A Possible Optimization Of Search Engine Base On Dependency Tree's Representation

Zhanfu Yang, Haonan Zhang and Chenghuan Huang, *Sun Yat-Sen university*

***Abstract*—we present a considerable method to optimize the result of the Chinese search engine, which uses the optimized dependency tree's representation to analyse syntactic of search content. The main contribution is the combination model between the dependency tree's representation and the reverse index of the Chinese search engine so as to optimize the index and the result's correlation. After optimized the search engine which based on the Wikipedia Chinese, it gain correlation between the result and the searching content, sometime it can also search some ralated articles while the source code finds nothing.**

Keywords*— *Search Engine, Dependency Tree, Syntactic Analysis, Segmantic Role Labeling.

I. Introduction

Totally speaking, in order to present a possible optimized method, our framework mainly contains the technology of semantic analysis and the basic search engine. In detailed, it includes the following 4 aspect:

A. Chinese Tree-bank

Lately, Chinese dependency tree-bank is very popular, many universities and companies keep perfecting Chinese Tree-bank, including Harbin Institute of Technology, Peking University and University of Pennsylvania. As we all know, Chinese syntactic analysis is far more puzzle than the English one, Although Pennsylvania is the first institution which build the Chinese Tree-bank(dating back to 1986), it 's English analysis structure can not completely adapt to Chinese sentence such as Coo(sharing-object coordinate) and Cos(non-sharing-object coordinate) , Harbin Institute of Technology and Peking University presented many methods to optimize the Chinese tree-bank, including Peking's Dependency Tree Representation of Predicate-Argument Structures and HIT 's LTP-Cloud which is an free dependency parsing tree-bank platform providing open source API for developers' designing. In this passage, I would build a web bases on html adopting HIT 's API so as to analyze the structure of a sentence.

B. Syntactic Analysis based on dependency depressing

Dependency depressing, a method based on the tree-bank, reveals the syntactic structure of a sentence by analyzing the dependence relation among the components in the language unit.

Intuitively speaking, Dependency depressing parses and recognizes the grammatical constituents of a sentence with tags like "SVO" and so on. Besides, it analyzes the relationship among the various components at the same time. Give an example as shown in figure 1, the core predicate of the sentence is "Ti Chu", it 's subject is "Li Ke Qiang" and it's object is "Zhi Chi Shang Hai ...","Li Ke Qiang" 's adjunct is "Guo Wu Yuan Zongli","Zhi Chi" 's subject is "Tan Suo". With the above analysis 's result, we can easily found that the action's agent is "Li Ke Qiang" rather than "Shang Hai" even it's closer to "Ti Chu".

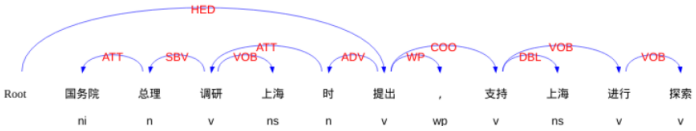


Fig. 1. A Dependency depressing example of a sentence. In this picture, "Zong Li" is the subject and the "Shang Hai " is the object.

C. Semantic Role Labeling

Semantic role labeling, sometimes also called shallow semantic parsing, is a task in natural language processing consisting of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. In this passage, we may adopt the optimized semantic method combining with the Chinese search engine.

D. Chinese Search Engine

Search engine is a kind of special lattice point in the network, which can search the information in the web server. As there are a large amount of search engine 's users in China, the quality and efficiency of the search engine becomes so important that it may affect the number of users.

Reverse index is one of the core method in the search index. Human 's large amount of researches based on the reverse index are mainly among Three filed. The first one is the compression algorithm. The second one is optimization of the the storage structure 's designing. The final one is the combination of the reverse index and the specific participle algorithm.

In this passage, we put forward a possible method point at increasing the quality and efficiency of the search engine by changing it's reverse index.

E-mail of Zhanfu Yang 865031716@qq.com.
Thanks the help of every partner

II. Model

A. Optimized semantic role labeling

Semantic role labeling can always promote the development of information extraction. Based on dependency tree-bank, an optimized Semantic Role Labeling can acquire more accuracy of the information extractionAcco. rding to Likun Qiu 's related work, we can get a better semantic role labeling methods base on his transmission model. In this passage, we will adopt the semantic role labeling method depend on the following list of semantic roles.

Label	Description
ADV	adverbial, default tag
BNE	beneficiary
CND	condition
DIR	direction
DGR	degree
EXT	extent
FRQ	frequency
LOC	locative
MNR	manner
PRP	purpose or reason
TMP	temporal
TPC	topic
CRD	coordinated arguments
PRD	predicate
PSR	possessor
PSE	possessee

TABLE I.
APPENDIX WITH LIST OF SEMANTIC ROLE WHICH MAY USE IN THE MODEL.

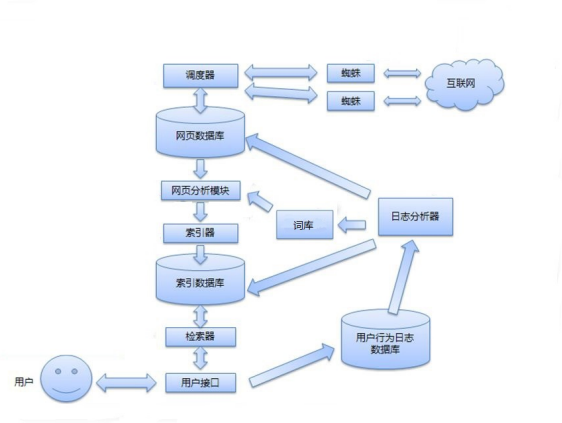


Fig. 2. A Dependency depressing example of a sentence. In this picture, "Zong Li" is the subject and the "Shang Hai " is the object.

B. Combination of Chinese search engine and optimized dependency tree-bank

Search engine is based on a strict schedule, as shown in figure 2. In a search engine, it mainly includes 4 part. They are Index Manager, Index Searcher, Indexer and Document Manager. When we are going to build the index, we need to search the passage. There are totally two main method to search for the whole passage, the first one is full text scanning and the other one is build the index. Since we are living in big data time, the article and files we need to search in a database some time may be very big, now we usually use inverse index to search for the whole passage in order to remain searching speed even when we search for more files.

Originally, the reverse index in the Chinese search engines often adopt n-some constant structure which bases on Maximum likelihood estimation as the core method.

$$p(w_i|w_1w_2...w_{i-1}) = \frac{C(w_1w_2...w_i)}{C(w_1...w_{i-1})}.$$

Give bi-gram algorithm(an algorithm bases on N=2) as an example, as shown in figure 3, the bi-gram algorithm will analyze the consistent two Chinese word and evaluatMaximum likelihood estimatione the possibility of it's appearance. With this method, it has a serious problem, N-gram model is basedon each other without any genetic properties of discrete element words constructed, which does not have a continuous space of word vectors to satisfy the semantic advantages. Similar meaning words have similar word vector, so when the system model for a certain word or word sequence adjustment parameters, the similarity of words and word sequences will also change.

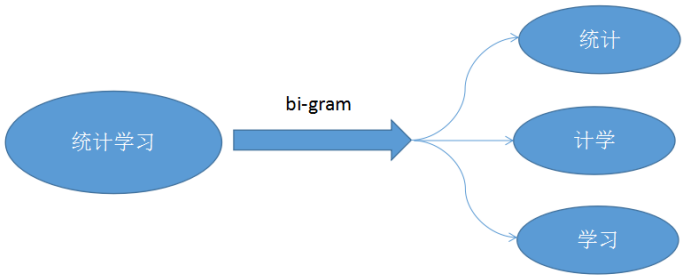


Fig. 3. An example about the participle result of sentence "Tong Ji Xue Xi", bases on the bi-gram algorithm

Different from the N-gram algorithm, Dependency tree can represent the structure of a sentence more precise and reasonable, so we want to combine the Dependency tree 's representation with the search engine like the way bi-gram algorithm adopts. All we need to do is to replace bi-gram with new algorithm when it comes to participle.

In our method, we firstly extract the word element from the sentence bases on the dependency tree's representation that is to save the position and extracted message(files id) from the files. We use Wikipedia 's entries as our database. Since those entries are all represented in the style of [UTF-8], we have to consider about the features of the character encoding. In the UTF-8, it represent a character with a length between one and four. In order to avoid the trouble in dealing with the UTF-8, we transform UTF-8 into UTF-32.

After we finish extraction, what we do is to create a reverse table for every lemmas, and add those table into the small reverse index, after the index reach a size, combine the index

in the storage together.

When it comes to the search process (given in the figure 4). In this way, we may finally get a list of result after we input the query we want to search.

Algorithm 1 Framework of dealing with search process.

Require:
The already finished reverse index bases on the privious work.

Ensure:
1: Divide the query into several words element.
2: Sort the divided words element in ascending order, according to the occurrence times in all files.
3: Acquire the reverse table of each words element, get the file's id and the list of position that the words element appears.
4: If all the words(or just parts of)element are appears in the same file, put this file into the result of the search.
5: Calculate the correlation between query with the file in the list of result.(in this method we use TF-IDF to calculate the correlation.)
6: Sort the Calculation value of the result in descending order.
7: **return** Take a number of files in the sorted list as the result to return.

III. Experiment

In this part, we will try to introduce our coding work based on the theme.

As the API of the semantic analysis is available in the LTP-Cloud platform. we make a html to get the participle result based on the HIT 's dependency tree-bank(given in the figure 4). Input a query into the html, what we get is a sequencial json/xml return which can be used in the program.

Fig. 4. The html platform we made to get the semantic analysis.

Hence we regard the participle as the words element and put in the rectified search engine, which bases on Wikipedia. Firstly, Download the articles from the https: // dumps.wikimedia.org/wiki/zhiki/20160701/ . Secondly, build the reverse index bases on the xml files. Next, We Input the query we want to search. Finally, we will get the result of files name in a list which is descending sorted. Give figure 5 as an example.

Compare with the original search engine, sometimes the original one can not find the article with key word "Tong Ji Xue Xi Fang fa", but we can find come related article in the new one simply because the new article divide the query much more better. Further judgement we will not discuss in this passage, we may adopt further work to completely finish our research.

```
yzf@yzf-ThinkPad-T440p:~/Downloads/IR/IR/trunk$ ./wiser -q "人工智能" wikip
2000.db
[time] 2016/07/20 08:30:15.000005
document_id: 117 title: 人工智能 score: 2023.918304
document_id: 8 title: 计算机科学 score: 105.474906
document_id: 956 title: 中国国家自然科学基金学科分类目录/F score: 75.820301
document_id: 827 title: 围棋 score: 69.544228
document_id: 520 title: Emacs score: 55.635383
document_id: 999 title: 中国学科分类国家标准/520 score: 41.726537
document_id: 9 title: Wikipedia:繁简分歧词表 score: 37.187017
document_id: 34 title: 心理学 score: 31.391158
document_id: 49 title: 克利斯登·奈加特 score: 27.817691
document_id: 53 title: 理查德·斯托曼 score: 27.817691
document_id: 109 title: 计算语言学 score: 27.817691
document_id: 170 title: 浙江省 score: 21.055779
document_id: 192 title: 逻辑 score: 18.448365
document_id: 951 title: 中国国家自然科学基金学科分类目录/A score: 17.482313
document_id: 73 title: 计算机程序 score: 13.908846
document_id: 147 title: 黑客 score: 13.908846
document_id: 584 title: Wikipedia:删除纪录/存档 score: 13.908846
Total 17 documents are found!
```

Fig. 5. An example about the result of the search engine with the query about "Ren Gong Zhi Neng".

IV. Future work

In this research, we just put two method together simpley withhout much deeper work, what we can do is mainly in the following two sides:

First, we can do more work and research to study the quality of the new method, including searching for the same database and query to compare their search result.

Second, try to add method to correct the wrong words search by the users which may be a good direction to gain the quality of the search engine.

V. Conclusion

We studied the open source search engine and the sytactic structures of the dependency tree's representation, presenting a combined search engine's model which may contribute to the correlation between the search and the results. We make our html platform which gained the dependency tree's representation and the originally source code based on "How to Develop a search engine" freely available at https://github.com/peterrich/Information-Retrive.

Acknowledgment

Thanks our teacher's wonderful information retrieval classes and the help of some seniors students. Thanks the help of teaching assistance. Thanks the instruction of the book named "How To Develop a Search Engine". Thanks the LDT=Cloud to provide open source API for developers.

References

[1] Likun Qiu, Yue Zhang and Meishan Zhang. *Dependency Tree Representations of Predicate-Argument Structures*. aaai2016, China.

[2] H Yamada, *How to Develop a Search Engine*. Japen, 2016.

[3] Wan Ling, Yang Xiudan, Du Xiaojing. *Evaluating Standard for Chinese Search Engine*. QingBaoKeCue, China, 2000.

[4] MA Jian, ZHANG Taihong, CHEN Yanhong. *New inverted index storage scheme for Chinese search engine*. Journal of Computer Applications, China, 2013.

[5] Anne Abeille,ed. *Treebank: Building and Using Parsed Corpora*. Kluwer Academic Publishers, 2016.

[6] Wang Jingjiang. *Comparative Study on the Evaluation Index Systems for search Engine*. Peking University, China, 2008.

[7] RamakrishnanG,BalakrishnanS,JoshiS.*Entity annotation based on inverse index operations*. Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2006:492-500.

[8] Burt C, John E. *An Analysis of Chinese Search Engine Filtering[J].*. Computer Science, 2011, 12(2):117-127.

[9] Wu J, Zhu H, Zhu H.*Systems and methods for translating Chinese pinyin to Chinese characters*. US, US 7478033 B2[P]. 2009.e. H Yamada , *How to Develop a Search Engine*. Japen,2016.

[10] Xu Y, Ringlstetter C, Goebel R, et al. *A Continuum-based Approach for Tightness Analysis of Chinese Semantic Units[J].*. Paclic, 2009.

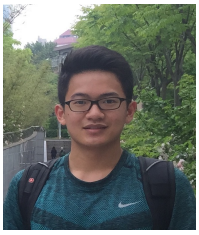
[11] Lepage Y, AuclercAmp N, SHIRAISatoshi. *A TOOL TO BUILD A TREEBANK FOR CONVERSATIONAL CHINESE[C].* The Proceedings of the 6(th) International Conference on Spoken Language Processing (Volume). 2000.

[12] Liu H, Zhao Y, Li W. *Chinese Syntactic and Typological Properties Based on Dependency Syntactic Treebanks[J].* . Pozna Studies in Contemporary Linguistics, 2009, 45(4):509-523.

[13] Harbin Institute of Technology. *ltp-cloud*. china, http://www.ltp-cloud.com/.

[14] Liu H, Huang W, Liu H, et al. *A Chinese Dependency Syntax for Treebanking[C]*. Asia, International computer's meeting. 2006.

[15] Xia F, Han C H, Palmer M, et al. *Comparing Lexicalized Treebank Grammars Extracted from Chinese, Korean, and English Corpora[C]*. Chinese Language Processing Workshop. 2002.



Zhanfu Yang Sophomore in Sun Yat-Sen university, gained much awards during the second Year in school including the first prize and application optimization award of the Asian students supercomputing 's competition,The first prize of IBM national campus's competition. One of the member in sysu-software igem team.

Haonan Zhang Sophomore in Sun Yat-Sen university, has a high ability of designing algorithm and coding, loves to design game, especially big and exciting games.

Chenghuan Huang Sophomore in Sun Yat-Sen university, gained much awards during the second Year in school including the first prize and application optimization award of the Asian students supercomputing 's competition, keen on coding and programing.