Generating a multilingual taxonomy based on multilingual terminology clustering

Chengzhi ZHANG¹,² *

¹ Department of Information Management, Nanjing University of Science and Technology, Nanjing, 210094, China
² Institute of Scientific & Technical Information of China, Beijing 100038, China

Abstract Taxonomy denotes the hierarchical structure of a knowledge organization system. It has important applications in knowledge navigation, semantic annotation and semantic search. It is a useful instrument to study the multilingual taxonomy generated automatically under the dynamic information environment in which massive amounts of information are processed and found. Multilingual taxonomy is the core component of the multilingual thesaurus or ontology. This paper presents two methods of bilingual generated taxonomy: Cross-language terminology clustering and mixed-language based terminology clustering. According to our experimental results of terminology clustering related to four specific subject domains, we found that if the parallel corpus is used to cluster multilingual terminologies, the method of using mixed-language based terminology clustering outperforms that of using the cross-language terminology clustering.

Keywords Multilingual taxonomy, multilingual terminology clustering, cross-language terminology clustering, parallel corpus, mixed language

1 Introduction

Taxonomy or a conceptual hierarchy is the basic structure of a knowledge organization system. It has important applications in knowledge navigation, semantic annotation and semantic search. For example, information seekers can expand or narrow their searching parameters in order to yield the desired results according to this conceptual hierarchy. There are four kinds of taxonomy generating methods: 1) A manual method, 2) customization of one or more off-the-shelf taxonomies, 3)
adoption of a taxonomy engine, and 4) a combination of any of the above methods[1].

It is a challenging task to study the concept of an automatic formation of multilingual taxonomy under a dynamic information environment in which massive amounts of information are processed and found. Multilingual taxonomy is the core component of a multilingual thesaurus or ontology. The goal of this study is to find how to generate the multilingual taxonomy more effectively via the method of multilingual terminology clustering.

The paper introduces two methods for bilingual taxonomy generating: The cross-language terminology clustering and the mixed-language based terminology clustering. According to the experimental result of terminology clustering related to four subject domains, we found that when the parallel corpus is used to cluster multilingual terms, the method of using the mixed-language based terminology clustering is superior to that of applying the cross-language terminology clustering.

2 Related works

The related works include primarily the automatic generation of both the monolingual and the multilingual taxonomy.

2.1 Automatic generation of a monolingual taxonomy

The approaches of automatic monolingual taxonomy generation include: 1) Clustering approaches, 2) matching lexicon-syntactic patterns, 3) using machine-readable dictionaries to search relations among searchable terms[1] and 4) heuristic rules-based approaches. The clustering approach is based on the distributional hypothesis, which assumes that words with similar meanings will tend to occur in similar contexts[1–6]. The approaches of matching lexicon-syntactic patterns are used for identifying lexicon-semantic relations between the words. For instance, Hearst[7] identified several syntactic patterns to find hypernym/hyponym pairs in text corpora. Approaches based on machine-readable dictionaries get the initial hierarchy with WordNet or similar ontology, and deduce relations among concepts through user feedback[1,8]. One case of heuristic rules-based approaches is the occurrence frequency of two terms (such as A and B). That is used by Sanderson and Croft to create a taxonomy, which assumes if $P(A|B) \geq 0.8$ and $P(B|A) < 1$, then A is hypernym of B[9].

2.2 Automatic generation of a multilingual taxonomy

Compared with the automatic generation of the monolingual taxonomy, there are relatively scant records of published materials available about the generation of
Generating a multilingual taxonomy based on multilingual terminology clustering

Chengzhi ZHANG
Research Papers

multilingual taxonomy. Hopfield network and associate constraint network were used by Yang and Wei to establish cross-lingual synonym dictionary with Chinese-English parallel corpus[10–11]. Hjelm and Buitelaar employed hierarchical clustering to create taxonomy with comparable corpus, and their result shows that taxonomy generated by comparable corpus is better than that generated by monolingual corpus[12]. Tsuji and Kageura proposed that translation equivalents can be clustered to generate Japanese-English synonym dictionary with parallel corpus according to the graph theory[13]. Akiko and Kyo generated Japanese-English bilingual keywords clustering from academic databases, which are also based on the graph theory[14].

This paper is aimed to generate the multilingual taxonomy automatically. The object of clustering in this work is the core set of multilingual terminologies. In addition, the paper introduces two methods in generating taxonomy with a bilingual concept: The cross-language terminology clustering and the mixed-language based terminology clustering.

3 Generation of the multilingual taxonomy based on multilingual terminology clustering

This paper focuses on using Affinity Propagation clustering (AP clustering) algorithm to process clustered terminologies. The key issues involved are the computation of terminology similarity and the selection of a clustering algorithm. The flowchart of generating such a taxonomy is shown in Fig. 1.

This paper calculates semantic similarity among terminologies based on Tongyici Cilin and a large scale of corpus statistics. Finally, the comprehensive similarity is calculated according to literal, semantic and corpus-based similarity. AP clustering algorithm is used to process clustered terminologies.

3.1 Taxonomy generation based on multilingual terminology clustering

There are two kinds of multilingual terminology clustering strategies according to the different clustering objects: i.e., the one with a generation strategy of a
cross-language taxonomy (the taxonomy in different languages can be mapped) and the other, with that of a mixed-language taxonomy.

### 3.1.1 Taxonomy generation with a cross-language mode

A taxonomy with a cross-language mode means that a taxonomic system can map out an appropriate grammatical relation of its constituent languages to one another. As shown in Fig. 2, the Chinese and English taxonomies can be generated based on the Chinese and English terminology clustering, respectively. Then, the taxonomy in these two different languages can be merged based on the Chinese-English bilingual probability dictionary or based on the manual editing methods. Finally, we can get the cross-language taxonomy based on the merging results.

In this paper, *Tongyici Cilin* (i.e., Chinese synonym dictionary)*[15]* is used to compute the semantic similarity among Chinese terminologies*[16]*. A large scale of corpus is used to compute the distribution similarity among terminologies in the corpus*[17–18]*. According to the literal similarity between terminologies*[19]*, semantic similarity and corpus-based similarity, we can get the comprehensive similarity between terminology A and terminology B via the linear combination*[16]* as follows:

\[
\text{Sim}(A, B) = \alpha \cdot \text{Sim}_W(A, B) + \beta \cdot \text{Sim}_S(A, B) + \gamma \cdot \text{Sim}_C(A, B) 
\]  

(1)

Where, \( \text{Sim}_W(A, B) \) is the literal similarity between A and B, \( \text{Sim}_S(A, B) \) is the semantic similarity between A and B, \( \text{Sim}_C(A, B) \) is the corpus-based similarity (after normalizing). And \( \alpha, \beta, \gamma \) are the similarity weight for three kinds of similarity, respectively, and \( \alpha + \beta + \gamma = 1 \).

In these experiments, the Chinese terminology similarity is computed based on the literal, semantic and corpus-based similarity; the similarity among English
terminologies and the similarity between the English and Chinese terminologies are computed based on the parallel corpus.

The merging process of bilingual taxonomy is as follows. The similarity between each element in the target cluster and exemplar of the source cluster is computed. If the similarity is higher than the threshold value (the minimum similarity between each element and exemplar in the source cluster), then, this element in the target cluster can be regarded as the element in the source cluster. If the proportion of such elements in the target cluster is more than a given threshold, the target cluster can be merged into the source cluster.

3.1.2 Taxonomy generation with a mixed-language mode

The taxonomy in a mixed-language mode is the taxonomy which contains multiple language terminologies. As shown in Fig. 3, we used parallel corpus to calculate the similarity among different terminologies. Then, the set of terminologies in different languages are clustered based on AP clustering algorithm. Finally, we can get taxonomy in the mixed language based on the clustering results.

The corpus-based similarity between Chinese and English terminologies can be computed based on the parallel corpus by mutual information, Dice, LogL values and so on [16]. Similarity of terminologies in a mixed language can be computed based on the multilingual ontology, a semantic system, or the parallel corpus. In the experiments of this paper, we only use parallel corpus to compute the similarity between Chinese-English terminologies. The clustering results in a mixed language are the bilingual clusters, which include the Chinese and English terminologies.

3.2 Terminology clustering by using AP clustering algorithm

The common clustering algorithm includes the K-Means, a hierarchical clustering method, and a model-based clustering method and so on. Affinity Propagation (AP) clustering is a novel clustering algorithm that was proposed in Science in 2007 [20]. Earlier experiment results show that AP clustering algorithm can get better clustering result compared with classical clustering algorithms, such as the K-Means and the
hierarchical clustering algorithm. AP clusters the data points according to their degree of similarities. These similarities can be symmetrical, which means the degree of similarity between two data points is the same (e.g. Euclidean distance). They can also be unsymmetrical, which means the similarity between any two data points is not the same. These similarities form a matrix $S$ of $N \times N$ ($N$ denotes the total number of data points)\cite{21-22}.

AP clustering algorithm does not need to set clustering numbers in advance. On the contrary, all the data points could be served as potential clustering centers, which are called exemplars. The value $s(j, j)$ of diagonal line in the matrix $S$ is used to judge whether data point $j$ can be a clustering center. It means the larger the value is, the higher the possibility of this data point is to become a clustering center. The value can also be called preference ($P$), which can affect the number of clusters. If every data point is likely to be a clustering center, then $P$ should use the same value. If the value $P$ is derived from the average degree of input similarity, then the clustering number is of medium quantity. If the value $P$ is taken at the minimum, the resulting clustering numbers are fewer\cite{21-22}.

AP algorithm delivers two types of messages. Suppose there are two data points $i$ and $j$, $r(i, j)$ denotes the value from point $i$ to clustering center $j$. It reflects whether the point $j$ can be the center of point $i$ and $a(i, j)$ denotes the value from the point $j$ to $i$. It reflects whether the point $i$ can be the center of point $j$. The bigger the $r(i, j)$ and $a(i, j)$ is, the higher the possibility of point $j$ is to be the center. AP algorithm will consistently update these two kinds of information of every data point through the iterative process, until it generates $m$ high-quality exemplars. Meanwhile it will divide the remaining data points into corresponding clusters\cite{21-22}.

In the AP algorithm, the corresponding parameters are explained as follows\cite{21-22}:

Similarity: The degree of similarity between data point $i$ and $j$ is denoted as $S(i, j)$.

Preference: The reference to or preference of data point $i$ is called $P(i)$ or $S(i, j)$, which refers to the preference of data point $i$ as a point of reference to a clustering center. Usually, the average value or the minimum value of the similarity degree of $S$ is adopted.

Responsibility: $R(i, j)$ is used to describe the fitness of data point $j$ being the clustering center of data point $i$.

Availability: $A(i, j)$ is used to describe the fitness of data point $i$ being the clustering center of data point $j$.

Damping factor: $\lambda$ is a damping factor used to avoid numerical oscillations.
Generating a multilingual taxonomy based on multilingual terminology clustering

Chengzhi ZHANG

Research Papers

National Science Library,
Chinese Academy of Sciences

http://www.chinalibraries.net

4 Experiment result and analysis

4.1 Testing data and the dataset processing

This paper collects bibliographic data (e.g. metadata including titles and keywords of academic papers in both Chinese and English) in the subject domains of “Library and Information Science (LIS)”, “Law”, “Economics”, and “Languages”. Chinese and English keywords are used as testing datasets. Firstly, core terminologies are extracted from Chinese and English keywords[24]. Then, multilingual taxonomy is generated via multilingual terminology clustering.

The detailed information about testing datasets is shown in Table 1. The testing scale of “Library and Information Science” is the largest. In all of the Chinese and English keywords, the total number of different Chinese keywords is 25,728, and

\[ R(i, j) \text{ is computed by Formula (2)} \]

\[ R(i, j) = S(i, j) - \max \{A(i, k) + S(i, k)\} \]

\((k \in \{1, 2, ..., N, \text{ but } k \neq j\}) \]

\( (2) \)

\[ A(i, j) \text{ is computed by Formula (3)} \]

\[ A(i, j) = \min \{0, R(j, j) + \sum_{k} \max(0, R(k, j))\} \]

\((k \in \{1, 2, ..., N, \text{ but } k \neq i \text{ and } k \neq j\}) \]

\( (3) \)

The process of AP algorithm is as follows:
Firstly, similarity among \(N\) data points is computed. Secondly, this similarity value is put into the matrix \(S\) and a preference value is set (usually it is set one or two times of the median value). At the same time, a maximum value of iterations is set. Thirdly, values of \(R\) and \(A\) are computed in each iteration, and we can judge whether the data point is to be a clustering center (when \(R(i, j) + A(i, j)>0\), it is a cluster center). When the iterations exceed the maximum value (i.e. \(\text{maxits}\) value) or when a clustering center will not change after repeated iterations, the computational process is terminated[21].

This paper used AP clustering tools provided by Frey for clustering terminologies. An initial result of AP clustering was used to serve as a candidate taxonomy. We can obtain different numbers of clusters via adjusting the value of \(P\). Smaller values for the preference parameter yield fewer clusters, while larger values yield more. Thus, after clustering with the small preference value, we can generate the next hierarchy of conceptual taxonomy by clustering with the big preference value[23].
English, 41,729. The testing scale of “Languages” is the smallest. The total number of different Chinese keywords is 12,467, and English, 14,047.

<table>
<thead>
<tr>
<th>Domain ID</th>
<th>Domain</th>
<th>Chinese Library Classification number</th>
<th>Scale of parallel corpus</th>
<th>Total keywords</th>
<th>Keywords without repetition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>LIS</td>
<td>G25, G35</td>
<td>42,954</td>
<td>124,060</td>
<td>25,728</td>
</tr>
<tr>
<td>9</td>
<td>Law</td>
<td>D9</td>
<td>7,047</td>
<td>26,046</td>
<td>12,743</td>
</tr>
<tr>
<td>F</td>
<td>Economics</td>
<td>F</td>
<td>31,928</td>
<td>113,024</td>
<td>42,872</td>
</tr>
<tr>
<td>H</td>
<td>Languages</td>
<td>H</td>
<td>5,858</td>
<td>22,055</td>
<td>12,467</td>
</tr>
</tbody>
</table>

These bilingual core terminologies in the four subject domains are extracted. This paper takes the top-10,000 terminologies in terms of achieved scores in every subject domain as clustering objects\[24\]. In the experiments, we set 5 kinds of terminology subsets: top-100, top-200, top-300, top-500 and top-1,000 to generate a bilingual taxonomy. In the AP clustering algorithm, in order to get the minimum clustering results, we take the preference value as the minimum similarity degree value.

4.2 Evaluation methods

The evaluation methods of taxonomy include both the internal evaluation and the external evaluation. The former is a direct evaluation while the latter is an indirect evaluation. This paper uses the internal evaluation to evaluate the bilingual taxonomy. The major evaluation indicators include the net similarity of clusters\[23\] and the number of clusters.

A net similarity denotes the maximal value of the cost function in AP clustering\[23\]. In this paper, the net similarity is used as one of the evaluation indicators. The higher the net similarity is, the better its clustering performance is.

The number of subcategories of clustering results reflects the distribution of terminologies. In this paper, the number of exemplars is used as such evaluation indicators. The bigger the number of the exemplars is, the more scattered the distribution of the terminologies is.

4.3 Experiment results and analyses

We generated bilingual taxonomy in the subject domain of “Library and Information Science”, “Law”, “Economics”, and “Languages”, respectively. We try to explain the point from the perspectives of different terminology clustering methods and affecting factors of the cross-language clustering performance.
4.3.1 Clustering result analysis of different multilingual terminology clustering methods

In order to analyze the clustering results of the two kinds of clustering strategies, we define the similarity as parallel corpus-based similarity, which is computed by mutual information. The merging threshold value is set at 1.

In the four subject domains, all monolingual terminologies are set at quantitative categories of 100, 200, 300, 500 and 1,000, respectively. Cross-language terminology clustering is implemented after Chinese and English terminology clustering are completed. The mixed-language terminology clustering is implemented via mixing a given number of Chinese and English terminologies, computing the degree of similarity based on the parallel corpus and clustering terminologies in different languages.

The results of four subject domains are shown as Table 2, which include the degree of net similarity and the number of clusters of the Chinese, English, cross-language and mixed language. The result of Chinese and English terminology clustering shows how many Chinese clusters are combined with English clusters: The number of clusters decreases after some similar Chinese and English terminologies are combined into the same cluster. As a whole, the degree of the net similarity of bilingual taxonomy clustering based on a mixed-language mode is higher than that based on a cross-language mode, but the number of resulting clusters of the former is less than that of the latter.

The results also show that if the parallel corpus is used to cluster multilingual terminologies, the method of mixed-language based terminology clustering is better than that of the cross-language based terminology clustering.

It needs to point out that without the parallel corpus, clustering based on a mixed-language mode cannot generate multilingual taxonomy because it cannot compute similarities among terminologies in different languages. By comparison, the clustering based on a cross-language mode is superior. As long as there is a corpus or semantics system for each language involved, the degree of terminology similarity in that particular language can be computed. As a result, by using an external bilingual dictionary or a probabilistic bilingual dictionary, we can assign each possible translation term a probability value to indicate how likely the translation might turn out to be. We can also map out or merge terminology clustering results in different languages.

In a word, if we use parallel corpus to compute similarity of multilingual terminologies, the performance of multilingual taxonomy clustering based on a mixed-language mode is better than that based on a cross-language mode. With an external bilingual dictionary or a probabilistic bilingual dictionary, the mixed-
### Table 2 Terminology clustering results from 4 subject domains

<table>
<thead>
<tr>
<th>Categories</th>
<th>Clustering category</th>
<th>Net similarity</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LIS</td>
<td>Law</td>
<td>Economics</td>
<td>Languages</td>
<td>LIS</td>
<td>Law</td>
<td>Economics</td>
</tr>
<tr>
<td>100</td>
<td>Chinese</td>
<td>976.4647</td>
<td>1917.2438</td>
<td>*</td>
<td>1095.4644</td>
<td>24</td>
<td>26</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>1520.7656</td>
<td>1957.1853</td>
<td>2224.1824</td>
<td>1086.1710</td>
<td>23</td>
<td>27</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Cross-language</td>
<td>1126.9561</td>
<td>2033.8512</td>
<td>*</td>
<td>1123.9259</td>
<td>45</td>
<td>49</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>Mixed-language</td>
<td>2967.1406</td>
<td>3052.8347</td>
<td>3456.6028</td>
<td>2247.9729</td>
<td>49</td>
<td>40</td>
<td>48</td>
</tr>
<tr>
<td>200</td>
<td>Chinese</td>
<td>*</td>
<td>6941.5396</td>
<td>7925.4922</td>
<td>3939.4097</td>
<td>*</td>
<td>47</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>5702.5732*</td>
<td>9055.7344</td>
<td>7842.3325</td>
<td>*</td>
<td>46</td>
<td>55</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Cross-language</td>
<td>*</td>
<td>7483.4126</td>
<td>8578.4023</td>
<td>*</td>
<td>*</td>
<td>92</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Mixed-language</td>
<td>11071.3525</td>
<td>10431.9346</td>
<td>12509.3564</td>
<td>8077.0718</td>
<td>84</td>
<td>78</td>
<td>87</td>
</tr>
<tr>
<td>300</td>
<td>Chinese</td>
<td>19941.2207</td>
<td>16793.1504</td>
<td>20205.7988</td>
<td>9245.7383</td>
<td>78</td>
<td>77</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>19941.9707</td>
<td>18759.5059</td>
<td>19742.4629</td>
<td>9973.5361</td>
<td>71</td>
<td>77</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Cross-language</td>
<td>21631.1250</td>
<td>17640.6367</td>
<td>20400.0977</td>
<td>9859.2607</td>
<td>132</td>
<td>145</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>Mixed-language</td>
<td>*</td>
<td>22970.8203</td>
<td>28978.9838</td>
<td>16509.6367</td>
<td>*</td>
<td>111</td>
<td>125</td>
</tr>
<tr>
<td>500</td>
<td>Chinese</td>
<td>37385.9023</td>
<td>48848.9102</td>
<td>60569.9063</td>
<td>26805.6094</td>
<td>109</td>
<td>131</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>40640.8086</td>
<td>57087.8555</td>
<td>58765.8242</td>
<td>28775.5371</td>
<td>100</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Cross-language</td>
<td>42688.9805</td>
<td>50547.1875</td>
<td>62438.5977</td>
<td>28195.1035</td>
<td>189</td>
<td>247</td>
<td>242</td>
</tr>
<tr>
<td></td>
<td>Mixed-language</td>
<td>72078.1797</td>
<td>63655.1602</td>
<td>*</td>
<td>44530.8086</td>
<td>204</td>
<td>175</td>
<td>*</td>
</tr>
<tr>
<td>1000</td>
<td>Chinese</td>
<td>218520.5781</td>
<td>283823.5625</td>
<td>262151.9375</td>
<td>118912.7656</td>
<td>225</td>
<td>285</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>*</td>
<td>297318.6250</td>
<td>264577.7500</td>
<td>147548.7969</td>
<td>*</td>
<td>258</td>
<td>272</td>
</tr>
<tr>
<td></td>
<td>Cross-language</td>
<td>*</td>
<td>277761.8750</td>
<td>258419.0781</td>
<td>129804.8594</td>
<td>*</td>
<td>494</td>
<td>510</td>
</tr>
<tr>
<td></td>
<td>Mixed-language</td>
<td>353323.5000</td>
<td>340132.1250</td>
<td>336053.5313</td>
<td>191058.8438</td>
<td>316</td>
<td>348</td>
<td>374</td>
</tr>
</tbody>
</table>

**Note:** * denotes that AP clustering algorithm is not converged with the current parameter.
Generating a multilingual taxonomy based on multilingual terminology clustering

Chengzhi ZHANG
Research Papers

National Science Library,
Chinese Academy of Sciences
http://www.chinalibraries.net

4.3.2 Affecting factors of the cross-language terminology clustering

(i) The impact of similarity computing  In this experiment, we used the terms in the subject domain of “Library and Information Science” as the testing dataset. The similarity computing includes three basic methods: The literal-, the semantic- and the corpus-based methods. According to different weighing scales and a combination of these methods, we generated different cross-language clustering results. The resulting clustering number of terminologies is totaling 500 and the merging threshold value is set at 1.

Table 3 shows different similarity algorithms affecting the cross-language clustering. From Table 3 we can see that the clustering number increased in the following order: semantic-based < literal-based < corpus-based.

<table>
<thead>
<tr>
<th>Similarity computing</th>
<th>Clusters of Chinese taxonomy</th>
<th>Clusters of cross-language taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literal</td>
<td>68</td>
<td>131</td>
</tr>
<tr>
<td>Semantic</td>
<td>29</td>
<td>123</td>
</tr>
<tr>
<td>Corpus</td>
<td>109</td>
<td>189</td>
</tr>
<tr>
<td>Literal + semantic (The weight is 0.3 and 0.7, respectively)</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>Literal + corpus (The weight is 0.7 and 0.3, respectively)</td>
<td>35</td>
<td>63</td>
</tr>
<tr>
<td>Semantic + corpus (The weight is 0.7 and 0.3, respectively)</td>
<td>29</td>
<td>124</td>
</tr>
<tr>
<td>Literal + semantic + corpus (The weight is 0.1, 0.8 and 0.1, respectively)</td>
<td>38</td>
<td>114</td>
</tr>
<tr>
<td>Literal + semantic + corpus (The weight is 0.2, 0.7 and 0.1, respectively)</td>
<td>46</td>
<td>129</td>
</tr>
<tr>
<td>Literal + semantic + corpus (The weight is 0.2, 0.6 and 0.2, respectively)</td>
<td>48</td>
<td>132</td>
</tr>
</tbody>
</table>

Different combinations of similarity algorithms can generate different clusters. If we only consider two kinds of similarity, the combination of literal and semantic-based similarity will generate relatively fewer clusters. If all of them are considered and if the semantic-based similarity computing is also added with a greater weighing scale, the resulting number of clusters will be even much less.

It needs to point out that we need to do further research to ascertain whether different numbers of clusters generated by different similarity computing methods can be actually used to judge the clustering result. Furthermore, since there are more algorithms and other resources for the measurement of the similarity of the target language, it seems natural that we will get a better result from such a taxonomic system than that of the original source language. We will, in our future work, explore
how to combine the clustering results from the original source language with those from the target language via a cross-language mapping.

(ii) The merging threshold value We took the subject domain of “Library and Information Science” as the testing dataset, the corpus-based similarity as similarity of terminologies and had the number of terminologies set at 500. We studied the impact of clustering result in different merging thresholds. Table 4 shows the different numbers of clusters when the merging threshold value is set at 0.5, 0.6, 0.7, 0.8, 0.9 and 1, respectively. It also shows that the smaller the merging threshold value is, the fewer clusters are. When the threshold value is set at a higher level (the maximum is 1), the resulting number of combined clusters is less. Thus, in our study, except for the parameter set (e.g. the preference value in AP clustering), we can control the final results via setting a merging threshold value.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Net similarity</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>21172.1641</td>
<td>121</td>
</tr>
<tr>
<td>0.6</td>
<td>30034.4453</td>
<td>141</td>
</tr>
<tr>
<td>0.7</td>
<td>32891.2305</td>
<td>162</td>
</tr>
<tr>
<td>0.8</td>
<td>36234.5781</td>
<td>181</td>
</tr>
<tr>
<td>0.9</td>
<td>42688.9805</td>
<td>189</td>
</tr>
<tr>
<td>1.0</td>
<td>75034.6797</td>
<td>210</td>
</tr>
</tbody>
</table>

5 Conclusion and future works

This paper explores multilingual taxonomy clustering from two kinds of strategies: the cross-language clustering and the mixed-language based clustering. According to the experiment results of terminology clustering related to four subject domains, we found that if the parallel corpus is used to cluster multilingual terminology, the mixed-language based terminology clustering is more productive than the cross-language terminology clustering. However, under the condition of not having a parallel corpus, the mixed-language based terminology clustering cannot work if it only depends on a machine translation system or a bilingual dictionary. Instead, the cross-language terminology clustering method can be used to generate the multilingual taxonomy.

Our future research work will move toward the following directions. First of all, we will study the possibility of integrating different languages into an integral taxonomy. Then, we will explore how we can generate the taxonomy of another language for its enhanced performance through such an integrated taxonomy. Furthermore, our future research efforts will also be dedicated to the study of the generation of a language taxonomy with multiple hierarchical levels and to integration.
Generating a multilingual taxonomy based on multilingual terminology clustering

Chengzhi ZHANG

Research Papers

of different multilingual taxonomy systems so as to improve the quality of their performance. Secondly, we will strive to develop an evaluation method for the multilingual taxonomy so as to compare it with the existing taxonomy of a given subject discipline or the portion of it with a machine-generated taxonomy. We will also try to evaluate the multilingual taxonomy from the perspective of its empirical application. Finally, we will combine many semantic resources (e.g. WordNet) with a large scale of corpus (e.g. WWW-based corpus) to improve the rate of computing accuracy with regard to terminology similarity.

References


22 AP Clustering Algorithm. Retrieved on December 31, 2009, from http://hi.baidu.com/river_hiter/blog/item/a5ae8c08fa34e02d6a60fbd7.html/cmtid/86b188a0f0a87b8746106413.


(Copy editor: Ms. Lin PENG, Language revision: Prof. Charles C. Yen)