# Research on Organizational Knowledge Structure's Construction Based on Text Mining

**Jiangnan Qiu[1] Chuangling Nian[1]**

[1]School of Management, Dalian University of Technology, Dalian, China
*Email: qiujiangnan@gmail.com, chuangling.nian@gmail.com*

**Abstract**

Organizational knowledge structure is a reflection of organizational knowledge system. In this paper, we study organizational knowledge structure's construction based on patent documents using text mining method. Its construction process is as follows: Firstly, patent documents of organization are collected and preprocessed. Secondly, key terms are extracted from patent documents as knowledge elements using a merge algorithm, and synonymous terms are merged using a Chinese thesaurus - TongYiCi CiLin and pattern matching. Finally, similarity between terms is counted based on co-occurrence of terms, and hierarchical relationships between terms are built using hierarchical agglomerative clustering algorithm (HAC), and then organizational knowledge structure is formed through visualization.

**Keywords**: organizational knowledge structure, text mining, patent documents, knowledge elements, HAC

## 1. Introduction

Organizational knowledge structure is a reflection of organizational knowledge system, which reflects organizational knowledge's basic composition and relationships between different knowledge elements. Organizational knowledge structure that plays the role of a metadirectory for effective knowledge organization and access. It contains information on what is available in the knowledge base, how knowledge is related to each other, and where knowledge resides [1]. It contributes to not only measuring organization's knowledge stock and knowledge gap, but also acquiring technology development trend of competitors by analysing their knowledge structure. Therefore, construction and visualization of organizational knowledge structure is very important and necessary for organization's decision making and development.

Currently, there are a few researches on organizational knowledge structure, but these researches are mostly qualitative descriptions of organizational knowledge structure and lack research on its construction and visualization. The related researches include visualization of individual and group' knowledge structure and organizational knowledge structure's measurement model. Literature [2] studys visualization of scientific researchers' knowledge structure using knowledge network based on papers published by them. Literature [3] presents weighted knowledge network model of individual and group' knowledge structure. Literature [4] presents a organizational knowledge structure's measurement model which uses directed and acyclic graph to show organizational knowledge structure. The researches on individual and group' knowledge structures lay the foundation for the construc-

tion of organizational knowledge structure.

The paper studys the construction of organizational knowledge structure based on patent documents using text mining method. Knowledge elements are automatically obtained from organization's patent documents, and relationships between knowledge elements are built to form organizational knowledge structure. Finally, the experiment based on patent documents published by some organization in LED field is conducted to validate the construction method of organizational knowledge structure.

## 2. Organizational knowledge structure' s construction process

Constructuring organizational knowledge structure is to organize knowledge in a variety of different knowledge carriers (material carriers and organizational members carriers) of organization to form knowledge structure of organization. In the actual research, constructuring organizational knowledge structure based on material carriers is an available way because of the difficulty to formalize, code and acquire tacit knowledge in the mind of organizational members. Material carriers of orgnization include patent documents, audio and video medias, softwares, databases, reports and so on. Patents are the most important data source of commercial operation, scientific research and technological development. Patent documents contain organizational significant research achievements. At the same time, patents as important symbol and reflection of technical innovation, largely represent the technical level and potential technology competitiveness, so patent literatures are the organization's core knowledge resources and the most important material carriers. Therefore, the paper constructures and visualises organizational knowledge structure based on patent documents.

Organizational knowledge structure's specific construction process is as follows: Firstly, patent documents of organization are collected and pre-processed. Secondly, key terms in the pre-processed patent documents are extracted using a merge algorithm, and synonymous terms are merged using a Chinese thesaurus — TongYiCi CiLin and pattern matching. Finally, relationships between terms are extracted using hierarchical agglomerative clustering algorithm (HAC), and then terms and relationships between them together are visualised to form organizational knowledge structure. Overall process is as shown in Figure 1.

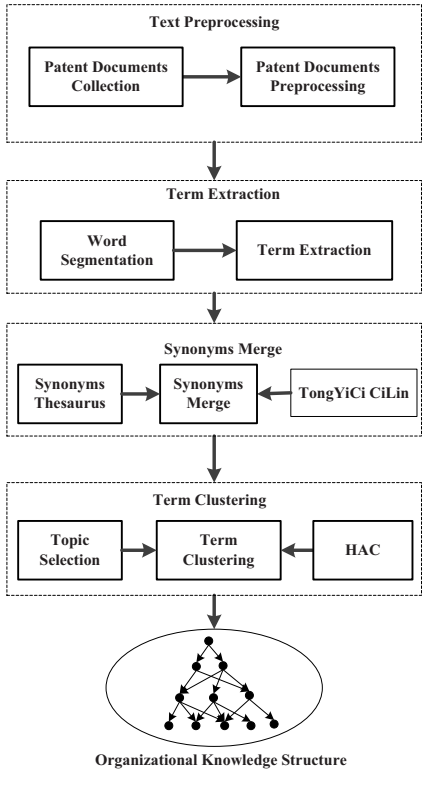


Fig. 1: Overall Process.

### 2.1. Patent documents's collection and preprocessing

Organizational knowledge structure's construction based on patent documents needs to collect all the patent documents which belong to the organization. We obtain all patent documents whose applicant is the given organization, and then preprocess them. Patent documents have a fixed structure, which provides convenience for knowledge discovery and analysis. However, patent documents is too long and some parts of them are difficult to identify and deal with using computers, so we select the most important five parts of patent documents, including *Invention Title, Abstract, Technology Field, Background Technology, Invention Content*. The collection of patent documents after presprocessing is formalized as: $D = \{d_1, d_2, \cdots, d_N\}, i = 1,2,\cdots,N$ , $d_i$ represents a patent document.

### 2.2. Term extraction

After patent documents of organization are collected and preprocessed, key terms which reflect technology knowledge of patent are extracted as the knowledge elements. At present, there are many term extraction's methods, such as: Pat_Tree method [5], N-gram algorithm [6], LDA (Latent Dirichlet Allocation) algorithm [7], etc. We use a merge algorithm to extract key terms. This algorithm is improvement and optimization of the keyword extraction algorithm by Yuen-Hsien Tseng et al [8] without using any dictionaries and can be used in all fields. The algorithm is based on one fact: If a word is a keyword, its each atom word's appearance frequency is greater than or equal to the frequency of the keyword in a patent document. The algorithm repeatedly merges back nearby words whose frequency is greater than a threshold to form a longer term, because long term is more likely to express precise meaning than short term. For example, Chinese term "照相机组件" has more precise meaning than "组件". If the frequency of "照相机组件" is equal to the frequency of "组件", the algorithm only extracts "照相机组件" as a candidate. Only if the frequency of "组件" is greater than the frequency of "照相机组件", the algorithm extracts both of them as candidates. The term extraction algorithm is described as follows:

**Input:** The original file of patent document and the POS tagging file;
**Output:** Key terms list *FinalList*;
**Step 1:** All words and their POS in POS tagging file are stored in a linked list named List.
**Step 2:** Computer each word's frequency. If its POS is noun (n) or verb (v) or adjective (a) or x (x represents unknown POS), the word's frequency is calculated, otherwize its frequency is assigned 0. In addition, word's frequency is respectively added a threshold T、A、D、C according to its appearance part in *Invention Title, Abstract, Technology Field, Background Technology, Invention Content*.
**Step 3:** Merge Process is in Figure 2.
**Step 4:** Filter the candidates of *FinalList* based on two rules:
1) Delete verb, the terms whose POS is @ and terms whose length is 1.
2) Reserve longer terms: If term1 literally contains term2 in FinalList and Freq(word1)==Freq(word2), term2 will be deleted.

All terms extracted from patent documents are considered as knowledge elements. Collection of all terms is formalized as:

$T = \{t_1, t_2, \cdots, t_n\}, i = 1,2,\cdots,n$ .

```
Do Loop
   Set MergeList to empty.
   For i =1 to NumOf(List)
   If Freq(List[i])>=t and Freq(List[i+1])>=t, Then
      Merge List[i] and List[i+1] intoTempWord.
      If List[i+1].POS is n or x, Then
            Set TempWord. POS to n.
      Else
            Set TempWord. POS to @.
            Put TempWord at the end of MergeList.
   Else
       If Freq(List[i-1])<t and Freq(List[i])>=t and Freq(List[i+1])<t , Then      //t is a threshold.
          Put List[i] into FinalList.
      Else
            If MergeList is not empty and the last element of MergeList is not X, then    // X is a seperator.
                  Add X into the end of MergeList, and set its frequency to 0.
    End of For Loop.
    List=MergeList
Until NumOf(List)<2
```

Fig. 2: Merge process for Step 3.

### 2.3. Synonyms merge

Organizational knowledge structure does not contain redundant knowledge elements, so it is necessary to merge synonymous terms together after terms are extracted from patent documents. If two terms $t_i$ and $t_j$ is synonymous, $t_j$ is deleted and $t_i$ becomes $t_i$ ($t_j$).

Currently, the main methods to identify Chinese synonyms include: 1) literal similarity algorithm based on single character; 2) literal similarity algorithm based on morpheme; 3) semantic similarity algorithm based on TongYiCi CiLin; 4) the algorithm based on words co-occurrence analysis; 5) the recognition method based on pattern matching [29]. We use "TongYiCi CiLin" and pattern matching method to identify synonyms.

We discover two kinds of synonyms pattern through analysing patent documents.

1) Chinese Term (English phrase or initials), for example, 阴极射线管 (CRT).

2) English words (Chinese words), for example, LED (发光二极管).

We extract synonyms manually because of less synonyms in each patent document, and then merge synonymous terms in the collection of terms.

### 2.4. Term clustering

Organizational knowledge structure shows hierarchy. Currently, semi-automatic methods to extract hierarchical relationships or taxonomic relationships between concepts include pattern-matching, concept clustering, dictionary and association rules mining [9]. We select concept clustering to extract hierarchical relationships between terms.

We use the hierarchical agglomerative clustering algorithm [10]. The similarity between two terms needs to be calculated before conducting clustering. Co-occurrence relationship between two terms is used to weigh correlation between them. We use the modified Dice Coefficient to measure the similarity between them.

Existing literatures studying co-occurrence between terms often consider a document as co-occurrence window, which can't inaccurately reflect correlation between them. Therefore, co-occurrence window is reduced to a sentence. If two terms co-occur in a sentence, they are considered as correlation according to words co-occurrence model [10]. Dice Coefficient is modified as follows:

$$Dice(t_{ij}, t_{ik}) = \frac{2 \times S(t_{ij} \cap t_{ik})}{S(t_{ij}) + S(t_{ik})} \quad (1)$$

$t_{ij}, t_{ik}$ represent two terms. $S(t_{ij} \cap t_{ik})$ represents the number of sentences in which $t_{ij}$ and $t_{ik}$ co-occur in a patent document $d_i$. $S(t_{ij})$ represents the number of sentences in which $t_{ij}$ occurs in patent document $d_i$. $In(1.72 + S_i)$ is added in order to avoid too big deviation between Dice Coefficient calculated in patent documents with different length because Dice Coefficient calculated in long document is lower than in short document.

The similarity between $t_j$ and $t_k$ is the sum of Dice Coefficient between $t_j$ and $t_k$ in all patent documents:

$$Sim(t_j, t_k) = \sum_{i=1}^{n} Dice(t_{ij}, t_{ik}) \tag{2}$$

The specific process of term clustering using HAC is as follows:

**Step 1:** Each term of terms collection $T$ is seen as a single cluster. $C_i = \{t_i\}, i = 1, 2, \cdots, n$. Set a variable $k = n$;
**Step 2:** Find out two clusters $C_i$ and $C_j$ between which the similarity is the largest, and merge them into a new cluser, then delete $C_i$ and $C_j$ and set $k = k - 1$;
**Step 3:** Find out the topic of each cluster which can represent the cluster;
**Step 4:** If $k > 1$, then turn into Step2, otherwise, end the procedure.

Calculating the similarity of two clusters use average value method, e.g. the similarity of two clusters equals the result of the sum of similarity of element pairs of two clusters divided by the product of the lengths of the two clusters.

Each time a new cluster generates in clustering process, it's necessary to select the topic of every cluster. For $C_i = \{t_{i1}, t_{i2}, \cdots, t_{im}\}$, Score of each element of cluster is calculated to weigh every element of $C_i$, then select the element whose Score is highest.

$$Score(t_{ij}) = \sum_{i=1}^{m} Sim(t_{ij}, t_{il}) * Freq(t_{ij}) \tag{3}$$

After extracting the topic of the cluster, the cluster's topic is seen as the parent node, other elements of cluster are its children nodes, thus forming a hierarchical tree that represent organizational knowledge structure.

## 3. Experiment

We conducted experiments to confirm the feasibility of our proposed method. The following considers all patent documents issued by an organization as the organization's knowledge resource to build the organizaiton's knowledge structure.

We retrieve certain organization's patent documents from Shanghai Intellectual Property (patent information) Platform. 27 pieces of patent documents published by the organization from 2003 to 2009 were collected as the organization knowledge resource. Then, the above method is used to build knowledge structure. Firstly, we choose five parts of patent documents of the organization and convert from PDF to TXT. Secondly, words segmenting and pos tagging are conducted on patent documents using word segmentation software ICTCLAS30 of Chinese Academy of Sciences, and then key terms are extracted using merge algorithm. Table 1 shows all key terms extracted from a patent document, the data following the terms is its frequency in the patent document.

| |
|---|
| 物体 33 |
| LED 50 |
| 单元 41 |
| 柔性 15 |
| 装置 41 |
| 照相机 60 |
| PCB 19 |
| 信号 16 |
| 连接器 33 |
| 部件 17 |
| 照相机组件 15 |
| 图象 62 |
| 通信终端 42 |
| 铰链单元 20 |
| 移动通信终端 41 |
| 照相机铰链单元 18 |
| 图象捕捉装置 43 |

Table 1: Key Terms Extracted from a Patent Document

For key terms extraction, it is necessary to adjust the frequency threshold in order to extract the key word more accuratly. In all, 247 key terms are extracted. After synonymous terms are merged, 244 terms are considered as knowledge elements of the organization. The hierarchical relations between them are extracted using HAC. Finally, the organzation's knowledge structure visualised by Pajek visualization software is as shown in Figure 4. Figure 3 is a part of the organization's knowledge structure.

The organizational knowledge structure visualised is validated by several LED experts who we invited, result is reasonable and above knowledge structure constructuring method is proved effective.

## 4. CONCLUSIONS AND OUTLOOK

Organizational knowledge structure plays an important role on organization's decision making, so it's very necessary to build organizational knowledge structure. This paper studies the specific process of constructuring organizational knowledge structure using text mining method based on patent documents. Firstly, organization's patent documents are collected and pre-processed. We select five parts of patent documents: Invention Title, Abstract, Technology Field, Background Technology, Invention Content. Secondly, key terms are extracted from patent documents as knowledge elements using a merge algorithm based some filter rules which has high accuracy and don't require artificial selection. Then synonymous terms are merged using TongYiCi CiLin and pattern matching method. Finally, terms clustering are conducted using HAC and topic of each cluster is selected to form a hierarchical tree. The trees are visualized as organizational knowledge structure.

However, organizational knowledge structure constructured based on patent documents is only the core part of organizational knowledge structure, but not the whole structure. Our method need further improvement and optimization in order to constructure more complete organization's knowledge structure.

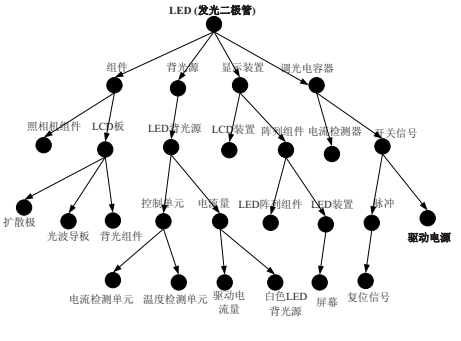


Fig. 3: A part of organizational knowledge structure.

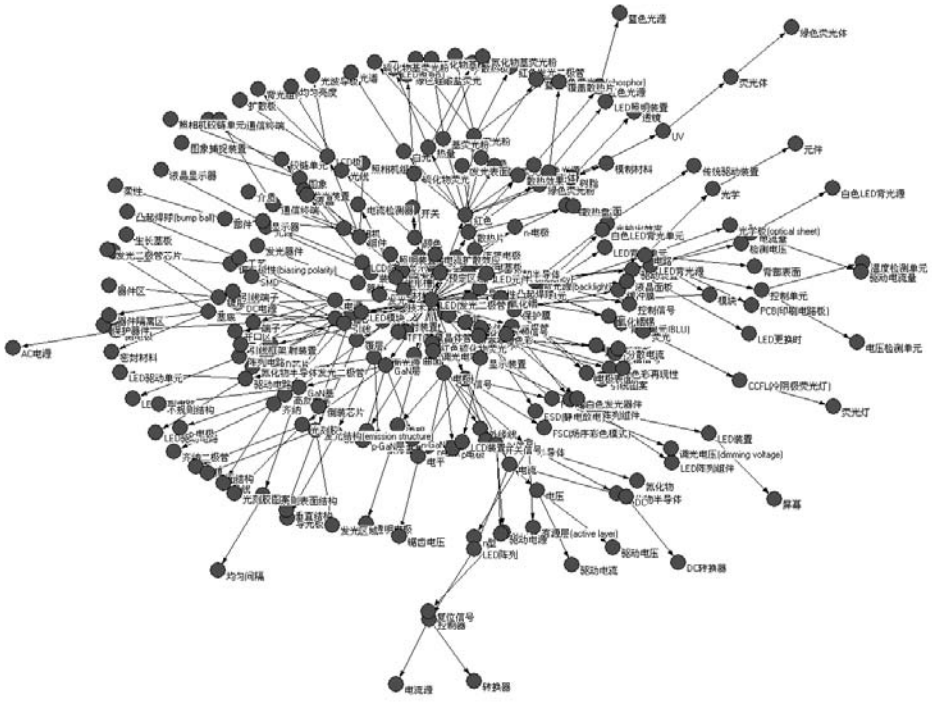

Fig. 4: Organizational knowledge structure.

## References


[1] Sulin Ba, Karl R. Lang, Andrew B. Whinston. Compositional Enterprise Modeling and Decision Support[J]. Information Systems and E-Business Management, 2008, 6(2): 137-160.

[2] Sun Jian-bin1, Zhang Peng-zhu. Visualization of Researcher Knowledge Structure Based on Knowledge Network [J]. Information Science, 2010, 28(3): 395-399,480.

[3] Xi Yunjiang, Dang Yanzhong. A Method to Analyze the Knowledge Structure of an Individual or a Group Based on Weighted Knowledge Network [J]. Journal of Industrial Engineering and Engineering Management, 2008, 22(3): 1-4,35.

[4] Wang Jun, Fan Zhiping. A Model for Measurement of Organizational Knowledge Based on Acyclic Digraph [J]. Systems Engineering, 2002, 20(5): 22-27.

[5] Kuanjiu Zhou, Lei Wang and Peng Qiu. Ontology Construction from Semi-Structured Text[C]. IEEE,2008: 1-4.

[6] Zhong Yunyun, Hou Hanqing, Du Huiping. Automatic Construction of Thesauri for e-Government [J]. Journal of Library Science In China, 2008(3): 97-102

[7] Jian-hua Yeh, Naomi Yang. Ontology Construction Based on Latent Topic Extraction in a Digital Library[C]. Proceedings of the 11th International Conference on Asian Digital Libraries, 2008: 93–103.

[8] Yuen-Hsien Tseng, Chi-Jen Lin, Yu-I Lin. Text mining techniques for patent analysis[J]. Information

Processing and Management, 2007(43): 1216–1247.
[9] Jia Xiuling, Wen Dunwei. A Study on Taxonomic Relation Extraction from Ontology Learning [J]. Computer Technology and Development, 2007, 17(10): 31-33,36
[10] Gen Huantong, Cai Qingsheng, Yu Kun et al. A Kind of Automatic Text Keyphrase Extraction Method Based on Word Co-occurrence [J]. Journal of Nanjing University (Natural Sciences), 2006(2): 156-162.