Lexical-semantic SLVM for XML Document Classification

Jun Long
School of Information Science and Engineering, Central South University, Changsha 410075, China
Email: jlong@csu.edu.cn

Luda Wang
School of Information Science and Engineering, Central South University, Changsha 410075, China
Xiangnan University, Chenzhou 423000, China
Email: wang_luda@csu.edu.cn

Zude Li, Zuping Zhang, Huiling Li and Guihu Zhao
School of Information Science and Engineering, Central South University, Changsha 410075, China
Email: zli@csu.edu.cn, zpzhang@csu.edu.cn, lihuiling2013@csu.edu.cn

***Abstract*—Structured link vector model (SLVM) and its improved version depend on statistical term measures to implement XML document representation. As a result, they ignore the lexical semantics of terms and its mutual information, leading to text classification errors. This paper proposed a XML document representation method, WordNet-based lexical-semantic SLVM, to solve the problem. Using WordNet, this method constructed a data structure for characterizing lexical semantic contents of XML document, and adjusted EM modeling to disambiguate word stems. Then, synset matrix of lexical semantic contents was built in the lexical-semantic feature space for XML document representation, and lexical semantic relations were marked on it to construct the feature matrix in lexical-semantic SLVM. On categorized dataset of Wikipedia XML, using NWKNN classification algorithm, the experimental results show that the feature matrix of our method performs F1 measure better than original SLVM and frequent sub-tree SLVM based on TF-IDF.**

***Index Terms*—Semi-structured document, SLVM, Lexical semantics, Classification, Feature matrix**

I. Introduction

In order to record semi-structured information, lots of document standards, such as HTML, BibTex and SGML/XML, are recommended by many main international standards organizations. The structural flexibility of these semi-structured documents offers an approach to representing information in specified structure. Contrasting conventional plain documents, semi-structured documents represent their syntactic structure via the use of document structural elements marked by user-specified tags, and the associated schema specified in Schema format [1]. Extensible Markup Language (XML) is a semi-structured document format with strong support via Unicode for different human languages. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures,[6] for example in web services.

Considering the structure property, classification methods for semi-structured document analysis have to consider the information embedded in both the element tags and their associated contents. Recently, the Structured Link Vector Model (SLVM) [2] was proposed for XML documents analysis. Especially, the extended version of SLVM uses the closed frequent sub-trees as structural units for content extraction from the XML document. For XML document representation, existing models depend on statistical term measure for feature extraction. However, in the information retrieval field, statistical term measures causes XML document analysis to perform on the level of term string basically, and neglect lexical semantic contents in the structured elements.

Semantic approach is an effectively used technology for document analysis. It can capture the semantic features of words under analysis, and based on that, characterizes and classifies the document. Close relationship between the syntax and the lexical semantics of words have attracted considerable interest in both linguistics and computational linguistics. For XML document classification, the design and implementation of lexical-semantic SLVM take account of the lexical semantics particularly. Unlike present models, using WordNet [3], our model developed a new term measure which can characterize lexical semantic contents and relations, and provides a practical method for XML document representation which can handle the impact of synonyms and other relations. Theoretical analysis and relevant experiments are carried out to verify the effectiveness of this model.

doi:10.4304/jsw.9.12.3028-3034

II. An Overview On Original Slvm And Frequent Sub-Tree Slvmhelpful Hints

Structured Link Vector Model (SLVM) was proposed by Jianwu Yang [2], which forms basis of our work, was proposed for representing XML documents. It was extended from the conventional vector space model (VSM) by incorporating document structures (represented as term-by-element matrices), referencing links (extracted based on IDREF attributes), as well as element similarity (represented as an element similarity matrix). On the other hand, based on original SLVM, an extended VSM utilize the closed frequent sub-trees as structural units for content extraction from the XML document [1], which are called frequent sub-tree SLVM in this context.

A. XML Document Representations

SLVM represents a XML document doc_x using a document feature matrix $\Delta_x \in R^{n\times m}$, given as [2]

$$\Delta_x = \left[\Delta_{x(1)}, \Delta_{x(2)}, \cdots\cdots, \Delta_{x(m)}\right], \qquad (1)$$

where m is the number of distinct XML document elements, $\Delta_{x(i)} \in R^n$ is the TFIDF feature vector representing the i^{th} XML element (e_i), given as $\Delta_{x(i,j)} = TF(w_j, doc_x.e_i) \cdot IDF(w_j)$ for all $j=1$ to n, and $TF(w_j, doc_x.e_i)$ is the frequency of the term w_j in the element e_i of doc_x.

In frequent sub-tree SLVM, each document is represented as a matrix based on SLVM, where the selected closed frequent substrees are regarded as structural unit [1]. In order to deal with exception that a document do not include any the selected closed frequent substrees, it adds the vector of the document based on VSM into the matrix as a column.

B. Similarity Measures

The SLVM-based document similarity between two XML documents doc_x and doc_y is defined as [2]

$$Sim(doc_x, doc_y) = \sum_{j=1}^{m}\sum_{i=1}^{m} M_{e(i,j)} \times \left(\tilde{\Delta}_{x(i)}{}^{T} \cdot \tilde{\Delta}_{y(j)}\right), \qquad (2)$$

where $\tilde{\Delta}_{x(i)}$ or $\tilde{\Delta}_{y(j)}$ is the normalized document feature matrix of documents doc_x or doc_y, and M_e is a matrix of dimension $m\times m$ and named as the element similarity matrix. The matrix M_e captures both the similarity between a pair of document structural elements as well as the contribution of the pair to the overall document similarity. To obtain an optimal M_e for a specific type of XML data, SLVM-based document similarity learn the matrix using pair-wise similar training data (unsupervised learning) in an iterative manner [4].

C. Analysis of Feature Extraction

In the above, these models for document representation are perceived as the mode using statistical term measures. As a sort of ontology methods [5], XML document representations based on statistical term measures ignore recognition of lexical semantic contents. It causes the document representation to lose the mutual information [6] of term meanings which comes from synonyms in different samples. Our comment on statistical term measures and XML document representation can be clarified by analyzing a small XML corpus *Example 1*.

Example 1

```
<!--Doc B.-->
<article>
<title>tree</title>
    <body>Human treasures trees.</body>
</article>
```

```
<!--Doc A.-->
<article>
              <title>maple</title>
  <body>Men prize maple.</body>
</article>
```

title body

$$\Delta_A = \begin{bmatrix}1\\0\\1\\0\\1\\0\end{bmatrix}\begin{array}{l}\text{men}\\ \text{human}\\ \text{prize}\\ \text{treasures}\\ \text{maple}\\ \text{trees}\end{array}$$

title body

$$\Delta_B = \begin{bmatrix}0\\1\\0\\1\\0\\1\end{bmatrix}\begin{array}{l}\text{men}\\ \text{human}\\ \text{prize}\\ \text{treasures}\\ \text{maple}\\ \text{trees}\end{array}$$

Figure 1. The SLVM feature matrices for Example 1

In *Example 1*, the two simple XML documents are viewed as two document samples, and these two documents comprise the small corpus. Evidently, the meanings of *Doc A* and *Doc B* are extremely equivalent. Thus, the correlation and semantic similarity between these two documents are considerable. But, SLVM and frequent sub-tree SLVM can not display the document similarity between *Doc A* and *Doc B*. Obviously, because document representation matrices shown in Fig.1 make each $\tilde{\Delta}_{A(i)}{}^{T} \cdot \tilde{\Delta}_{B(j)} = 0$, so the $Sim(doc_A, doc_B) = 0$, using the Eq. (2). Then, on behalf of statistical term measures, the document representations on *Example 1* did not perform well for semantic similarity.

III. Proposed Program

A. Preliminary Conception and Theoretical Analysis

For document analysis, document representations which depend on statistical term measures shall lose mutual information of term meanings. Besides, in different documents, term meanings are relevant to specific synonyms which are involved by lexical semantic contents. Thus, our model resorts to WordNet [3], a lexical database for English, for extracting lexical semantics Then, the method of document representation will construct a lexical-semantic SLVM of XML document in order to define feature matrix for classification.

In WordNet, a form is represented by a string of ASCII characters, and a sense is represented by the set of (one or more) synonyms that have that sense [3]. Synonymy (syn same, onyma name) is a symmetric relation between word forms [3]. Synonymy is WordNet's basic relation, because WordNet uses sets of synonyms (synsets) to represent word senses. Videlicet, shown as Fig 1, one word refers to several sets of synonyms (synsets).

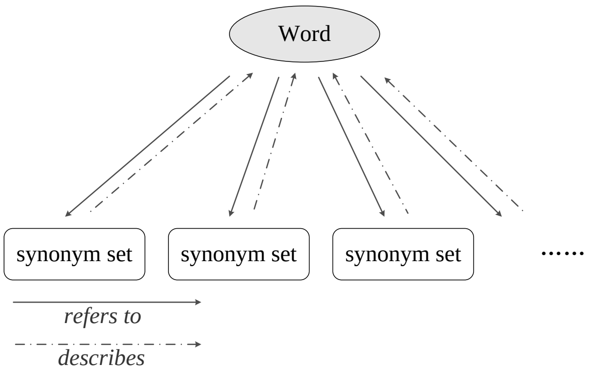

Figure 2. Word and synonym sets

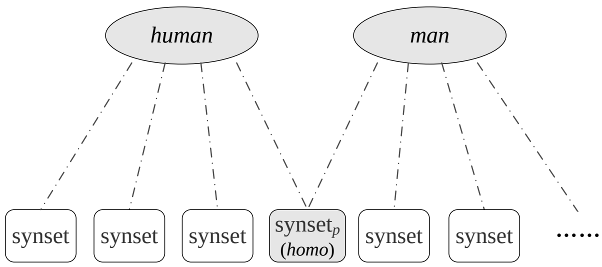

Figure 3. Common semantic-factor of words

WordNet contains 117659 sets of synonyms (synsets) [3]. In Fig. 2, because one word or term refers to particular synonym sets, our preliminary conception is that several particular synsets can strictly describe the meaning of one word for characterizing lexical semantic contents. Furthermore, those particular synsets can indicate lexical semantic between words, using Antonymy, Hyponymy, Meronymy, Troponomy and Entailment between synsets [3]. Then, our method defines these particular synonym sets as the semantic-factors of the word.

Based on the above definition, involved semantic-factors can character the lexical semantic contents of *Example 1*, which shall accomplish feature extraction of lexical semantic contents. For instance, in Fig. 2, the words *human* and *man* belong to different document samples in *Example 1*, and the common semantic-factor synset$_p$ (*homo*) that simultaneously describes the meanings of *human* and *man* can gain mutual information [6] between term meanings. Moreover, our document representation is able to capture the lexical semantic mutual information between samples which lies with a number of synonyms in different documents.

According to the statistical theory of communications, our conception needs further analysis for theoretical proof. The analysis first introduces some of the basic formulae of information theory [6, 7], which are used in our theoretical development of samples mutual information. Now, let x_i and y_j be two distinct terms (events) from finite samples (event spaces) *X* and *Y*. Then, let X or Y be random variable representing distinct lexical semantic contents in samples *X* or *Y*, which occur with certain probabilities. In reference to above definitions, mutual information between X and Y , represents the reduction of uncertainty about either X or Y when the other is known. The mutual information between samples, $I(X;Y)$, is specially defined to be [7]

$$I(X;Y) = \sum_{x_i \in X}\sum_{y_j \in Y} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}. \quad (3)$$

In the statistical methods of SLVM, probability $P(x_i)$ or $P(y_j)$ is estimated by counting the number of observations (frequency) of x_i or y_j in sample *X* or *Y*, and normalizing by *N*, the size of the corpus. Joint probability, $P(x_i, y_j)$, is estimated by counting the number of times (related frequency) that term x_i equals (is related to) y_j in the respective samples of themselves, and normalizing by *N*.

Taking the *Example 1*, according to SLVM feature matrices (shown in Fig. 1), between any term x_i in *Doc A* and any term y_j in *Doc B*, there is not any counting of times that x_i equals y_j . As a result, in *Example 1*, the statistical term measures indicate $P(x_i, y_j) = 0$ so the samples mutual information $I(X;Y) = 0$. Thus, the analysis verifies that the statistical methods of feature extraction lose mutual information of term meanings.

On the other hand, for feature extraction of lexical semantic contents, our method uses several particular semantic-factors to describe the meaning of one word or term. In different samples, words can be related to other words which are described by same synsets. Then, lexical semantic mutual information between samples, $I(X;Y)$, is re-defined to be

$$I(X;Y) = \sum_{x_i \in X} \sum_{y_j \in Y} F(s_{x_i,y_j}) \bmod N \log \frac{F(s_{x_i,y_j}) \bmod N}{F(s_{x_i}) \bmod N \cdot F(s_{y_j}) \bmod N} . \quad (4)$$

To denote probability $P(x_i)$ or $P(y_j)$, function $F(s_{x_i})$ or $F(s_{y_j})$ is estimated by calculating the frequency of semantic-factors that describe the meaning of x_i or y_j in sample *X* or *Y*, and modulo *N*, the total of semantic-factors in corpus. Meanwhile, to denote joint probability $P(x_i, y_j)$, function $F(s_{x_i,y_j})$ is estimated by calculating the frequency of common semantic-factors that simultaneously describe the meaning of x_i and y_j, and modulo *N*.

$$\Delta_A = \begin{array}{c} \text{title body} \\ \begin{bmatrix} \vdots \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ \vdots \end{bmatrix} \end{array} \begin{array}{l} \\ \texttt{synset}_1 \\ \vdots \\ \texttt{synset}_p \ (homo) \\ \vdots \\ \texttt{synset}_q \ (value) \\ \vdots \\ \texttt{synset}_n \end{array}$$

$$\Delta_B = \begin{array}{c} \text{title body} \\ \begin{bmatrix} \vdots \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ \vdots \end{bmatrix} \end{array} \begin{array}{l} \\ \texttt{synset}_1 \\ \vdots \\ \texttt{synset}_p \ (homo) \\ \vdots \\ \texttt{synset}_q \ (value) \\ \vdots \\ \texttt{synset}_n \end{array}$$

Figure 4. The lexical-semantic matrices for the corpus shown in Example 1

In *Example 1*, joint probability $P(x_i, y_j)$ is estimated by calculating the frequency of common semantic-elements that relate to lexical semantic contents or relations of x_i and y_j, and modulo *N*. For instance, shown in Fig. 3, the words *human* and *man* are described by the common semantic-factor $synset_q$ (*homo*). Actually, according to document representation matrices shown in Fig. 4, $P(human, man) = F(\texttt{synset}_p) \bmod N > 0$. In addition, joint probability of semantic relation $P(tree, maple) = F(\text{Hyponymy}(tree, maple)) \bmod N > 0$. As a result, $I(X;Y) > 0$, so mutual information from lexical semantic contents and relations between *Doc A* and *Doc B* is positive. Thus, the analysis proves that the semantic-factors and feature extraction of lexical semantic contents can provide the probability-weighted amount of information (PWI) [7] between XML documents on the lexical semantic level.

B. Lexical-semantic SLVM

In our work, XML documents are represented using the lexical-semantic SLVM. In this model, each XML document is represented as a document feature matrix in the lexical-semantic structured link vector space. For organizing the lexical-semantic SLVM, the procedures are as follows. First, **(1)** a data structure of semantic-factor information is composed for feature extraction of lexical semantic contents. Secondly, **(2)** the EM modeling is used to disambiguate word stems. Furthermore, **(3)** this work constructs the feature space of lexical-semantic SLVM and builds the synset matrix in the space to characterize lexical semantic contents of XML. Lastly, **(4)** to characterize lexical semantic relations, it marks each vector in synset matrix with weights of 5 semantic relations between synsets [3]. Thus, the feature matrix in lexical-semantic SLVM is constructed via marking semantic relations on synset matrix.

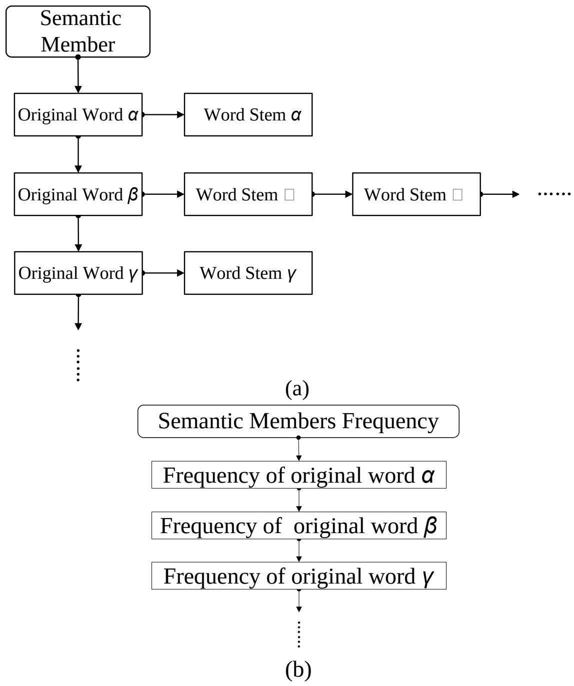

Figure 5. The linked lists of Semantic Member (a) and Semantic Members Frequency (b)

(1) The data structure of semantic-factor information comprises relevant information of each snyset in a document, which is formalized as a data element, shown in Table I. It can record all essential information of semantic-factors in a XML document, such as synset ID, weight, sample ID and relevant information of words. Note that, in a record of the data structure, each original word in inflected form [8] referring to the semantic-factor and its word stem(s) in base form [8, 9] are recorded by linked list of *Semantic Member* (shown in Fig. 5(a)). And, according to WordNet framework [3], when original word refers to more than 1 word stems, the list of *Semantic Member* will expend the node of original word to register all word stems.

Meanwhile, the list of *Semantic Members Frequency* is shown in Fig. 5(b). It records the frequency of each original word one by one in their order of *Semantic Member*.

TABLE I
DATA STRUCTURE OF SEMANTIC-FACTOR INFORMATION

Item	**Explanation**
Synset ID	Identification of synonym set
Set of Synonym	Synonymy is WordNet's basic relation. WordNet uses sets of synonyms (synsets) to represent word senses.[3]
Weight (Frequency)	Frequency of semantic-factor in a element (sum of Semantic Members Frequency)
Sample ID	Identification of semi-structured document sample
Element ID	Identification of structural element or unit in semi-structured document
Semantic Member	A linked list (shown in Fig. 3) which carries all Original Words of Terms referring to the semantic-factor and their Word Stem(s)
Semantic Members Frequency	A linked list (shown in Fig. 4) which carries frequency of each Original Words of Terms (that refer to the semantic-factor) one by one

(2) On the basis of data structure of semantic-factor information, *Semantic Member* needs to disambiguate word stems of original word. In case of an original word referring to more than 1 word stem in base form, semantic-factors must ensure that one original word refers to only 1 word stem. Then, in order to select only 1 word stem for an original word (shown in Fig. 6), we employ the Maximum Entropy Model [10].

ME modeling provides a framework for integrating information for classification from many heterogeneous information sources [11]. In our model, we put an assumption that diversity [12] of *Semantic Member* implies the significance of the semantic-factor and the rationality of existing *Semantic Members* in a document.

Assume a set of original words *X* and a set of its word stems *C*. The function $cl(x): X \to C$ chooses the word stem *c* with the highest conditional probability, which makes sure original word *x* only refers to: $cl(x) = \arg\max_c p(c \mid x)$. Each feature [11] of original word is calculated by a function that is associated to a specific word stem *c*, and it takes the form of Eq. (5), where S_i is the number of *Semantic Member* in semantic-factor *i*, P_j is the proportion of the *Frequency of original word j* to *Weight* in semantic-factor *i*, and the $-\sum_{j=1}^{S_i} P_j \log_2 P_j$ indicates *Semantic Member* diversity of semantic-factor *i* in a document, in the form of Shannon-Wiener index [12].

The conditional probability $p(c \mid x)$ is defined by Eq. (6). The parameter of the semantic-factor *i* [11], α_i , is the *Frequency* of original word *x* in semantic-factor *i*. *K* is the number of semantic-factors which word stem *c* refers to, and $Z(x)$ is a value to ensure that the sum of all conditional probabilities for this context is equal to 1.

$$f_i(x,c) = \begin{cases} -\sum_{j=1}^{S_i} P_j \log_2 P_j & \text{if } x \text{ refers to } c \text{ and } c \text{ refers to semantic-factor } i, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$$p(c \mid x) = \frac{1}{Z(x)} \prod_{i=1}^{K} \alpha_i^{f_i(x,c)} . \tag{6}$$

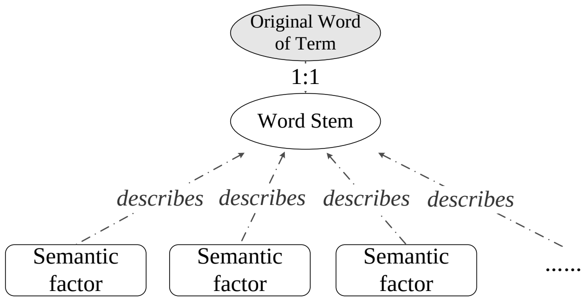

Figure 6. 1:1 reference of original word

Above equations aim at finding the highest conditional probability $p(c \mid x)$, and using the function $cl(x)$ to ensure that original word *x* refers to only 1 word stem (like Fig. 6). After semantic-factors characterizing lexical semantic contents of XML document preliminarily, the specified ME modeling is applied to implement disambiguation of word stems. Necessarily, the relevant items in the data structure of semantic-factor information shall be modified, such as the *Semantic Member*, the *Frequency of original word*, and the *Weight*. Furthermore, some relevant semantic-factors shall be eliminated.

(3) As for XML document dataset, all referred semantic-factors are fixed by disambiguation of word stems. Then, in feature space of lexical-semantic SLVM, a XML document doc_x is preliminary represented using a synset matrix $\Delta_x \in R^{n \times m}$, defined as

$$\Delta_x = \left[\Delta_{x(1)}, \Delta_{x(2)}, \cdots\cdots, \Delta_{x(m)}\right] , \tag{7}$$

$$\Delta_{x(i)} = \left(\Delta_{x(i,1)}, \Delta_{x(i,2)}, \cdots\cdots, \Delta_{x(i,n)}\right), \tag{8}$$

where *m* is the number of distinct XML document elements and units, $\Delta_{x(i)} \in R^n$ is the synset vector representing the i^{th} XML element or unit, given as $\Delta_{x(i,j)} = FS(s_j, doc_x.e_i) IDF(s_j)$ for all *j=1* to *n*, and $FS(s_j, doc_x.e_i)$ is the frequency of the semantic-factor s_j in e_i , in which e_i is the i^{th} XML element or unit of doc_x.

(4) In feature space of lexical-semantic SLVM, to characterize lexical semantic relations, our method marks *Antonymy*, *Hyponymy*, *Meronymy*, *Troponomy* and *Entailment* on each dimension of the synset matrix. The processing is formulized as

$$\Delta\overline{\Delta}_{x(i,q)} = \sum_{p=1}^{n} R(p,q)\ \overline{\Delta}_{x(i,p)}\ , \tag{9}$$

$$R(p,q) = \begin{cases} 0.5 & \textit{Antonymy} \\ 0.2 & \textit{Hyponymy, Meronymy, Troponomy} \text{ or } \textit{Entailment} \\ 0 & \text{Unrelated} \end{cases}, \tag{10}$$

where p and $q=1$ to n, n is dimensional number of the synset vector of the i^{th} XML element, and $\overline{\Delta}_{x(i,p)}$ is value of the p^{th} synset vector element of the i^{th} XML element. $\Delta\overline{\Delta}_{x(i,q)}$ is semantic relation increment to the q^{th} dimensional value of the synset vector, and function $R(p,q)$ denotes semantic relation coefficient for $\Delta\overline{\Delta}_{x(i,q)}$. Specifically, when the synset of q^{th} dimension is related to synset of p^{th} dimension via semantic relation such as Antonymy, Hyponymy, Meronymy, Troponomy or Entailment, the $R(p,q)$ assignment is shown in equation (7). The assignments of $R(p,q)$ reflect the semantic relations which are organized into synsets by WordNet.

As for XML documents, all synset dimensions carry the corresponding semantic feature values. Then, lexical-semantic SLVM represents a XML doc_x using a feature matrix $\underline{\Delta}_x \in R^{n\times m}$, defined as

$$\underline{\Delta}_x = \left[\underline{\Delta}_{x(1)}, \underline{\Delta}_{x(2)}, \cdots\cdots, \underline{\Delta}_{x(m)}\right] , \tag{11}$$

where m is the number of distinct XML elements and units. $\underline{\Delta}_{x(i)} \in R^n$ is the lexical-semantic vector representing the i^{th} XML element, given as

$$\underline{\Delta}_{x(i)} = \left(\underline{\Delta}_{x(i,1)}, \underline{\Delta}_{x(i,2)}, \cdots\cdots, \underline{\Delta}_{x(i,n)}\right), \tag{12}$$

$$\underline{\Delta}_{x(i,j)} = \overline{\Delta}_{x(i,j)} + \Delta\overline{\Delta}_{x(i,j)} . \tag{13}$$

where n is the number of identical *Synset ID* of all semantic-factors in XML dataset.

IV. Experiment And Result

A. The Experiment

In our work, experiments use three sorts of matrices to represent document sample: 1) document feature matrix based on original SLVM, 2) document feature matrix based on frequent sub-tree SLVM, 3) the lexical-semantic matrix in the lexical-semantic feature space based on lexical-semantic SLVM

In the XML document classification, the dataset consisting of 20 categories is composed of XML documents of the Wikipedia XML, the training set is composed of 5000 XML documents, and the test set is composed of 3000 XML documents.

In the experiments, to tackle unbalanced text dataset, we select an optimized KNN classification, the NWKNN (Neighbor-Weighted K-Nearest Neighbor) algorithm defined to be Eq. (14) [13]. As for NWKNN, each XML document d is considered to be a feature matrix based on original SLVM, frequent sub-tree SLVM, or lexical-semantic SLVM.

$$score(doc, c_i) = \mathrm{Weight}_i \left(\sum_{d_j \in KNN(d)} Sim(doc, doc_j)\,\delta(doc_j, c_i) \right) \tag{14}$$

$$\textit{subjected to} \qquad \delta(d_j, c_i) = \begin{cases} 1 & d_j \in c_i \\ 0 & d_j \notin c_i \end{cases}$$

In the process of Eq. (8), this algorithm uses the SLVM-based document similarity between feature matrices of *doc* and doc_j [2] to calculate the $Sim(doc, doc_j)$. Besides, according to experience of NWKNN algorithm [13], the parameter of Weight_i, Exponent [13], is equal to 3.5.

B. The Result

To evaluate the classification systems, we use the $F1$ measure [14]. Then, we can observe the effect of different kinds of data on a classification system [14]. For ease of comparison, we summarize the $F1$ scores over the different categories using the macro-averages and micro-averages of $F1$ score, and display mean average precision. Table II summarizes the experiment result.

TABLE II
THE EXPERIMENT RESULTS

Approaches	Macro-$F1$	Micro-$F1$	Mean Average precision
Original SLVM [1]	0.241853	0.295418	0.486702
Frequent sub-tree SLVM [1]	0.479748	0.518635	0.701861
Lexical-semantic SLVM	0.492090	0.521977	0.727203

V. Conclusion

In the work, a data structure of semantic-factor information is constructed to record relevant information of each semantic-factor in element of XML document. It can characterize lexical semantic contents and be adapted for disambiguation of word stems. Furthermore, in the the feature space, lexical semantic relations are marked on the synset matrix. Using the NWKNN algorithm, lexical-semantic SLVM achieve better performance of classification than document feature matrix built by original SLVM and frequent sub-tree SLVM which stand for the typical statistical method of feature extraction.

Our future research includes using more current algorithms based on the lexical-semantic matrix for XML document analysis, and developing a method for analyzing WSDL document on the basis of semantic-factor.

ACKNOWLEDGMENT

This work was supported by National High Technology Research and Development Program (863 Program) of China (No. 2012AA011205), National Natural Science Foundation of China (No. 61272150, 61472450, 61379109, 61103034), Research Fund for the Doctoral Program of Higher Education of China (No. 20120162110077), and Excellent Youth Foundation of Hunan Scientific Committee, China (No. 11JJ1012).

REFERENCES

[1] Jianwu Yang, Songlin Wang. "Extended VSM for XML Document Classification Using Frequent Subtrees ", In proceeding(s) of the 8th International Conference on Initiative for the Evaluation of XML Retrieval (INEX'09), 2010.

[2] Jianwu Yang, Xiaoou Chen. "Semi-Structured Document Model for Text Mining", Journal of Computer Science and Technology, vol.17, no.5, 2002, pp.603-610.

[3] George A. Miller. "WordNet: a lexical database for English" , Communications of the ACM, vol.38, no.11, 1995, pp. 39-41.

[4] Jianwu Yang, William Cheung, Xiaoou Chen. "Learning element similarity matrix for semi-structured document analysis", Knowledge and Information Systems, vol.19, no.1, 2009, pp.53-78.

[5] David Sánchez, Montserrat Batet. "A semantic similarity method based on information content exploiting multiple ontologies", Expert Systems with Applications, vol.40, no.4, 2013, pp.1393–1399.

[6] Akiko Aizawa. "An information-theoretic perspective of tf–idf measures", Information Processing and Management, vol.39, no.1, 2003, pp.45-65.

[7] Wen Zhang, Taketoshi Yoshida, Xijin Tang. "A comparative study of TF*IDF, LSI and multi-words for text classification", Expert Systems with Applications, vol.38, no.3, 2011, pp.2758–2765.

[8] Mihai Lintean and Vasile Rus. "Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics", In Proceeding(s) of the 25th International Florida Artificial Intelligence Research Society Conference, 2012, pp. 244-249.

[9] MIT Java Wordnet Interface, Version 2.3.3. http://projects.csail.mit.edu/jwi/api/edu/mit/jwi/morph/WordnetStemmer.html

[10] Armando Suárez, Manuel Palomar. "A maximum entropy-based word sense disambiguation system", In proceeding(s) of the 19th international conference on Computational linguistics (COLING '02). 2002, pp.1-7.

[11] Myunggwon Hwang, Chang Choi, Pankoo Kim. "Automatic enrichment of semantic relation network and its application to word sense disambiguation", IEEE Transactions on Knowledge and Data Engineering, vol.23, no.6, 2011, pp.845-858.

[12] C. J. Keylock. "Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy", Oikos, vol.109, no.1, 2005, pp.203-207.

[13] Songbo Tan. "Neighbor-weighted k-nearest neighbor for unbalanced text corpus", Expert Systems with Applications, vol.28, no.4, 2005, pp. 667-671.

[14] C. J. van Rijsbergen, Information retrieval [M], London: Butterworths Press, 1979.

Jun Long was born in Anhui Province, China in 1972. He is a professor and Ph. D. supervisor of Central South University of Information Science and Engineering. His research interests include distributed computation, software architecture and trusted computing et al.

Luda Wang was born in Jinan, Shandong Province, China in 1981. He received the M.S. degree in Computer Application Technology from Hunan University in 2009, and he is currently a Ph. D. candidate of Central South University of Computer Science and Technology.

He has been a faculty member of Computer Science at Xiangnan University, Chenzhou, China, since 2005, where he is currently a lecturer. His research interests include AI, data analysis and information retrieval.

Zude Li is an Assistant Professor at Central South University (China). He obtained his Ph.D. degree in 2010 from The University of Western Ontario (Canada). His research interests are in the fields of software architecture, evolution and quality. He collaborates with IBM Canada and is appointed as a Software Engineering expert in SKANE SOFT.