Research of Hierarchy Calculation Based Semantic Dimensionality Reduction

Q. Zhang
Shenhua Helishi Information Technology Co, Ltd. Beijing, China

X. Guo
School of Computer and Information Technology, Shanxi University, Taiyuan, China

D.D. Lv, S.H. Yuan, Y.Q. Zhang,
Department of Computer Science and Technology, Tongji University, Shanghai, China

T. Pan
Shenhua Helishi Information Technology Co, Ltd. Beijing, China

ABSTRACT: The text data contain rich semantic information. This paper proposed a semantics based on method of text dimension reduction based on ontology. This method utilized Dictionary Library WordNet to discovery semantic relationships in the feature graph of text data and constructed text feature ontologies for each text dataset, including concepts, semantic relations, attributes. Then for text data dimension reduction, we set hierarchy operations on text feature ontologies were given. Finally, an example is demonstrated to show the efficiency and accuracy of the proposed method.

KEYWORD: ontology; hierarchy calculation; dimensionality reduction

1 INTRODUCTION

In the Internet of Things and cloud computing era, the mass of data and high-dimensional data make data dimensionality reduction more practical in application value. For this type of complex text data, efficient and accurate dimensionality reduction methods can remove redundant and irrelevant dimensions, greatly reducing data processing and analysis calculation. The so-called data dimensionality reduction refers to the process that a linear or nonlinear mapping is used to map the samples from high-dimensional space to low-dimensional space to obtain a meaningful low-dimensional representation of high-dimensional data. There are many data dimensionality reduction methods, such as filter strategy [1], wrapper strategy [2] and so on. It is worth noting that the linear and nonlinear methods make calculation convenient and the results easily explained, such as principal component analysis (PCA)[3], linear discriminant analysis (LDA)[4], manifold learning based local linear embedding (LLE)[5], Laplacian Eigenmap[6] ,Isomap[7], Self-Organizing Maps[8] and so on.

In this paper, common ontology WordNet is used to construct text feature ontology for each text feature graphs and express the semantics of text features graphs. In order to find the best dimensionality reduction program, a Hierarchy Calculation method based ontology for text data dimension reduction on text feature ontologies were given. Fianlly, an instance is demostrate to show the efficiency and accuracy of the proposed approach.

2 ONTOLOGY ALGEBRA SYSTEM

Prasenjit Mitra proposed ontology algebra system ONION[9], wherein ontology is expressed as *O*=(*G*, *R*), wherein *G* is a directed graph, and *R* is a set of rules. *G = (V, E)* is composed by set of vertices and edges. Vertex set *V* is a conceptual noun or noun phrase. Edge set *E* is expressed in (*Subject r Object*) form, *Subject*, *Object* ∈ *V*, wherein *r* is the relationship between *Subject* and *Object*. The algebraic system mainly defines a unary operator "select" and three binary operators "Intersection", “Union” and “Difference”. The algebra system proposed by Saket Kaushik expresses ontology as a five-tuple *(C, P, CN, I, IN)*, *C, P, CN, I, IN* respectively represent concepts, attributes, concept name, instance and instance name[10]. This paper proposes a definition of ontology-based hierarchical operation. First, the basic concepts are defined as follows:

Definition 2.1: text concept set *C*.

$$C = \{c_1, c_2, ...\} \tag{1}$$

Definition 2.2: Concept relation set *E*.

$$E = \{e_1, e_2, ...\} \tag{2}$$

$\forall e \in E$, $e = (c\_1, c\_2, asso\_Type, weight)$,

$c\_1 \in C$, $c\_2 \in C$. $asso\_Type$ is the type of relationship between concepts, and *weight* is the relationship weight.

Definition 2.3: Set E' without consideration into the type of the relationship between the concepts.

$$E' = \{e_1', e_2', ...\} \tag{3}$$

$\forall e' \in E', \ e' = (e.c\_1, e.c\_2)$.

Definition 2.4: Set E'' without consideration into the relationship weight between the concepts.

$$E'' = \{e_1'', e_2'', ...\} \tag{4}$$

$\forall e'' \in E'', \ e'' = (e.c\_1, e.c\_2, e.asso\_Type)$.

Definition 2.5: text feature ontology *O*.

$$O = (C, E \tag{5}$$

We use the OWL language to describe ontology of the text feature, view the point which represented feature words in the text as a class (Class), query the two point which have relationship represented feature words in the WordNet whether there is semantic relations. If that is corrected, we regard the semantic relations as the first node ObjectProperty, relationship weight is DataProperty. For example, the two points-"corporation and distributor"- queried in WordNet have the semantic relations hypernym, the relationship weight is 87.65.The text features ontology of part of the OWL text file are as follows:

```
   <owl:Class rdf:ID="distributor"/>
      <owl:Class rdf:ID="corporation">
         <hy rdf:resource="#distributor"/>
         <relatetodistributor
rdf:datatype="&xsd;float"
         >87.65</relatetodistributor>
      </owl:Class>
      <owl:ObjectProperty rdf:ID="hy">
         <rdf:type
rdf:resource="&owl;TransitiveProperty"/>
      </owl:ObjectProperty>
      <owl:DatatypeProperty
rdf:ID="relatetodistributor">
         <rdfs:range rdf:resource="&xsd;float"/>
      </owl:DatatypeProperty>
```

In fact, the semantic relations queried is attributeOf or attribute in WordNet, that is the attribute relationship and attribute value relationship, which is similar to the DataProperty meaning of the OWL ontology language; the semantic relations queried is hypernymin WordNet, and is similai to the subclass meaning of OWL ontology language. But in order to facilitate the programming, we still view it as the relationship between concepts that is ObjectProperty. In other word, all concepts DataProperty only have relationship weights and the semantic relations queried in WordNet are ObjectProperty. Because some relationships are symmetry, in order to reflect semantic structure of the text feature Graphs more concise and hierarchical, we only use top-down relationship.

For example, Hyponymy -"hypernym, hyponym"- only uses hypernym, the relationship - "holonym, meronym" only uses the holonym. Type of relationship which cannot find in WordNet is defined as relation. In addition, the relationship will be used synonyms sometimes.

Table 1. The semantic types When using WordNet to build the text feature ontology

type of semantic	abbreviation	speech
synonym	sy	nouns, verbs, adjectives and adverbs
hypernym	hy	nouns, verbs
holonym	ho	nouns
cause	ca	verbs
attribute	at	Nouns (1st word), adjectives (2nd one)
instance	i	nouns
relation	r	nouns, verbs, adjectives and adverbs

2.1 *Hierarchy Calculation*

Secondly, we define the Hierarchy Calculation. Hierarchy calculation is binary calculation based on the concept relationship between the feature ontology and the concept, including the ontology concept calculation. Assuming that there is a text feature ontology, $O_1 = (C_1, E_1), c \in C_1$, $O=(C, E)$ is obtained through hierarchy calculation of the text feature ontology O_1 and concept *c*.

Definition 2.8 Drill-up calculation: $O = O_1 \oplus \mathrm{c}$.

```
If $\neg\exists e^* \in E$, such that $e^*.asso\_Type=$hy or ho and $e^*.c\_2 = c$, then
    $\forall e \in E$, if $e.asso\_Type=$hy or ho and $e.c\_1=c$, then
        $\forall \bar{e} \in E$, if $\bar{e}.asso\_Type \neq hy \ or \ ho$ and ($\bar{e}.c\_1 = c$ or $\bar{e}.c\_2 = c$), then
                if $\bar{e}.c\_1 = c$, then
                    $\bar{e}.c\_1 = e.c\_2$
                if $\bar{e}.c\_2 = c$, then
                    $\bar{e}.c\_2 = e.c\_2$
                    if $\exists \hat{e} \in E$, such that $\hat{e}' = \bar{e}'$, then
            if $\hat{e}.asso\_Type \neq \bar{e}.asso\_Type$, then
                if $\bar{e}.asso\_Type = r$, then
                    $\bar{e}.asso \ \ Ty\neq p\hat{e}. \ \ e \ ass$
            if $\bar{e}.asso\_Type \neq r$ and $\hat{e}.asso\_Type \neq r$, then
                    if $\hat{e}.weight > \bar{e}.weight$, then
                        $\bar{e}.asso\_Type = \hat{e}.asso\_Type$
                    $\bar{e}.Weight \leftarrow \max(\bar{e}.weight, \hat{e}.weight)$
                        $E \leftarrow E - \{\hat{e}\}$
            $E \leftarrow E - \{\bar{e}\}$
        $E = E - \{e\}$
    $C = C - \{c\}$
```

Definition 2.9 Drill-down calculation: $O = O_1 \odot c$.

```
if $\neg\exists e^* \in E$, such that $e^*.asso\_Type$ = hy or ho and $e^*.c\_1 = c$, then
    $\forall e \in E$, if $e.asso\_Type$= hy or ho and $e.c\_2=c$, then
        $\forall \bar{e} \in E$, if $\bar{e}.asso\_Type \neq$ hy or ho and ($\bar{e}.c\_1=c$ or $\bar{e}.c\_2=c$), then
            if $\bar{e}.c\_2 = c$, then
                $\bar{e}.c\_2 = e.c\_1$
            if $\bar{e}.c\_1 = c$, then
                $\bar{e}.c\_1 = e.c\_1$
            if $\exists \hat{e} \in E$, such that $\hat{e}' = \bar{e}'$, then
                if $\hat{e}.asso\_Type \neq \bar{e}.asso\_Type$, then
                    if $\bar{e}.asso\_Type = r$, then
                        $\bar{e}.asso\_Type = \hat{e}.asso\_Type$
                    if $\bar{e}.asso\_Type \neq r$ and $\hat{e}.asso\_Type \neq r$, then
                        if $\hat{e}.weight > \bar{e}.weight$, then
                            $\bar{e}.asso\_Type = \hat{e}.asso\_Type$
                $\bar{e}.Weight \leftarrow \max(\bar{e}.weight, \hat{e}.weight)$
                $E \leftarrow E - \{\hat{e}\}$
            $E \leftarrow E - \{\bar{e}\}$
        $E = E - \{e\}$
    $C = C - \{c\}$
```

3 EXPERIMENTS

WordNet 2.1 programming interface is used to obtain semantic relationship of both ends of each edge of text feature graphs and construct the text features graphs to text feature ontology in the OWL format. Ontology is processed by Jena API to achieve calculation and semantic dimensionality reduction based on these calculations.

Fig.1 and Fig.2 are two text feature ontology O_1 and O_2 O_1 = {C_1, E_1} , O_2 = {C_2, E_2}. Fig.3 is result of O_3=$\otimes O_1$, namely that the text feature ontology O_1 is subjected to Drill-up calculation and dimensionality reduction to obtain text feature ontology O_3. Fig. 4 is result of O_4=$\odot(\otimes O_2)$, namely that that the text feature ontology O_2 is subjected to two hierarchy calculation(Drill-up calculation and Drill-down calculation) and dimensionality reduction to obtain text feature ontology $O_{4.}$ As can be seen from the experimental results, the proposed hierarchy calculation can be flexibly combined to obtain different semantic dimensionality reduction results and achieve semantic dimensionality reduction in terms of the internal of the text feature graphs.

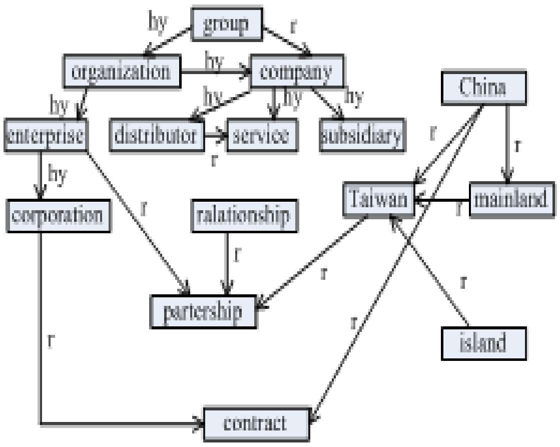

Figure 1. text feature ontology O1

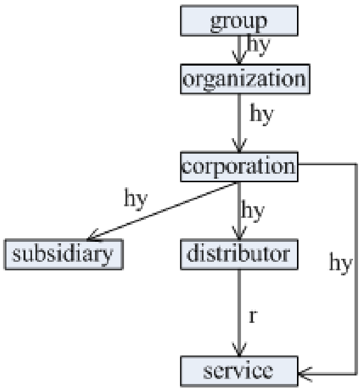

Figure 2. text feature ontology O2

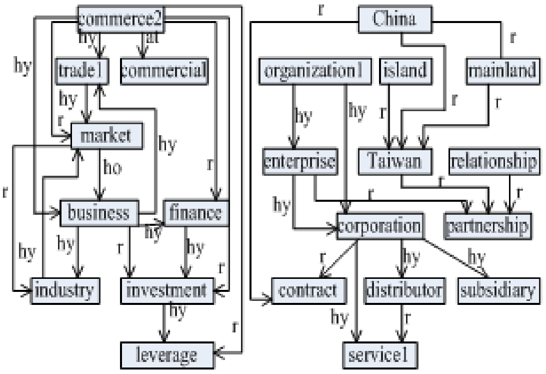

Figure 3. text feature ontology O_3= $\otimes O_1$

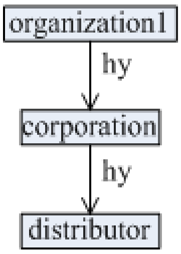

Figure 4. text feature ontology O_4=$\odot(\otimes O_2)$

4 CONCLUSION

From the semantic point of view, this paper achieves semantic-based text data dimensionality reduction by building text feature ontology, defining hierarchy calculation of ontology. This approach is reflected in the semantics, demonstrating that this method is relatively high in performance-price ratio, and its unique ontology hierarchy calculation for dimensionality reduction as the text feature ontology is featured in high flexibility and relatively strong scalability.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China(71171148, 61403238), the National Key Technology R&D Program (2012BAD35B01, 2012BAH13F04) , the National High-Tech Research and Development plan of China (2012AA062203), the National Basic Research Program of China(2014CB340404), and the Natural Science Foundation of Shanxi (2014021022-1).

REFERENCES

[1] Estévez, P. A.. Tesmer M, Perez C A, et al 2009. Normalized mutual information feature selection. IEEE Transactions on Neural Networks20(2): 189~201.

[2] Fodor, I. K. 2002. A survey of dimension reduction techniques. California: Lawrence Livermore National Lab.

[3] Abdi, H. & Williams, L. J.2010. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2(4): 433~459.

[4] Zhao, H. & Yuen, P. C.2008. Incremental linear discriminant analysis for face recognition. Systems, IEEE Transactions on Man, and Cybernetics, Part B: Cybernetics 38(1): 210~221.

[5] Zhang, Y. et al. 2008. Local linear embedding in dimensionality reduction based on small world principle. 2008 International Conference on Computer Science and Software Engineering. Piscataway, NJ: IEEE, 394~398.

[6] Luo, W.2011. Face recognition based on Laplacian Eigenmaps. 2011 International Conference on Computer Science and Service System. Piscataway, NJ: IEEE, 416~419.

[7] Cho, M & Park, H.2009. Nonlinear dimension reduction using ISOMap based on class information. International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 566~570.

[8] Sagheer, A.2010. Piecewise one dimensional Self Organizing Map for fast feature extraction. 10th International Conference on Intelligent Systems Design and Applications. Piscataway, NJ: IEEE, 633~638.

[9] Mitra, P. & Wiederhold, G. 2004. An ontology-composition algebra. Handbook on ontologies. Berlin Heidelberg: Springer, 93~113.

[10] Kaushik, S. et al.2006 An algebra for composingontologies. 2006. Conference on Formal Ontology in Information Systems. Amsterdam: IOS Press, 265~276.