Multi-Classification of Theses to Disciplines Based on Metadata

Jianling Li[1,1], Shiwen Yu[1], Shasha Li[1], Jie Yu[1]

[1]School of Computer Science, National University of Defense Technology, Changsha 410073, China
jianlingl@mail.nwpu.edu.cn,{yushiwen14,shashali,yj}@nudt.edu.cn

Abstract. Thesis classification is fundamental to a wide range of efficient research management. Current thesis classification is limited to major, research direction and classification number manually labeled by students themselves, which lacks standard and accuracy. Furthermore, previous auto-classification studies do not take account of interdisciplinary. This study intends to make a major contribution to Chinese thesis classification by taking advantage of the metadata such as title, keywords in the thesis. We propose a novel hierarchical classification model based on methods in metadata semantic representation and the corresponding similarity calculation. Experiments on 4K+ Theses show our methods have significant effect.

Keywords: thesis metadata semantic representation; similarity calculation; multi-classification

1 Introduction

The classification of academic papers has been existing for ages. Depending on the classification, people hope to store, search, manage and study academic papers more efficiently. Currently, theses in China are labeled in various granularities，such as “major”, “research direction” and “Chinese Library Classification Number”. However, these labels have at least three problems unsolved:

- Research institutions often disagree on the domain and the naming of a major. For example, School of Computer Science in Northwestern Polytechnical University has four departments, named “Computer Systems and Microelectronics”, “Computer Science and Software”, “Computer Information Engineering” and “Information Security and Electronic Commerce Technology”. However, School of CS in National University of Defense Technology has “Computer Science and Technology”, “Software Engineering”, “Electronic Science and Technology” and “Cyberspace Security”.
- There is no strict rule for students to fill in those labels. For example, during the preprocessing of metadata, we find it is not rare for students to write the wrong “Chinese Library Classification Number” in their theses.
- Current theses classifications have not considered interdisciplinary.

According to the ***Classification and Code of Disciplines***[1], the thesis showed in Fig.1 involves at least three disciplines— "Radar Engineering", "Wireless Communication Technology" and "Military Information Engineering and Information Countermeasure Technology", any of which is reasonable, but no one can replace another.

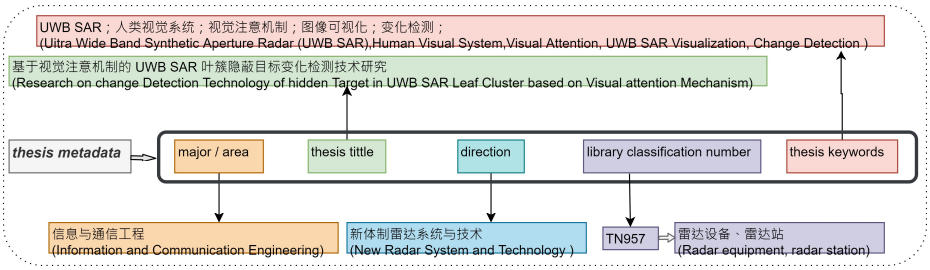

Fig. 1. Example of metadata of a thesis

The 2018 theses statistical report by ***WanFang Data***[2] shows that about 97.8% of the theses included by it are written in Chinese. Therefore, we focus on Chinese theses classification. In China, two taxonomies are the most popular and commonly used—***Chinese Library Classification*** and ***Classification and Code of Disciplines.*** After investigation, we believe the ***Classification and Code of Disciplines*** are more suitable for theses classification [1]. The reasons are: 1, it has a clearer structure and is easier to use; 2, it is internationally accepted; 3, it is designed for academic purpose while the other is designed for general books. An example of ***Classification and Code of Disciplines*** is shown in Fig.2—the second and third level disciplines under "Computer Science" (which itself is a first-level discipline).

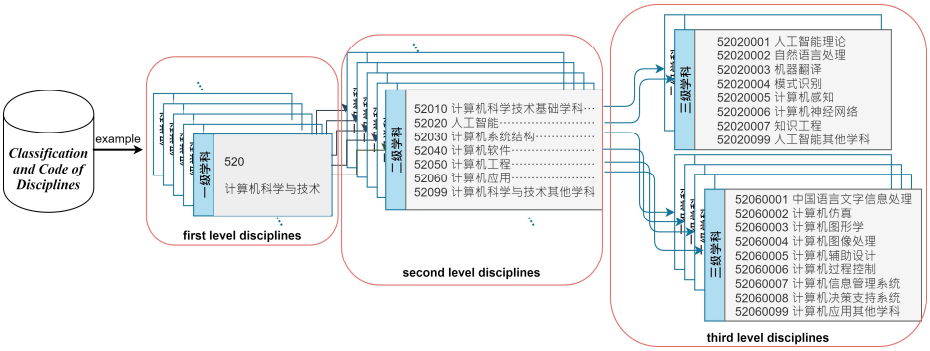

Fig. 2. example of Classification and code of disciplines

Previous studies were focused on journal papers [1]. Researches on multi-

[1] The version we are using is GB/T13745-2009. This taxonomic hierarchy is divided into three levels: first-level disciplines, second-level disciplines, and third-level disciplines.

[2] Wanfang Data is one of the most popular knowledge service platforms in china.

classification of theses to disciplines are certainly inadequate considering the importance of this field.

Although current labels in theses metadata are not the direct goal we desire, they can be of great use to the multi-classification task. Our multi-classification system is based on the semantic representation of metadata of theses. The main contributions are:

a) We proposed several methods for high-level semantic representation by weighted splicing low-level semantic representations. The weights are determined by the relevance between the characters (, words or phrases) of the metadata term and the central words[3].
b) We proposed a novel method for similarity measure of vectors with different lengths. This method is based on cosine similarity and especially aimed at the weighted spliced semantic vectors.
c) An open-source multi-classification software system of theses to disciplines, available at http://git.trustie.net/jianlingl/thesis_muti-classification.git

2 Related work

In natural language processing, semantic representation is crucial for tasks such as word disambiguation, similarity calculation, and analogy reasoning. Popular representation methods include one-hot representation based on bag of words, distribution-based representation based on counting, and word embedding representation by expressing words as dense low-dimensional real-value vectors [2]. But in Chinese, things are more complicated. Characters in Chinese may have multiple meanings. As a result, centralized models for ambiguity of words is proposed, such as semantic modeling that considers the position information of characters together with their multiple meanings [3], modeling that focus on multi-meaning words [4], modeling that is powered by HowNet [5] , modeling that considers Chinese character component (radical) [6], etc. these methods are dependent on semantic dictionary, which makes them time-consuming, laborious and difficult to scale.

As mentioned above, the metadata of theses are useful to the multi-classification task. But theses metadata are usually phrases. For English, A phrase semantic representation is composed of the semantic representations of words, where the traditional compositional model is faced with low accuracy and data sparsity. Consequently, the model which learns the representations of words and phrase simultaneously [7][8] and model which extends the semantic are proposed. Inspired by previous works and focused on our task and language, we proposed several methods for high-level semantic representation by weighted splicing low-level semantic representations.

The similarity in natural language can be divided into semantic similarity and distribution similarity. The former is based on the similarity of cognitive taxonomy,

[3] We define the title and keywords of a thesis as its central words.

and the latter is based on the similarity of the topic [10]. We focus on semantic similarity. The existing similarity calculation methods include edit distance, Jaccard similarity coefficient, cosine similarity, TF-IDF coefficient, similarity measure method of vectors with different lengths—Distance correlation, etc. Chinese phrase text similarity measure methods consider the position of the same word in the phrase text [11]. Inspired by previous works, we proposed a novel method for similarity measure of vectors with different lengths, which is especially aimed at our weighted spliced semantic vectors.

Multi-classification of text can be of great use in scientific research and application. The earliest text multi-classification system appeared in 1999 as an automatic classification system for e-mail. It mainly uses information entropy theory and Bayes algorithm to realize multi-classification of text. While the researches of multi-classification of Chinese text started rather late, the methods of which are mainly based on the similarity comparison in semantic vector space.

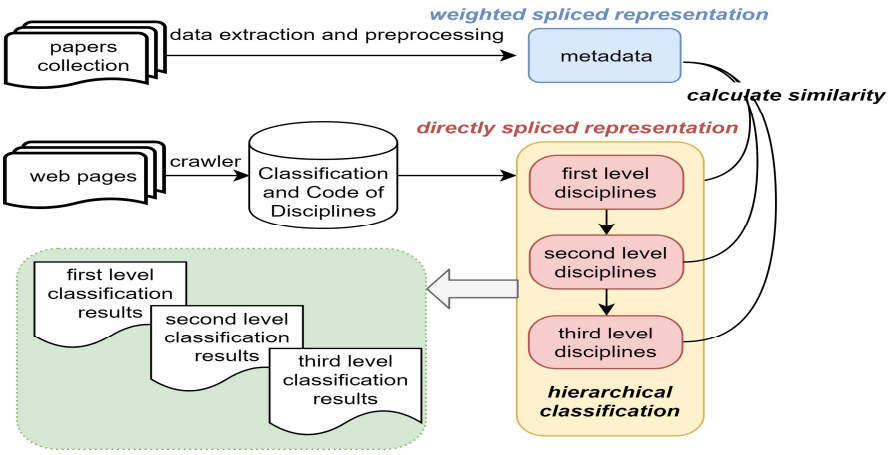

Fig. 3. Model Framework.

The framework of our multi-classification system is shown in Fig.3, where the semantic representations of the metadata are weighted spliced with low-level representations (words or characters) and the representations of the discipline phrases are directly spliced with low level representations. By calculating the similarity between the paper metadata and the discipline phrases, we can obtain all three level discipline classifications.

3 Model

There are four parts in our multi-classification model: semantic representation of theses metadata, semantic representation of discipline phrase, similarity measure method and hierarchical classification algorithm of theses.

3.1 Semantic representation of theses metadata

We represent theses metadata as vectors, which splice the weighted semantic vectors of the character (, words or phrases) of the metadata term. The weights are determined by the relevance between the characters (, words or phrases) and the central words because the influences of every character (, word or phrase) for the meaning of metadata term are usually not the same. We quantize influences of the characters (, words or phrases) by the similarity between them and the central words. In our work, the central words are defined as the title and keywords of the thesis.

For example, In thesis “具有深度信息的视频图像中的人物步态识别技术研究”(“Research on Character Gait Recognition Technology in Video Images with Depth Information”), the research direction of the thesis “图形与图像处理技术”(“graphics and image processing technology”) contains very obvious features for classification. In our model, we give character “图”(“image”) a higher weight than character “形”(“shape”) since “图” is more relevant to the central word of this thesis— the title. The same story happens between words “图像”(‘image and picture’) and “图形”(“graphics”), in which the former is more relevant.

After given the weights, the semantic vectors of characters (or words) are spliced into one vector representing the metadata term. Algorithm details are shown in pseudocode below.

algorithm 1 weighed low-level representation splice for metadata

input: *metadata*，*central_words*

output: *metadata representation*

```
function METADATA_REP(metadata, central_words)
    MetadataEm ← []
    for low_level_structure in metadata and not in StopWords do
        weigh ← GET_WEIGH(low_level_structure, central_words)
        Em ← weigh * W2Vmodel[low_level_structure]
        MetadataEm.append(Em)
    end for
    return MetadataEm
end function

function GET_WEIGH(low_level_structure, central_words)
    Weigh ← 0.00001
    try :
        Weigh ← Similarity(low_level_structure, central_words)
    return Weigh
end function
```

Semantic vectors of characters and words are obtained by word embedding model which is already implemented in gensim (a python library). We also build a list of stop-words to ignore noisy characters and words. The parameter *central_words* is the key to weights determination. We chose the titles and keywords of the theses as *central_words*.

The proposed algorithm can not only be used to represent metadata but also be used to represent any hierarchical compositional semantic representation. For instance, the semantic representation of a paragraph can be weighted spliced with

the semantic vectors of sentences. The only thing needs making efforts is to determine the weights of the low-level terms according to the task.

3.2 Semantic Representation of Discipline Phrase

This algorithm is slightly different from that in the representation of theses metadata. Instead of weighting characters and words and splicing their semantic vectors, we simply splice these vectors. For example, we splice vectors of "图像" ("image") and "处理" ("processing") to represent the third-level discipline "图像处理" ("image processing").

During experiments, we realize that second-level disciplines often have abundant semantic information. It is not easy for us to represent them only by splicing their characters or words. So, we splice together all the third-level disciplines under this second-level discipline into a long-phrase and use the same splicing method to obtain the semantic representation of the long-phrase.

3.3 Similarity measure method

Previous phrase similarity measure methods are not suitable for multi-classification of theses to disciplines. Therefore, we proposed a novel method for similarity measure of vectors with different lengths based on cosine similarity and specially aimed at the weighted spliced semantic vectors. The similarity measure method can be described by the equation below:

$$Sim\_(mdEmb, dpEmb) = \frac{\sum_{i=0}^{lm-1} \left(\max_{0 \le j \le ld-1} sim(mdEmb[i], dpEmb[j]) \right)}{lm} \quad (1)$$

In the equation, $mdEmb$ is the semantic vector of the metadata; $dpEmb$ is the semantic vector of the discipline; $mdEmb[i]$ is the i-th low-level semantic vector (the vector of a character or word) of the metadata; $dpEmb[j]$ is the j-th low-level semantic vector of the discipline; lm is the number of low-level semantic vectors, which together compose into the metadata; ld is the number of discipline's low-level semantic vectors; $sim()$ in the right part is the cosine similarity function. If the vector embeddings of the metadata and discipline are the same, the similarity value calculated by this equation will be 1. This equation calculates two high-level semantic representations by taking the average of the maximum of the similarity values of their low-level semantic vectors.

3.4 Hierarchical classification Algorithm of Thesis

Because ***Classification and Code of Disciplines*** has obvious hierarchical characteristics, we use Hierarchical classification in our system. We first sort all first-level disciplines by the similarity between them and metadata of the thesis. Since the metadata are multiple, we accumulate all similarities between the discipline and all metadata as the similarity. Then, instead of traversing all the second-level disciplines,

we only consider those whose parent disciplines are among the top-N first-level disciplines sorted before, where N is defined by user interests. We define N as 3 in our experiments. The top-N second-level disciplines are obtained in the same way as the first-level, and so do the third-level disciplines. Hierarchical classification is timesaving and very helpful in noise filtering; it can also give us a systematic classification result. details are shown below:

algorithm 2 Hierarchical Classification of theses

input: $theses$, $1st\_level\_disciplines\_all$, $TopN$

output: $1st\_level\_disciplines$, $2nd\_level\_disciplines$, $3rd\_level\_disciplines$

function HIERARCH_CLASSIFY($theses$, $1st\_level\_disciplines\_all$, $TopN$)
    $1st\_level\_disciplines \leftarrow get\_disciplines(theses, grade1disciplines, TopN)$
    $2nd\_level\_disciplines\_subsets \leftarrow get\_2nd\_level\_disciplines\_subset(1st\_level\_disciplines)$
    $2nd\_level\_disciplines \leftarrow get\_disciplines(theses, 2nd\_level\_disciplines\_subsets, TopN)$
    $3rd\_level\_disciplines\_subsets \leftarrow get\_3rd\_level\_disciplines\_subset(2nd\_level\_disciplines)$
    $3rd\_level\_disciplines \leftarrow get\_disciplines(theses, 3rd\_level\_disciplines\_subsets, TopN)$
  return $1st\_level\_disciplines$, $2nd\_level\_disciplines$, $3rd\_level\_disciplines$
end function

4 Experiment

We extracted all metadata from the theses corpus and all discipline phrases in the ***Classification and Code of Disciplines***, together with their hierarchical structures, into a database. The metadata we concern are the title, the keywords, the area, the major, the research direction, the Chinese Library Classification Number and the degree type.

4.1 Dataset

The dataset is built with 4146 theses published between 2013 and 2017 in National University of Defense Technology. These theses cover a wide range of topics, such as Aerospace, Computer Science, Electronic Information, Control Science, and Management Science.

Training set.
All 4146 master theses are in the training set. The preprocessing of theses is simple: first split the text of the thesis into characters and words separately, then ignore special characters and those in the list of stop-words (or characters). We use ***jieba*** (a python library) to help us split text into words. This data set is used to train the word2vec models.

Test set.
We randomly selected 190 theses from the training set and manually labeled them by 2 volunteers independently. Each thesis is labeled with multiple first-level,

second-level and third-level disciplines. After the labeling, another volunteer decided the final labels of the theses on which the 2 volunteers disagreed. We evaluated the consistency between the 2 volunteers and found they agreed on 89.01% first-level disciplines, 86.54% second-level disciplines and 88.89% third-level disciplines. At last, we got 307 first-level, 457 second-level and 497 third-level manually labeled disciplines in the 190 theses.

4.2 Evaluation

The results of multi-classification of theses to disciplines are shown in table 1, where the R is the recall rate and MAP (mean average precision) is used to deal with limitations of point estimation. The MAP algorithm is shown below:

algorithm 3 MAP for TopN results
input: $labeledDisciplines$, $classifiedDisciplines\_TopN$
output: MAP

```
1:  function MAP_TopN(labeledDisciplines, classifiedDisciplines_TopN)
2:      for labeleldata in labeledDisciplines do
3:              score_ ← 0
4:          if labeleldata in classifiedDisciplines_TopN then
5:                  score_ ← 1.0/(classifiedDisciplines_TopN.index(labeleldata) + 1)
6:                  hitCount ← hitCount + 1
7:          end if
8:          Score ← Score + score_
9:      end for
10:     if hitCount bigger than 0 then
11:         return Score/hitCount
12:     else
13:         return 0
14:     end if
15: end function
```

labeledDisciplines are manually labeled disciplines, *classifiedDisciplines* are the predicted disciplines from our system. *classifiedDisciplines_TopN.index(x)* will return the ranked place number of *x* in the predicted disciplines. For each manually labeled discipline, if it is near the top of the disciplines given by our system, then we will give it a high score; If it is near the bottom, we give it a low score; If it is not in the classification results, we ignore it. In the end, we take the average of all scored disciplines as the final score. For example, if a manually labeled discipline is in the second place of the predicted disciplines, it will be scored as 0.5.

For the recall calculation, if one of the top-N predicted disciplines is the same as one of the manually labeled disciplines, we say the prediction is true positive. In this paper, we did several controlled experiments and listed the results in table 1.

The MAP value can reflect the ranking accuracy of the correct disciplines in the prediction results, which means that the higher MAP the more accurate the prediction is. As shown in the table, the semantic representation of metadata we proposed has brought a significant improvement. ***Character based weighted splicing method*** is better than ***words based weighted splicing method. Phrase based weighted splicing method*** gets the best R and MAP score in the first-level classification, which we believe is because the metadata terms of the theses are often in phrase form.

Table 1. multi-calssification results of theses based on different methods(xxx_R means recall of xxx-level disciplines classification,xxx_MAP means MAP of xxx-level disciplines classification)

Racall and MAP / *Metadata_rep+ similarty*	1st_R	1st_MAP	2nd_R	2nd_MAP	3rd_R	3rd_MAP
character based accumulating+ W2V.n_Similarity	75.79%	53.38%	-	-	-	-
*Character based splicing +**Sim_***	84.21%	68.07%	-	-	-	-
Word based accumulating+ W2V.n_Similarity	69.51%	50.24%	-	-	-	-
*Word based splicing+ Sim_*	73.68%	55.66%	-	-	-	-
Character based weighted splicing+ Sim_	92.11%	67.81%	85.78%	59.74%	81.14%	54.14%
Word based weighted splicing+ Sim_	89.77%	57.42%	80.10%	50.30%	73.25%	46.90%
Phrase based weighted splicing+ Sim_	**94.21%**	***72.85%***	88.01%	68.71%	78.63%	***56.81%***
Phrase based weighted splicing*+ Sim_	92.63%	67.46%	***89.54%***	***68.84%***	***80.41%***	56.01%

Phrase based weighted splicing method* is composed with two steps: represent phrase semantic by weighted splicing characters, and then represent metadata by weighted splicing phrases. This method gets the best scores in the second-level and the third-level classifications. In practice, we can use ***Phrase based weighted splicing method*** in the first-level classification and use ***Phrase based weighted splicing method**** in the rest. A prediction instance is shown in Fig.4.

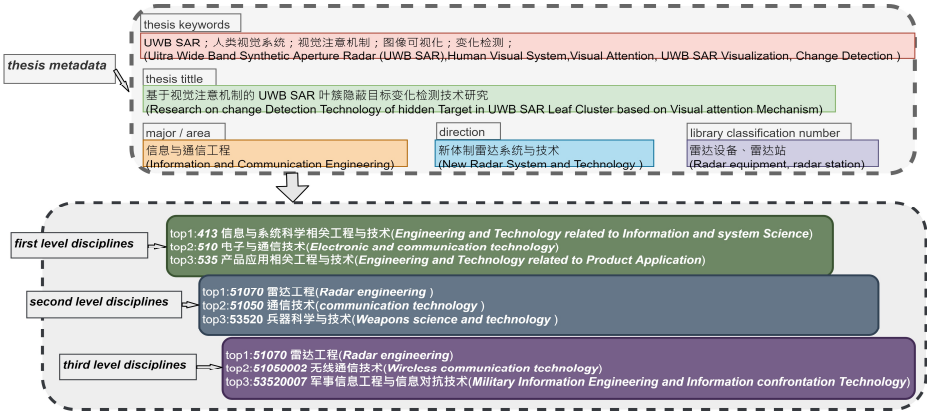

Fig. 4. Thesis multi-classification example

5 Conclusion

Multi-classification of theses to disciplines is of great use in scientific research and application. Following the ***Classification and Code of Disciplines***, we proposed a multi-classification system based on metadata and semantic embedding. A significant improvement in this task has been brought by our weighted splicing algorithms,

of which the extensibility is also promising. Considerably more work will need to be done to this field.

Acknowledgements

The research is supported by the National Key Research and Development Program of China (2018YFB1004502) and the National Natural Science Foundation of China (61532001, 61303190).

Reference

1. 邱均平, 赵岩杰, 罗力. 科学评价中的论文分类方法研究[J]. 情报学报, 2011, 30(5).
2. Mikolov T , Sutskever I , Chen K , et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013.
3. Chen X , Xu L , Liu Z , et al. Joint learning of character and word embeddings[C]// International Conference on Artificial Intelligence. AAAI Press, 2015.
4. Xinxiong Chen, Zhiyuan Liu,Maosong Sun.A Unified Model for Word Sense Representation and Disambiguation[C].//Conference on empirical methods in natural language processing, vol. 2: Conference on empirical methods in natural language processing (EMNLP 2014), 25-29 October 2014, Doha, Qatar.2014:1025-1035.
5. Xie R , Yuan X , Liu Z , et al. Lexical Sememe Prediction via Word Embeddings and Matrix Factorization[C]// Twenty-Sixth International Joint Conference on Artificial Intelligence. AAAI Press, 2017.
6. Sun Y , Lin L , Tang D , et al. Radical-Enhanced Chinese Character Embedding[J]. Lecture Notes in Computer Science, 2014, 8835:279-286.
7. Mikolov T , Sutskever I , Chen K , et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013.
8. Hashimoto K , Tsuruoka Y . Adaptive Joint Learning of Compositional and Non-Compositional Phrase Embeddings[J]. 2016.
9. Passos A , Kumar V , Mccallum A . Lexicon Infused Phrase Embeddings for Named Entity Resolution[J]. Computer Science, 2014.
10. Utsumi A , Suzuki D . Word Vectors and Two Kinds of Similarity[C]// International Conference on Acl. DBLP, 2006.
11. 王莹莹, 任贤, 龙鹏飞. 中文短语文本相似度计算新方法[J]. 软件导刊, 2011, 10(1):79-81.
12. 王莹莹. 中文短语相似度计算方法研究及应用[D]. 长沙理工大学