterprise Ontology Project (TOVE) method, the Methontology method, and the seven-step method. The skeleton method was proposed by Uschold [4] based on the research and development experience of the enterprise ontology model. This method mainly includes the following processes: determining the domain and scope of the ontology application, analyzing and evaluating the target ontology, encoding the ontology and integrating with the ontology, evaluating the ontology, and assembling the document. The TOVE method [5] was proposed by the University of Toronto Enterprise Integration Lab when constructing the Toronto Virtual Enterprise Ontology Project. This ontology includes the ontology of enterprise design, engineering, plan, and service, and uses first-order logic for integration. The Methontology method [6] uses the concept of the ontology life cycle to organize the development process of the entire ontology. This method is divided into three phases: the management phase, the development phase, and the maintenance phase. The seven-step method [7] is a classical method of ontology construction proposed by researchers at Stanford University. The seven-step method is a relatively logical method that first advocates the reuse of the existing ontology as much as possible, extracting the relationship between the concepts by a step-by-step method; the method then constructs the generic relationship between concepts, examining the relationship between concepts, and ultimately fills in the instance object of the concept.

At present, the construction methods and applications of various domain knowledge graphs have been developed. Chen et al. [8] proposed a domain ontology construction model and a personalized knowledge search and recommendation system to enhance knowledge integration and application. Sulthana [9] built an ontology with a neuro-fuzzy classification and evaluated the context and determined comments based on the ontology and context recommendation system. The proposed method was reported to improve the accuracy of the recommendation system. To solve the ambiguity problem, Run et al. [10] proposed a financial news recommendation algorithm to help users find interesting articles. Uma et al. [11] built a job recommendation system for job seekers by collecting work portal data. Castells et al. [12] presented a comprehensive personalized retrieval framework where the advantages of ontologies are exploited in different parts of the retrieval cycle. Based on network encyclopedia resources, Kramer [13] extracted more than two million conceptual concept networks from Wikipedia and mapped these concepts to more than three million terms. Shinzato and Torisawa [14] proposed a method for automatically acquiring a relationship-assisted knowledge graph from HTML documents through the shortcomings of traditional methods. KnowItAll [15] and Nell [16] used iterative methods to learn high-quality triples from web page data to construct a knowledge graph.

However, these studies are not fully applicable to the construction and application of a water knowledge graph. At present, the monitoring data of water structure mainly exists in the Oracle database and composes millions of data volumes arranged in a complex structure. Unstructured data mainly include data sources such as text, images, and video. Based on the current distribution of water data and the demand for integrated data by water users, it is necessary to develop a model that can integrate a large number of multisource heterogeneous data points and apply them.

In this paper, for the user's demand for water structural data and unstructured data integration, first, a model of water structured monitoring data and water unstructured text data is constructed to construct a knowledge graph. Then, an information recommendation system based on the water knowledge graph is constructed. Finally, according to the knowledge graph construction method, the recall rate, accuracy rate, and F comprehensive results are compared to evaluate the effectiveness and accuracy of the proposed method. At the same time, a set of verification data sets is used to verify the effectiveness of the information recommendation system based on the water knowledge graph.

The rest of this paper is organized as follows. Section 2 presents the construction and application model of the water knowledge graph. Section 3 proposes the construction of the water knowledge graph and the method of information recommendation. Section 4 introduces the results of the construction and recommendation of the water knowledge graph and carries out the evaluation verification experiment. Section 5 presents the conclusions and provides guidance for future work.

2. Materials and Methods

To realize the construction and recommendation of the water knowledge graph, this section establishes a personalized adaptation model of the water affair knowledge graph, as shown in Figure 1. The model is mainly divided into four parts, namely, the data source layer, knowledge graph layer, reasoning layer, and application layer. Each part is discussed as follows. Data source layer: Wordnet dictionaries, Dbpedia thesaurus, water industry standards, and water expert experience are used to construct the top-level knowledge graph of water affairs. Structured monitoring data and unstructured text data are used to supplement the knowledge graph to form the final natural domain knowledge graph of water affairs. The knowledge graph layer mainly includes the construction of a top-level knowledge graph in the natural field of water affairs and the refinement of the knowledge graph in the natural fields of water affairs. The analytical layer analyzes the water affair knowledge graph and formulates an effective method of recommendation. Finally, the application layer mainly includes recommendations based on the knowledge graph of the water affair field.

Figure 1. Water affair knowledge graph model technology framework.

The technical model is shown in Figure 2, where the red dashed box is the technical route of the data layer, the green dashed box is the technical route of the knowledge map construction layer, and the black dashed box is the technical route of the analysis layer and application layer. First, the Protégé (Stanford University, Stanford, CA, USA) ontology construction software is used to build a top-level conceptual model, and the conceptual model of the top knowledge graph of water affair is formed based on the relationship established between the concepts. Second, the D2RQ tool is used to extract

the structured database to form the RDF triple data, and the text extraction tools (such as POI or jieba) are used to extract the unstructured text data to form the RDF triple data. In addition, the two sets of RDF triplet data are linked to the water affair top knowledge graph conceptual model to form a water affair knowledge graph. Finally, this paper creates a water affair information recommendation based on the methods of the knowledge graph.

Figure 2. Technical model.

2.1. Materials and Tools for the Construction of a Water Affair Knowledge Graph

2.1.1. Construction of the Top-Level Knowledge Graph of Water Affairs

The water affair knowledge graph construction process includes the construction of the top-level knowledge graph, extraction of the database, extraction of the text, and attachment of the knowledge graph. The construction of the top-level knowledge graph is a key link that directly affects the quality of the entire knowledge graph. This research mainly constructs the top-level knowledge graph of water resources based on the Wordnet [17] dictionary, Dbpedia [18] thesaurus, water industry standards, and water expert experience; the examples of their corresponding natural concepts and examples referred to herein are shown in Table 1. The construction of the top-level knowledge graph of water affairs is completed in the Protégé tool [19]. The content of the top-level knowledge graph of the water business constructed in this paper is shown in Table 2, which mainly includes the concepts of water affairs and the hierarchical structure between the concepts.

Table 1. Instance and adoption of Wordnet, Dbpedia, water industry standards, and water expert experience.

Source	Concept	Instance	Adoption
Wordnet	natural	in accordance with nature; relating to or concerning nature; "a very natural development"; "our natural environment"; "natural science"; "natural resources"; "natural cliffs"; "natural phenomena"	natural resources
Dbpedia	natural	natural attraction; water body; lake; flowing water body, well, spring, river	lake, well, spring, river
water industry standards	natural	surface class; underground class	surface class; underground class
water expert experience	natural	surface class (surface water channel, water storage unit); underground class (underground water conveyance)	surface class (surface water channel, water storage unit); underground class (underground water conveyance)

Table 2. Concept and structure of the top knowledge graph in the water affair sector.

First Level Concept	Second Level Concept	Third Level Concept	Fourth Level Concept
natural	surface class	surface water channel water storage unit	river, spring lake
	underground class	underground water conveyance	well

2.1.2. Extraction Tools for Structured Monitoring Data

D2RQ [20] is a tool used for a large number of structured data extractions; it extracts structured data and transforms them into RDF files for constructing a knowledge graph on the protégé platform. Due to the huge amount of water affair monitoring data, ordinary extraction methods cannot achieve large-scale structured data extraction. To achieve an effective extraction of structured data, this paper applies the D2RQ tool for extracting a large number of water structured monitoring data. The D2RQ language is implemented in the Table 3.

Table 3. D2RQ automatic extraction language.

D2RQ Automatic Extraction Language
generate-mapping -u sw -p sw -d oracle.jdbc.OracleDriver -o sw.ttl dump-rdf -f RDF/XML -o sw.rdf sw.ttl

2.1.3. Parsing Tools of Texts

POI [21] is a tool for parsing text that can screen out nontext content such as tables and images in text, leaving only the contents of the text. Jieba [22] is an open-source Python Chinese word segmentation tool that is divided into three modes: precise mode (default), full mode, and search engine mode. Jieba is more accurate than other commonly used open-source word segmentation tools (such as mmseg4j). Due to the large granularity of the experimental text concepts, this paper uses the precise mode of jieba to conduct our experiments. CN-Dbpedia [23], a large-scale general-domain structured wiki developed and maintained by the Knowledge Factory laboratory of Fudan University, covers tens of millions of entities and hundreds of millions of relations. Jena has a Java API. As an ontology parsing tool, Jena converts the extracted text information into the RDF text format, which is used to realize ontology visualization in Protégé. This article uses these tools to parse water texts.

For the first time, CN-Dbpedia is applied to water data to improve the water affair knowledge graph. In addition, the combination of these tools greatly improves the efficiency and integrity of the construction of a water affair knowledge graph.

2.2. Construction Algorithms for a Water Affair Knowledge Graph

The mapping algorithm of a knowledge graph is the key to its construction. The algorithm maps the extracted table names of structured data tables, the content of unstructured texts, and the top-level knowledge graph of the constructed water service. Then, these data are integrated with the corresponding concepts to supplement the top-level knowledge graph of the water service with instances and attributes and attribute values. The mapping algorithm of the water affair knowledge graph is divided into an edit distance algorithm applied to the mapping of structured monitoring data and a latent Dirichlet allocation (LDA) text classification method used to map the unstructured text.

2.2.1. Edit Distance Algorithm

The edit distance is the minimum number of edit operations required to switch from one string to another. If the strings are more distant, then they are considered more different. Permissive editing operations include replacing one character with another, inserting one character, and deleting one character. The algorithm mainly adds the content of step1 on the basis of the original edit distance

algorithm [24], and this paper combines regularization with the original edit distance algorithm to realize the construction of the water affair structured data knowledge graph. The algorithmic process is as follows:

Step 1: First, the table name is formatted by regularization and conversion of all letters to lowercase, and the concept words in the table name are extracted;

Step 2: For two concept words to be compared, set one of them to be the source string s with length n and the other as target string t with length m . Based on these two strings, this paper constructs a matrix named d $[m+1, n+1]$ and initializes the first row of the matrix to $0, 1, 2, \dots, n$, and the first column is initialized to $0, 1, 2, \dots, m$;

Step 3: Compare each pair of characters in s (i from 1 to n) and t (j from 1 to m);

Step 4: If $s[i] = t[j]$, the edit cost $\text{cost} = 0$; if $s[i] \neq t[j]$, the edit cost $\text{cost} = 1$;

Step 5: Set the value of the cell $d[i, j]$ in the matrix by the following algorithm:

- a. The value of the cell immediately above is incremented by 1, which is $d[i-1, j] + 1$;
- b. The value of the left cell is incremented by 1, which is $d[i, j-1] + 1$;
- c. The value of the diagonal cell is increased by the value of the edit cost, which is $d[i-1, j-1] + \text{cost}$.

Step 6: Iterating the second, third, and fourth steps, $d[m, n]$ is the value of the last edit distance of the two concept words being compared. Then, the similarity between the two strings s and t is

$$\text{Sim}(s, t) = 1 - \frac{d[i, j]}{\max(m, n)} \quad (1)$$

2.2.2. LDA Text Classification Algorithm

The LDA topic model is a three-layer Bayesian production probability model proposed by D. M. Blei et al. in 2003. The model assumes that the text is a random mixture of a series of potential themes and that the topic is a mixture of all of the words in the vocabulary. The main difference between different texts is the different assortment of topics. The model implements the probability distribution at the document-subject level through the Dirichlet function. Documents are seen as a set of probabilistic topics that are combined with each other, and words have probabilities assigned to each topic. The algorithm improves the application of the original algorithm [25] in step 4. In step 4, the frequency P of the occurrence of the i -th word in the document d is expressed as the similarity between the text *text* and the concept c . The specific generation process is as follows:

Step 1: The word is the basic unit of text data and is a subitem of a word list indexed with $\{1, 2, \dots, V\}$. The v th word in the vocabulary is represented by a V -dimensional vector w , where for any $u \neq v$, $w_v = 1$, and $w_u = 0$.

Step 2: A document is a sequence of N words, which is denoted by $d = \{w_1, w_2, \dots, w_n\}$, where w_n is the n th word in the sequence.

Step 3: A document set is a collection of M documents expressed as $D = \{d_1, d_2, \dots, d_M\}$. Assuming that there are k topics, the probability of the i th word w_i in document d can be expressed as follows:

$$P(w_i) = \sum_{j=1}^T P(w_i | z_i = j) P(z_i = j) \quad (2)$$

where z_i is the latent variable, indicating that the i th word w_i is taken from this topic. $P(w_i | z_i = j)$ is the probability that the word w_i belongs to the subject j , and $P(z_i = j)$ gives the probability that the document d belongs to the subject j .

Step 4: The result $P(w_i)$ in (3) is used as the similarity between the document text and the concept c in the water affair knowledge graph:

$$\text{Sim}(\text{text}, c) = \sum_{j=1}^T P(w_i | z_i = j) P(z_i = j) \quad (3)$$

The j th topic is expressed as a polynomial distribution $\varphi_{w_i}^j = P(w_i | z_i = j)$ of V words in the vocabulary, and the text is represented as a random mixture $\theta_j^d = P(z_i = j)$ on K implicit topics, so the probability of the vocabulary w “occurring” in the text d is as follows:

$$P(w|d) = \sum_{j=1}^T \varphi_w^j \times \theta_j^d \quad (4)$$

This paper finds the maximum likelihood function by EM (expectation maximization algorithm):

$$l(\alpha, \beta) = \sum_{i=0}^M \log p(d_i | \alpha, \beta) \quad (5)$$

The maximum likelihood estimators of Equation (5) are α and β , and the parameter values of α and β are estimated to determine the LDA model. The conditional probability distribution of where the text d “occurs” is as follows:

$$P(d | \alpha, \beta) = \frac{\Gamma(\sum_i a_i)}{\prod_i \Gamma(a_i)} \int \left(\prod_{i=1}^k \theta_i^{a_i-1} \right) \left(\sum_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \quad (6)$$

There are θ, β pairings, and the analytical formula cannot be calculated; thus, an approximate solution needs to be obtained. In the LDA model, an approximated inference algorithm such as Laplace approximation, variational inference, Gibbs sampling, or expectation propagation can be used to obtain the estimated parameter values.

2.2.3. Water Affair Knowledge Graph Recommendation Algorithm Based on the Semantic Distance

The “semantic distance” [26] is a quantitative expression of the strength of the relationship between concepts. The definition of semantic similarity is based on the length of the path between concepts, which determines the degree of semantic similarity. The semantic distance and semantic similarity are different representations of the same relational features of a pair of concepts. If the semantic distance between two concepts is closer, then the concepts are considered more similar. The semantic similarity of concepts is related not only to the distance between concepts, but also to the depth of the concept in the knowledge graph. Considering these factors comprehensively, this paper proposes a water affair knowledge graph recommendation algorithm based on the semantic distance. The specific definition is as follows:

Step 1: For the two concepts C_1 and C_2 , if the semantic similarity is $\text{Sim}(C_1, C_2)$ and the semantic distance is $\text{Dis}(C_1, C_2)$, then

$$\text{Sim}(C_1, C_2) = \frac{\alpha}{\alpha + \text{Dis}(C_1, C_2)} \quad (7)$$

where α is an adjustable parameter that represents the semantic distance value when the similarity is 0.5.

Step 2: Introduce the hierarchical depth of the node:

$$\text{Sim}(C_1, C_2) = \frac{\alpha \times \min(\text{depth}_{C_1}, \text{depth}_{C_2})}{\alpha \times \min(\text{depth}_{C_1}, \text{depth}_{C_2}) + \text{Dis}(C_1, C_2)} \quad (8)$$

where $\min(\text{depth}_{C_1}, \text{depth}_{C_2})$ is the minimum depth of C_1 and C_2 . Thus, in the case where the path distances are the same, the nodes having deeper levels have higher similarities.

Step 3: Following reference [27], this study set the parameter $\alpha = 1.6$ and recommends the water affair knowledge graph area with a $\text{Sim}(C_1, C_2)$ value greater than 0.6.

The rules that the semantic similarity should follow are as follows: the range of the semantic similarity value $\text{Sim}(C_1, C_2)$ of the concepts C_1 and C_2 is $[0, 1]$. When the two concepts are identical, their semantic similarity is 1; when the two concepts are completely different, their semantic similarity is close to zero.

3. Results

3.1. Implementation Environment

In this paper, the top-level knowledge graph of waterworks was implemented in the Protégé software, and the D2RQ tool was used to extract water monitoring data. Water documents were extracted using the POI tool to read the word, and jieba tools were used for text segmentation. The last part of the construction algorithm and the recommendation algorithm were implemented in python.

3.2. Experimental Data Preparation

3.2.1. Preparation of the Water Affair Knowledge Graph

The construction of a knowledge graph in the natural field of water affair involves the application of 10 water-related structured monitoring data tables from an Oracle database and 10 unstructured texts (including rivers and lakes, for example), as shown in Tables 4 and 5. The construction also involves a top-level knowledge graph of the water affair constructed in Section 2. The concept in the knowledge graph is based on the concept of extraction and the concept information of the upper level in CN-Dbpedia, as shown in Table 6.

Table 4. Experimental table information.

No.	Name	Type	Explanation
1	ST_RIVER_R	table	river water
2	ST_RIVER1_R	table	Beijing river water table
3	LK_LAKE_B	table	lake general information table
4	ST_DAM_R	table	gate dam water situation table
5	WRS_ZBWELL_P	table	self-prepared well water consumption
6	WRS_SYWELL_P	table	water source well water monthly report
7	ATT_SPRING_B	table	spring head basic situation table
8	ATT_RIVER_B	table	river basic situation table
9	ATT_LAKE_B	table	basic characteristics of the lake
10	ATT_RLWELLBASE_S	table	human well statistics

Table 5. Experimental text information.

No.	Name	Type	Explanation
1	text1	text	Beijing river information document
2	text2	text	north canal profile document
3	text3	text	Beijing lake information document
4	text4	text	lake feature document
5	text5	text	Beijing suburban well document
6	text6	text	well category document
7	text7	text	spring basic information document
8	text8	text	Beijing rivers and lakes document
9	text9	text	northern canal feature document
10	text10	text	business configuration document

Table 6. The concept of extraction and the upper concept information in CN-Dbpedia.

Concept	CN-Dbpedia Upper Layer Concept1	CN-Dbpedia Upper Layer Concept2
river	hydrology and water	natural resources
lake	water storage unit	topography (natural resources)
well	underground class	natural resources
spring	surface water channel	natural resources

3.2.2. Experimental Data Based on Information in the Water Affair Knowledge Graph

The information recommendation experiment based on the water affair knowledge graph uses the example of the river, its concept, and the upper concept. The water affair structured monitoring data tables include the following three tables: ST_RIVER_R, ST_RIVER1_R, WRS_YSLQKXJZBJM_P, whose specific explanation is shown in Table 4. Among them, ST_RIVER_R, ST_RIVER1_R is a table related to the river concept, and WRS_YSLQKXJZBJM_P is almost irrelevant to the river concept. The water affair documents select text1, text2, and text10, whose specific explanation is shown in Table 5, where text1 and text2 are related to the river concept and text10 is almost irrelevant to the river concept.

3.3. Experimental Results

3.3.1. Construction Results of the Water Affair Knowledge Graph

This section uses the top-level knowledge graph of water affairs in Section 2.1.1 and the water affair structured monitoring data and unstructured text in Section 3.2.1 to complete the construction of a water affair knowledge graph. In this section, formula (1) is used to complete the construction of the structured data knowledge graph, and formula (3) is used to complete the construction of the unstructured text knowledge graph. The process calculates the similarity of the ten water database tables and the corresponding concepts of the ten water affair texts and knowledge graph; the process then performs the corresponding mounting. The calculation results are shown in Table 7, and the completed water affair knowledge graph is shown in Figure 3.

Table 7. Similarity between table/text and concepts.

Sim (Table/Text, Concept)	River	Lake	Well	Spring
ST_RIVER_R	1	0.2	0	0
ST_RIVER1_R	0.83	0.17	0.17	0
LK_LAKE_B	0.2	1	0	0
ST_DAM_R	0	0.25	0	0
WRS_ZBWELL_P	0.17	0.17	0.67	0
WRS_SYWELL_P	0.17	0.17	0.67	0.17
ATT_SPRING_B	0.17	0	0	1
ATT_RIVER_B	1	0.2	0	0
ATT_LAKE_B	0.2	1	0	0
ATT_RLWELLBASE_S	0.2	0.3	0.4	0
text1	0.85	0.23	0.1	0.35
text2	0.78	0.1	0.1	0.05
text3	0.1	0.8	0.2	0.05
text4	0.21	0.78	0.3	0.1
text5	0.2	0	0.8	0
text6	0.1	0	0.85	0
text7	0.1	0.4	0	0.95
text8	0.8	0.12	0	0
text9	0.68	0.1	0	0
text10	0	0	0	0

Figure 3. Water affair knowledge graph.

The water affair natural field knowledge graph contains the information of the knowledge graph at the top level of the water affair and the conceptual information of the structured monitoring data and the unstructured text. This graph is the basis for the water affair information recommendation.

3.3.2. Water Affair Information Recommendation Results Based on the Water Affair Knowledge Graph

To recommend river information, this section uses the three water monitoring data sheets and three water documents described in Section 3.2.2. From Equation (8), the final similarity between the river and three water monitoring data sheets and the three water documents are available, as shown in Figure 4.

Figure 4. Similarity results.

The results of the final recommendation algorithm indicate that table ST_RIVER_R and table ST_RIVER1_R have the greatest correlation with the river concept, exceeding 90%. The correlation between text1, text2, and the river concept follows, exceeding 80%. The correlation between table WRS_YSLQKXJZBJM_P and text10 and the river concept is the smallest, which with similarity values at 30% between WRS_YSLQKXJZBJM_P and text10, and 20% between text10 and the river concept. These results are consistent with our expectations. The recommended order is arranged according to the similarity value from largest to smallest. Furthermore, results having a similarity of less than 60% are removed; the final results are shown in Table 8.

Table 8. The recommendation sequence of river information.

Name	Type	Explanation	Similarity (%)
ST_RIVER_R	table	River regime	95
ST_RIVER1_R	table	Beijing river water	90
text1	text	Beijing river system	85
text2	text	North canal overview	80

3.4. Method Assessment

3.4.1. Assessment of the Knowledge Graph Construction Results in Water Affairs

This paper evaluates the edit distance algorithm in water structured data mapping and the LDA text classification algorithm in water text mapping. The edit distance algorithm, Jaccard coefficient algorithm, and Euclidean algorithm are compared in terms of the recall rate, accuracy, and F synthesis results, and by using LDA and TF-IDF algorithms and LSI algorithms. The algorithms are as follows:

$$\text{Precision rate : } P = \frac{TP}{FP} \times 100\% \quad (9)$$

$$\text{Recall rate : } R = \frac{TP}{TP + FN} \times 100\% \quad (10)$$

$$\text{Harmonic mean of exact and recall : } F - \text{measure} = \frac{2 \times P \times R}{P + R} \times 100\% \quad (11)$$

where TP is the correct number of matches, FP is all matches found by the method, and FN is the number of matches recommended by the method.

As determined by formulas (9)–(11), the comparisons between the edit distance and the Jaccard coefficient algorithm and the Euclidean metric algorithm are shown in Figure 5. The comparisons between the LDA model algorithm and the LSI model algorithm and the TF-IDF model algorithm are shown in Figure 6.

Figure 5. Mapping algorithm evaluation of structured data.

Figure 6. Mapping algorithm evaluation of the unstructured text.

Figure 5 shows that the accuracy P, recall rate R, and F-measure value of the edit distance algorithm are all above 90%, which exceeds the results for the Jaccard algorithm and the Euclidean metric algorithm because structured tables are named with concept words and because the edit distance

method can effectively improve the efficiency of conceptual noun matching. In addition, Figure 6 show that the recall rate R of the LDA model algorithm is lower than that of the LSI model algorithm, which occurs because the LSI model more fully considers the synonym problem, so the value of the recall rate R is increased. However, because the LDA model contains three layers of words, topics, and documents, the text is analyzed more precisely. The evaluation results also indicate that the accuracy P and F -measure value of the LDA model are both above 90%, which is superior to the results for the LSI model algorithm and the TF-IDF model algorithm.

3.4.2. Evaluation of a Water Affair Knowledge Graph Information Recommendation Method Based on the Semantic Distance

From Section 3.3.2, this paper obtains the recommendation results of river by the water affair knowledge graph recommendation method based on the semantic distance. To prove the reliability of this method, this section uses the lake-concept-related tables LK_LAKE_B and ATT_LAKE_B and the lake-related information documents text3, text4, and text8. Furthermore, this section uses the tables ST_RIVER_R, ST_RIVER1_R, ATT_SPRING_B, ST_DAM_R, WRS_ZBWELL_P, WRS_SYWELL_P, ATT_SPRING_B, ATT_RIVER_B, and ATT_RLWELLBASE_S, and the texts text1, text2, text5, text6, text7, text9, and text10; both the tables and the texts are not related to the lake for verification. The specific explanation information is shown in Tables 4 and 5.

This paper uses the water distance knowledge map recommendation method based on the semantic distance to recommend the lake-concept-related information. The similarity result is shown in Figure 7.

Figure 7. Results of the lake concept recommendation information.

The verification results indicate that the similarity between the lake-related database tables and the lake concept and the similarity between the lake-related documents and the lake concept exceed 75%. Furthermore, the similarity between the database tables that is not related to the lake and the lake concept is no more than 15%. These results demonstrate that the water affair knowledge graph recommendation algorithm based on the semantic distance, which was proposed in Section 2.2.3 of this paper, is valid.

4. Discussion

This paper proposes a knowledge graph construction method for combining water affair structured and unstructured data. This method is applicable to all data construction knowledge graphs similar to water affair structured and unstructured data, and the method has universality. Moreover, the recommendation system based on the knowledge graph can improve the efficiency

of water affair information searches and provide a certain reference value for the accuracy of the construction of a water affair knowledge graph.

This paper proposes a method for constructing a knowledge graph of water affairs in the natural field by constructing a combined knowledge graph based on structured monitoring data in the water affair and unstructured text information. To realize the recommendation of water affair information, this paper also proposes a personalized recommendation method based on the natural field knowledge graph of water affair. Finally, these methods are tested and evaluated. The main results and contributions of this paper are summarized below:

(1) Construction of a top-level knowledge graph of water affair: This paper builds a top-level knowledge graph of water affair based on Wordnet, Dbpedia, water industry standards, and expert experience, and provides a solid foundation for building a complete water affair knowledge graph.

(2) Methodology for constructing a water affair knowledge graph: In this paper, the water affair knowledge graph is improved by the edit distance algorithm and the LDA text semantic similarity algorithm, and the water affair information recommendation algorithm based on the water affair knowledge graph is prepared.

(3) Methodology of a water affair knowledge graph recommendation based on semantic distance: The recommendation method considers the water affair knowledge graph to calculate the similarity of the water affair information, and the water affair knowledge graph information that exceeds a given threshold which is recommended.

5. Conclusions

With the continuously increasing demand for water conservancy information in social production and life, the complexity of data involved in water conservancy information is increasing, which leads to challenges in data integration and low data utilization. To meet the needs of users for the integration of water data information, a method of constructing a knowledge graph by combining water affair structured and unstructured data is proposed. For meeting the needs of a water information search, an information recommendation system for a water affair knowledge graph is proposed. In this paper, the editing distance algorithm and LDA algorithm are used to construct the water affair-structured data and unstructured data combination knowledge graph, and the water affair-based knowledge graph is recommended based on the semantic distance algorithm. Finally, this paper uses the recall rate, accuracy rate, and F comprehensive results to compare the algorithms. The evaluation results in Section 3.4 shows that the evaluation results of the edit distance algorithm and the LDA algorithm are greater than that of the comparison algorithm, which verifies the validity and accuracy of the construction of the water affair knowledge graph. Furthermore, a set of water affair verification sets is used to verify the recommendation method, which proves the effectiveness of the recommended method. The results of this study promote the recommendation of water affair information. They also enhance the integrity of knowledge integration in the water affair sector and meet the user's need for knowledge in the water affair sector, thereby increasing the practical value of the water affair knowledge graph.

Based on the models and methods proposed in this study, planned further work will proceed in two directions. In this study, the mapping of the knowledge graph is used to supplement the structure of the natural knowledge graph of water affair. However, due to the continuous improvement in technologies and processes, this method does not take into account the development of a knowledge graph in the water affair field. Therefore, a future domain knowledge graph adaptation should analyze and capture the concepts in the new water affair field knowledge graph. In addition, the present study also explored water affair information recommendation methods based on a knowledge graph of water affairs to recommend water affair information to users. However, this method is somewhat tedious to implement. Therefore, we seek better ways to increase the efficiency of the recommendation process.

Author Contributions: T.L. designed the construction of the water affair knowledge graph, applied the model, and performed all of the experiments. J.Y. and Y.Y. provided comments and helped modify the manuscript. All of the authors have read and approved the final manuscript.

Funding: This research was funded by [Water Pollution Control and Treatment Science and Technology Major Project] grant number [2018ZX07111005] and the APC was funded by [Water Pollution Control and Treatment Science and Technology Major Project].

Acknowledgments: The authors thank Jianzhuo Yan and Yongchuan Yu for their thoughtful advice and suggestions on research design and implementation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kwok, C.; Jane, H.; John, D. MatSeek: An Ontology-Based Federated Search Interface for Materials Scientists. *IEEE Intel. Syst.* **2009**, *24*, 47–56.
2. Fabiano, B.R.; Giancarlo, G.; Ricardo, A.F.; Cássio, C.R.; Victor, A.S. From reference ontologies to ontology patterns and back. *Data Knowl. Eng.* **2017**, *109*, 41–69.
3. Hao, W. *Research on the Construction of TCM Health Knowledge Map*; Beijing Jiaotong University: Beijing, China, 2017.
4. Uschold, M. Ontologies principles, methods and applications. *Knowl. Eng. Rev.* **1996**, *11*, 93–155. [[CrossRef](#)]
5. Tham, K.D.; Fox, M.S.; Gruninger, M. A cost ontology for enterprise modeling department of industrial engineering. In Proceedings of the Third Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, Morgantown, WV, USA, 17–19 April 1994; pp. 111–117.
6. Fernandez, M.; Gómez-Perez, A.; Juristo, N. METHONTOLOGY: From ontological art towards ontological engineering. In Proceedings of the Spring Symposium on Ontological Engineering of AAAI, Stanford, CA, USA, 24–26 March 1997; pp. 33–40.
7. Noy, N.F.; McGuinness, D.L. *Ontology Development 101: A Guide to Creating Your First Ontology*; Stanford Knowledge Systems Laboratory Technical Report KSL-01-05; Stanford University: Stanford, CA, USA, March 2001.
8. Chen, Y.G.; Chu, H.C.; Chen, Y.M.; Chao, C.Y. Adapting domain ontology for personalized knowledge search and recommendation. *Inf. Manag.* **2013**, *50*, 285–303. [[CrossRef](#)]
9. Razia, S.; Subburaj, R. Ontology and context based recommendation system using Neuro-Fuzzy Classification. *Comput. Electr. Eng.* **2018**. [[CrossRef](#)]
10. Ren, R.; Zhang, L.L.; Cui, L.M.; Deng, B.; Shi, Y. Personalized Financial News Recommendation Algorithm Based on Ontolog. *Procedia Comput. Sci.* **2015**, *55*, 843–851. [[CrossRef](#)]
11. Kethavarapu, U.P.K.; Saraswathi, S. Concept Based Dynamic Ontology Creation for Job Recommendation System. *Procedia Comput. Sci.* **2016**, *85*, 915–921. [[CrossRef](#)]
12. Pablo, C.; David, V.; Phivos, M.; Yannis, A. Self-tuning personalized information retrieval in an ontology-based framework. In *On the Move to Meaningful Internet Systems*; Springer: Berlin, Heidelberg, 2005; pp. 977–986.
13. Gregorowicz, A.; Kramer, M.A. *Mining a Large-Scale Term-Concept Network from Wikipedia*; Mitre Corporation: McLean, VA, USA, 2006.
14. Shinzato, K.; Torisawa, K. Acquiring Hyponymy Realations from Web Documents. In Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Meeting, Boston, MA, USA, 2–7 May 2004; pp. 73–80.
15. Etzioni, O.; Cafarella, M.; Downey, D. Web-scale information extraction in knowitall: (Preliminary results). In Proceedings of the International Conference on World Wide Web (WWW 2004), New York, NY, USA, 17–20 May 2004; pp. 100–110.
16. Carlson, A.; Betteridge, J.; Kisiel, B. Toward an Architecture for Never-Ending Language Learning. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010), Atlanta, GA, USA, 11–15 July 2010; pp. 529–573.
17. Eduard, B. Property type distribution in Wordnet, corpora and Wikipedia. *Expert Syst. Appl.* **2015**, *42*, 3501–3507.

18. Sarah, D.; Abderrahim, E.Q.; Hamid, B. Enriching User Queries Using DBpedia Features and Relevance Feedback. *Procedia Comput. Sci.* **2018**, *127*, 499–504.
19. Andrew, L.C.; Ridha, K. Conto: A Protégé Plugin for Configuring Ontologies. *Procedia Comput. Sci.* **2016**, *83*, 179–186.
20. Konstantinos, N.V.; Theofanis, K.G.; Pericles, A.M. RDOTe—Publishing Relational Databases into the Semantic Web. *J. Syst. Softw.* **2013**, *86*, 89–99.
21. Zhang, R.Q.; Xu, C.; Cheung, S.C.; Yu, P.; Ma, X.X.; Lu, J. How effectively can spreadsheet anomalies be detected: An empirical study. *J. Syst. Softw.* **2017**, *126*, 87–100. [[CrossRef](#)]
22. Li, W.; Guo, K.; Shi, Y.; Zhu, L.Y.; Zheng, Y.C. DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. *Knowl.-Based Syst.* **2018**, *146*, 203–214. [[CrossRef](#)]
23. Arup, S.; Ujjal, M.; Utpal, B. Towards Bengali DBpedia. *Procedia Technol.* **2013**, *10*, 890–899.
24. Zeina, A.A.; Romain, R.; Jean-Yves, R.; Patrick, M. A parallel graph edit distance algorithm. *Expert Syst. Appl.* **2018**, *94*, 41–57.
25. Tu, S.Z.; Huang, M.L. Mining microblog user interests based on TextRank with LDA factor. *J. China Univ. Posts Telecommun.* **2016**, *23*, 40–46.
26. Li, W.J.; Xia, Q.X. A Method of Concept Similarity Computation Based on Semantic Distance. *Procedia Eng.* **2011**, *15*, 3854–3859. [[CrossRef](#)]
27. Liu, Q.; Li, S. Calculation of Lexical Semantic Similarity Based on HowNet. *Chin. Comput. Linguist.* **2010**. [[CrossRef](#)]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).