

Automatic Text Summarization of Chinese Legal Information

Dmitry Lande¹[0000-0003-3945-1178], Zijiang Yang², Shiwei Zhu², Jianping Guo²
and Moji Wei²

¹ Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine

² Information Research Institute of Shandong Academy of Sciences, Jinan, China
dwlände@gmail.com

Abstract. Article is devoted to a method of automatic text summarization of the legal information provided in Chinese. The structure of the abstract and the model of its formation is considered. Two approaches are suggested. First one is determination of weight of separate hieroglyphs instead of words in the texts of documents and abstracts for sentences importance level determination process. Second approach is to consider a model of document as a network of sentences for detection of the most important sentences by parameters of this network. Various methods of automatic text summarization are performed and tested. A cosine measure and Jensen-Shannon divergence are applied as two estimates of quality of the paper abstracts without participation of experts. Compared to other summarizing methods, given one on the basis of the suggested network model of the document was the best by criteria of a cosine measure and Jensen-Shannon's distances for abstracts which volume exceeds 2 sentences. The suggested approach, with minimal modifications, can be applied to texts on any subject of scientific, technical or news information.

Keywords: Automatic Text Summarization, Legal Information, Chinese Language, Cosine Measure, Jensen-Shannon Divergence.

1 Introduction

Processing of natural languages practically began with statement of problems of the artificial translation and automatic text summarization. The first fundamental works on automatic text summarization appeared in the middle of the last century [1].

The task is connected with the solution of the most important problem – reduction of the volumes of information consumed by the person, fight against information noise. This task is very relevant today due to the constant growth of the information space. Automatic referencing is known to all users of network search engines – in response to the request they receive not only the title of documents, but also their short automatically created descriptions (snippets). Mobile users want to see a brief description of the articles before they go on to read more. Persons who make impor-

tant management decisions have to familiarize themselves with thousands of documents a day, deliberately dismissing information noise.

Now there are hundreds of industrial systems of automatic text summarization, for example, such packages as Office Word AutoSummarize, Mac OS X Summarize, IBM Tivoli Monitoring Summarization and Pruning Agent, Oracle Text, plug-ins for browsers Chrome, Mozilla.

Numerous approaches to automatic text summarization are known, recently, neural network technologies, deep training are applied more and more widely. There are also numerous linguistic approaches associated with automatic parsing of sentences submitted in different languages. Traditional type of systems of automatic text summarization – extractive (quasi summarizing) at which the paper consists from of the separate, sometimes poorly connected among themselves sentences of the initial document. He is succeeded by abstractive type of text summarizing at which the systems close to the systems of artificial intelligence in a short form retell contents of the initial document "by the own words".

However, it should be noted that today still practically all industrial systems of automatic text summarization belong to extractive systems.

It would seem, the subject of automatic text summarization of texts is already rather studied, the main results are received. However, and in this article it is about creation of system of automatic text summarization.

There are several reasons for development of new system of automatic text summarization. First the problem of automatic text summarization of legal information is solved. And it is texts which can't fully be considered free, unstructured. There is a structure of separate types of documents and use of the best universal systems of summarizing doesn't yield satisfactory results. Secondly, the authors deal with the texts of documents presented in Chinese, which significantly narrows the range of possible ready-to-use systems. For processing Chinese texts, as a rule, segmentation of words is required – in the Chinese language words are often not separated by separators.

Thirdly, the program capable in corporate system to process big data flows with an acceptable productivity and quality, built in the existing system of document flow has to be developed.

Besides, retelling of documents in this case is unacceptable. Any "imaginings", liberties of retelling by the computer of legal acts it isn't admissible. Exit one – to develop some hybrid algorithm and, respectively, the program of extractive type capable to consider features of legal acts of the People's Republic of China. At the same time the program has to be capable to process separate documents which unite to big documentary arrays. This program has to be capable to allocate obviously set objects in the parts of documents marked with semantic markers, to reveal the most important parts of documents (including by statistical criteria), to form networks of sentences and to remove the necessary volume of target information in the abstract.

2 The Suggested Approach

In addressing the problem, two approaches were proposed that could be considered new in this area. To solve the problem of determining the level of importance of individual parts of documents (in our case, sentences) it was suggested to move to the definition of weight values of separate hieroglyphs, not words in the text of documents and abstracts. It was also suggested that the document model should be considered as a network of sentences to identify the most important sentences for the parameters of the network. The weight of the links of the two sentences in this network is determined by the weight of the common hieroglyphs included in them.

Within the traditional statistical approach to the processing of natural languages, the weight of sentences is usually calculated on the basis of the estimated weights of lexical units (words, phrases) included in these sentences [2] - [5]. In this study it is proposed as such elements for the Chinese language to use separate hieroglyphs.

The transition from the words considered in the classical model to hieroglyphs allows avoiding the relatively complex procedure of words segmentation the text, which is inevitable with all other meaningful methods of Chinese texts automatic analysis. Of course, this approach is not applicable to European languages, where the number of different letters does not exceed several dozens. However, for the purpose of automatic text summarization of Chinese texts, the proposed approach provides acceptable results, which will be shown below.

It is known, that in the Chinese language there are more than 40 thousand hieroglyphs, therefore each of them (though not always, fully reflecting a semantic unit) it is possible to attribute a weight value calculated on the known formulas, for example, $TF \cdot IDF$ [6].

$TF \cdot IDF$ (TF — term frequency, IDF — inverse document frequency) is a statistical measure used to evaluate the importance of a word (in this case, not a word, but a hieroglyph) in the context of a document that is part of an array of documents. The weight of some hieroglyph is proportional to the number of its use in the document, and is inversely proportional to the frequency of occurrence of this character in all documents of array.

Thus, the measure $TF \cdot IDF$ depends on the word t (hieroglyph), the document d , the whole array of documents D , and is a product of two factors:

$$TF \cdot IDF(t, d, D) = tf(t, d) \times idf(t, D).$$

Here the expression $tf(t, d)$ is the ratio of the number of occurrences of some hieroglyph to the total number of characters in the document (to the length of the document, actually). Thus, the frequency of the hieroglyph within a single document is estimated.

The second factor, $idf(t, D)$ (inverse document frequency — the reverse frequency of the document) is the inversion of the frequency with which some hieroglyph occurs in the documents of array D .

IDF accounting allows you to reduce the weight of hieroglyphs that occur very often. There is only one IDF value for each t within the entire array of the documents D :

$$idf(t, D) = \log \frac{|D|}{|\{d \in D \mid t \in d\}|}.$$

In addition, unlike classical approaches to the definition of weight values of sentences, a new, network model is proposed. Under this model, a non-directional network is considered, with nodes appearing as separate sentences in the document, between which the links are established if they have common hieroglyphs. The weight of the relationship between the two sentences is defined as the sum of the weights common to these sentences. For this network, the weight of each sentence of the text is calculated as the sum of the weights of the links of all links that emanate from the node. Naturally, the weight of the proposals is then normalized, since long sentences without this procedure will on average have a deliberately greater weight. Practice has shown that a good normalization is the division by the logarithm of the length of the corresponding sentence.

3 Automatic Text Summarization of Legal Information

Procedures of automatic text summarization of an extractive class are based on determination of weight values (importance degree) of separate sentences which, in turn, depend on scales of words. In the study, as weight word meanings, the classical criterion $TF \cdot IDF$ was used though it is not only possible for the solution of a problem of summarizing approach [7]. Traditionally for definition of weight word meanings two known algorithms were used – in the first case the weight of the sentence was considered as the sum of weights, rated on length of this sentence, of the words entering it, and in the second case was used, so-called, a symmetric summarizing algorithm [8]. In this case the weight of the sentence was defined as the sum of weights of its links with the previous and subsequent sentence.

In addition, this paper proposes a network algorithm, which, unlike the second case, calculates the relationship not only between adjacent sentences, but also between all the sentences in the text. This approach, of course, is computationally more complex than the first two, but, as practice has shown, leads to better results. At the same time, the complexity of the algorithm, in the case of the considered approach of texts summarization in Chinese, is compensated by the fact that instead of words (segmentation of which in this case is not required) are considered only separate characters.

So let's present the basic steps of three considered algorithms of definition of weight values of sentences:

Step 1. For each hieroglyph t_i the value $DF = df(t_i, D)$ is calculated as the number of documents d_j from the documentary array D that contain this hieroglyph, that is

$$DF := |\{d_j \in D : t_i \in d_j\}|.$$

Step 2. For each hieroglyph t_i and document d the value as frequencies $TF = tf(t_i, d)$ of emergence of this term in the document is calculated:

$$TF := \frac{\#\{t_i \in d\}}{|d|}.$$

Then hieroglyph weight is calculated

$$w_i = TF \cdot IDF = TF \cdot \log \frac{|D|}{DF}.$$

Step 3. Segmentation of sentences, i.e. the text of the document is divided into separate sentences p_i and then the definition of their weight values wp_i . Let's introduce notation: let the sentence p_i of the set of sentences P ($p_i \in P$) consists of hieroglyphs $t_{i,k}$ with weight $w_{i,k}$. Let's write in a brief form the essence of three different algorithms.

Step 4a) Algorithm of the sum of hieroglyphs weights ($\sum tf \cdot idf$):

$$wp_i = \frac{1}{|p_i|} \sum_{k=1}^{|p_i|} w_{i,k}.$$

Step 4b) Symmetrical algorithm for calculating the power of connection of the sentence p_i with the nearest sentences (Nearest):

$$wp_i = \frac{1}{\log |p_i|} \sum_{k=1}^{|p_i|} (w_{i,k} w_{i-1,k} + w_{i,k} w_{i+1,k}),$$

where T is the general composition of the hieroglyphs of the array. If the character is not present in the document, its weight in it is equal to zero.

Step 4c) Network algorithm of calculation of force of link of the sentence (Network):

$$wp_i = \frac{1}{\log |p_i|} \sum_{\substack{j=1 \\ j \neq i}}^{|p_i|} \sum_{k=1}^{|p_j|} w_{ik} w_{jk}.$$

Step 5. The weight of the sentence is corrected depending on its location in the document. Weight values of initial and last sentences of the document artificially increase.

It should be noted that the specifics of legal information, requirements to the structure and volume of the abstract, allowed to use the above-mentioned universal approaches to the solution of a private special task.

The structure and volume of the abstract of the legal document (examples of such documents can be found on the website <http://www.gov.cn/> in the section /zhengce) are put forward requirements that have found their programmatic implementation:

1. Abstract start with the title of the document, given almost without changes.
2. The abstract notes the type of document (announcement "通告", report "报告", results of work "工作成果", provisions "政策" etc.).
3. If the document indicates its purpose ("目的", "奖补目的", "调整目的", "普查的目的和意义", etc.), it is also reflected in the abstract.
4. If the first or second sentence of the document identifies the subjects of appointment of documents (which is also visible by special markers), such a proposal is also included in the abstract.
5. If in the title of the document or in the designation of its purpose explicitly there are objects from the number of the previously known (included in the base objects table), these objects should be highlighted in the abstract.

6. If the document belongs to the type not subject to further processing (awards "表彰", announcements of bids"招标", letters "函"etc.), the abstract is considered prepared.
7. All sentences containing the objects selected from the title and purpose are selected from the text of the document. If such proposals are less than the required number (given in advance or calculated on the basis of the volume of the document), they are presented in the abstract in the same sequence as in the primary document. The abstract is considered prepared.
8. If the sentences are more than the required number, they are weighed according to the above algorithm (based on the results of testing the network algorithm is selected). After that the sentences ranked by weight and are presented in the abstract in the same sequence as in the primary document. The abstract is considered prepared.

According to the submitted requirements the program of automatic text summarization of the legal information provided in Chinese was developed. The web interface of the user of this program is given in the Figure 1.

4 Adjacent Tasks

Automatic text summarization of texts is one of important problems of technologies of the deep analysis of texts which includes some more directions, such as extraction of entities (Information extraction), creation of networks of words (Language Networks) reflecting features of subject domains, a clustering (Cluster Analysis).

The algorithm offered for summarizing leans on some set of in advance prepared words reflecting the main objects presented in legal documents (for example, "人口" – the population, "产业" – the industry, "儿童" – children, etc.).

At the same time, if you apply the algorithm of words segmentation, and then rank them, it is easy to identify the most common "extensions" of starting objects, for example, the concept of "organization" (组织) to expand to the concept of "international organization" (国际组织), "public organization" (社会组织), and the concept of "defense" (事业) to the concept of "people's air defense" (人民防空事业). As a result, the documents of the array of legal information have been put in line with the basic concepts that can act as "keywords", descriptors, basis for the construction of domain models (Subject domain).

As one of the types of domain models can be considered a words network, the nodes of which correspond to separate concepts. There were proposed and implemented such simple rules of building this network, i.e. rules of communication between nodes:

1. All objects from the base, pre-prepared list, included in one document are linked by links.
2. If two objects are in N different documents, the force of link between them equals N.

3. Concepts that are extensions of concepts from the starter kit are linked with the corresponding basic concepts.



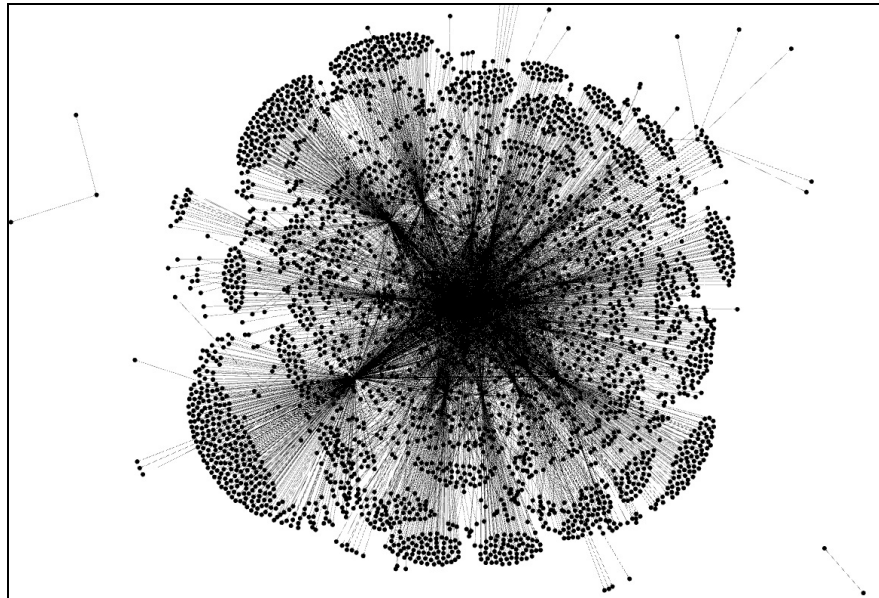
Fig. 1. The web interface of the system of automatic summarization.

With the help of the program Gephi (<http://gephi.org>) [9] the built network has been visualized (Figure 2) and received such parameters of the built network: number of nodes: 3364 (number of objects from the starting set – 220); - number of links: 10167; network density: 0.001; number of connected components: 6; average path length: 3,013; average clustering factor: 0.859.

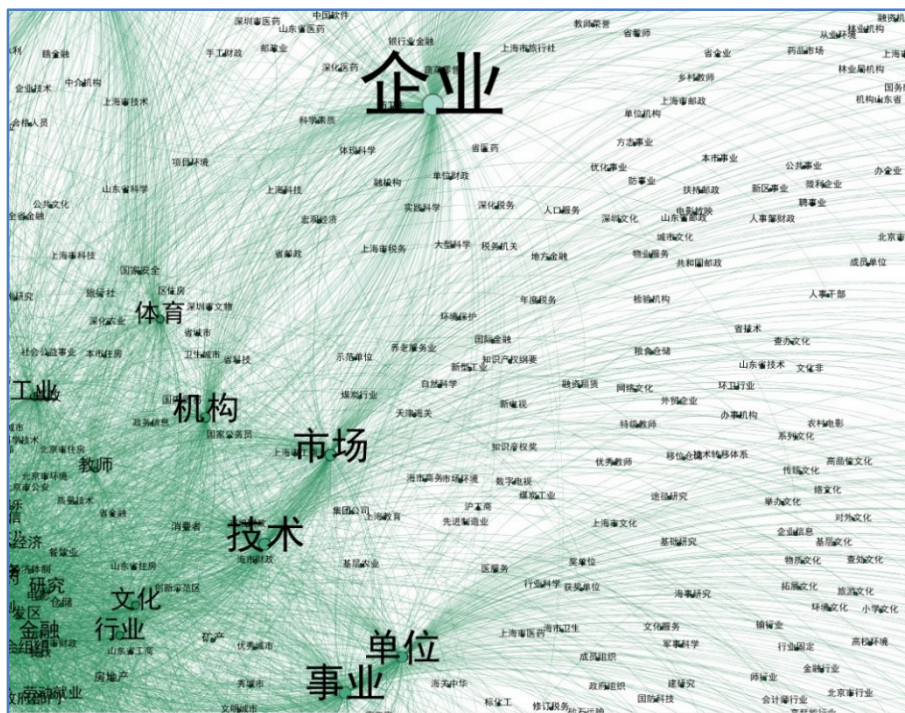
The topology features of the built network include a very large average clustering factor. This is due to the ode of a large number of concepts related only to the natal of their concept (the absence of other neighbors), and on the other hand the strong cohesion of objects from the start list. The small average length of the path indicates that the network is a Small World [10].

With the help of the program Gephi also received lists of the most important nodes in accordance with the criterion of PageRank and the greatest hubs by the criterion of HITS [11] (Figure 3).

The general view of network of words given on the Figure 2 clearly demonstrates a further possibility of a clustering of network, the choice of subsets – clusters from words (concepts). This procedure allows to allocate thematic subsets within the considered subject domain.



General view of words network



A fragment of the words network

Fig. 2. The web interface of the system of automatic summarization.

5 Methods of Results Evaluation

To evaluate the results, two assessments of the quality of the abstract are applied without experts – the cosine measure and the divergence of Jensen-Shannon (Jensen – Shannon), the justification of which is substantiated in the work [12].

Label	PageRank	Label	Hub
水利	0.001548	残疾人	0.04637
“十一五”	0.00154	租赁	0.044748
扶贫	0.00149	科技创新中心	0.043809
毕业生	0.001408	农民工	0.043738
银行	0.001405	统计	0.04244
上海市财政	0.001383	电信	0.04112
邮政	0.001374	省财政	0.040688
林业	0.001352	经济特区	0.039118
信息传输	0.001349	山东省财政	0.039081
城市规划	0.001337	作业	0.038172
文物	0.001323	食品药品	0.036646
医生	0.001315	北京市财政	0.036476
省财政	0.001242	水利	0.036002
技术服务	0.001215	试验区	0.035958
山东省财政	0.001194	电影	0.031812
农民工	0.001126	人工智能	0.031356
县域	0.00111	娱乐	0.03121
农田	0.001097	邮政	0.028793
电信	0.001072	物流业	0.027323
经济特区	0.001055	海关	0.026766
科技创新中心	0.001045	社会信用体系建设	0.026739
试验区	0.001	餐饮	0.02619
北京市财政	0.000989	深圳市市场	0.02569
食品药品	0.000964	干部	0.025645
电影	0.000949	公共管理	0.025336
房地产	0.000897	金融业	0.025252
矿产	0.00088	食品药品监管	0.025083
供销	0.000877	经济体制	0.025021

PageRank

HITS

Fig. 3. The web interface of the system of automatic summarization.

Let us explain the possibilities of using these approaches. The document's d hieroglyphic dictionary is supposed to consist of N elements $\{t_1, t_2, \dots, t_N\}$. Each hieroglyph corresponds to its weight, calculated according to the rule $TF \cdot IDF$. An array of these weights can be represented as a vector: $\vec{d} = (w_1, w_2, \dots, w_N)$. Accordingly, the hieroglyphic dictionary of the abstract r consists of a subset of the dictionary of the document and the abstract can also be put in line with the vector of weight values: $\vec{r} = (\hat{w}_1, \hat{w}_2, \dots, \hat{w}_N)$. In this case, we give a natural definition:

$$\hat{w}_i = \begin{cases} w_i, & \text{if } t_i \in r; \\ 0, & \text{if } t_i \notin r. \end{cases}$$

It is known that the scalar product of two nonzero vectors in Euclidean space A and B is defined by a formula:

$$\vec{A} \cdot \vec{B} = \|\vec{A}\| \|\vec{B}\| \cos \theta$$

Here θ – a corner between the considered vectors. It is natural if the direction of vectors coincides, the value θ becomes equal to zero (respectively, $\cos \theta = 1$). I.e. than closer $\cos \theta$ to unit, the direction of vectors is closer to those that is easily substantially interpreted for a case of the document and its abstract (the short summary). It is accepted function of proximity between vectors A and B to designate as $Sim(\vec{A}, \vec{B})$ (from word Similarity). In case of studying of a cosine measure of proximity we have:

$$Sim(\vec{A}, \vec{B}) = \cos \theta = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

where A_i and B_i – components of vectors $\vec{A}$ and $\vec{B}$, respectively.

According to definition of a cosine measure for calculation of proximity of the document and the paper it is possible to use a formula:

$$Sim(d, r) = \frac{\sum_{i=1}^N w_i \hat{w}_i}{\sqrt{\sum_{i=1}^N w_i^2} \sqrt{\sum_{i=1}^N \hat{w}_i^2}},$$

Proceeding from the fact that $\sum_{i=1}^N w_i \hat{w}_i = \sum_{i=1}^N \hat{w}_i \hat{w}_i$, we receive a formula which is used at practical calculations:

$$Sim(d, r) = \frac{\sum_{i=1}^N \hat{w}_i \hat{w}_i}{\sqrt{\sum_{i=1}^N w_i^2} \sqrt{\sum_{i=1}^N \hat{w}_i^2}} = \sqrt{\frac{\sum_{i=1}^N \hat{w}_i^2}{\sum_{i=1}^N w_i^2}}.$$

Another used criterion for formal identification of the degree of closeness – Jensen-Shannon divergence, which is based on the formalism of information theory and ma-

thematical statistics, in particular, on the relative entropy of Kullback-Leibler [13], [14].

The Kulbak-Leibler entropy is generally defined as a non-negative functional, which is an asymmetric measure of the distance between two probability distributions defined on a common space of elementary events.

The divergence distribution Q relatively P is designated $D(P\|Q)$. Distribution Q often serves as distribution P approach. This measure of distance in the theory of information is also interpreted as the size of losses of information when replacing true distribution P to distribution Q . The functional value can be understood as the number of unaccounted information of distribution Q if it was used for approach the distribution P .

For discrete probability distributions $P = \{p_1, p_2, \dots, p_n\}$ and $Q = \{q_1, q_2, \dots, q_n\}$ the Kulbak-Leibler entropy is defined as follows:

$$D(P\|Q) = \sum_{i=1}^n \log \frac{p_i}{q_i} p_i.$$

The entropy of Kulbak-Leibler, substantially close to the concept of distance, could be called a metric in the space of probability distributions, but this would be incorrect, since it is not symmetrical $D(P\|Q) \neq D(Q\|P)$ and does not satisfy the inequality of the triangle. In the future, we will use Jensen-Shannon divergence (JSD), which is based on Kulbak-Leibler entropy, but is a metric [15], [16], so it is also called "Jensen-Shannon distance" [17], [18], [19].

Jensen-Shannon's divergence is defined as follows:

$$JSD(P\|Q) = \frac{1}{2} (D(P\|M) + D(Q\|M)),$$

where $M = \frac{1}{2}(P + Q)$.

In case of application of distance of Jensen-Shannon to a problem of assessment of quality of abstracts the number of the lost information in the abstract in comparison with the initial document is estimated. As well as in a cosine measure, it is supposed that to the document d there corresponds the vector of the hieroglyphsweights $\vec{d} = (w_1, w_2, \dots, w_N)$, and to the abstract r – a vector of weight values: $\vec{r} = (\hat{w}_1, \hat{w}_2, \dots, \hat{w}_N)$. The "average" vector used in Jensen-Shannon's method is presented in the following form:

$$\vec{M} = \frac{1}{2}(\vec{d} + \vec{r}).$$

Respectively,

$$JSD = \frac{1}{2} (D(\vec{d}\|\vec{M}) + D(\vec{r}\|\vec{M})) = \frac{1}{2} \left(\sum_{i=1}^N \log \left(\frac{w_i}{\frac{1}{2}(w_i + \hat{w}_i)} \right) w_i + \sum_{i=1}^N \log \left(\frac{\hat{w}_i}{\frac{1}{2}(w_i + \hat{w}_i)} \right) \hat{w}_i \right).$$

Let's consider the given sums on two areas of index values: the 1st area where hieroglyphs of the document and the abstract coincide and 2nd, where do not coincide, i.e. where $\hat{w}_i = 0$:

$$JSD = JSD_1 + JSD_2 .$$

In the first area, obviously,

$$JSD_1 = \frac{1}{2} \sum_{i=1}^N \log \left(\frac{w_i}{\frac{1}{2}(w_i + w_i)} \right) w_i + \frac{1}{2} \sum_{i=1}^N \log \left(\frac{w_i}{\frac{1}{2}(w_i + w_i)} \right) w_i = 0.$$

In the second area, respectively,

$$JSD_1 = \frac{1}{2} \sum_{i=1}^N \log \left(\frac{w_i}{\frac{1}{2}w_i} \right) w_i + \frac{1}{2} \sum_{i=1}^N \log \left(\frac{\hat{w}_i}{\frac{1}{2}(\hat{w}_i)} \right) \cdot \hat{w}_i \rightarrow \frac{1}{2} \sum_{i=1}^N w_i.$$

Strictly speaking, the second term in the latter formula is not correct (you can consider the limit of expressions under the sign of the sum of when $\hat{w}_i \rightarrow 0$), but at the same time, we can make a fairly obvious conclusion that the Jensen-Shannon measure corresponds to the loss of information when summarizing and proportional to the total weight of the words (in our case – characters) included in the document, but missing in the abstract.

6 Comparison of Methods

When summarizing the new idea of determination of weight values of sentences on the basis of weights of separate hieroglyphs, but not words as it is standard was realized. Therefore, the quality of summarizing is checked not only proceeding from accounting of scales of separate hieroglyphs, but also taking into account scales of the whole words included in the documents and abstracts to be convinced that the offered approach is satisfactory also by criteria of traditional systems of summarizing. Naturally, this had to perform resource-intensive procedure of segmentation of words [20]. It should be noted that this procedure was performed only for quality check of algorithms of summarizing and is not a part of these algorithms.

The tests were conducted on a real array of legal information of the People's Republic of China in the amount of 10 thousand documents.

In Fig. 4-7 the results of the conducted tests are shown. In Fig. 4 and 6 are the results, when the models of documents and abstracts corresponded to vectors, elements of which - weights of individual hieroglyphs from the text of the document by $TF \cdot IDF$. In Fig. 5 and 7 – results, when elements of vectors correspond to weight values of words, segmented from texts of documents and abstracts. In Fig. 4 and 6 the results are given in accordance with the cosine measure of the proximity of the document and the abstract, and in Fig. 5 and 7, according to the Jensen-Shannon distance.

On the horizontal axis on all figures the number of sentences included in the abstract is marked. The vertical axis shows the values of the corresponding criteria, which are averaged throughout the document array. It should be noted that in all ex-

amples, as the first sentence of the abstract includes the title of the document, so the values with argument 1 for all four types of algorithms ($\sum tf \cdot idf$, Nearest, Network, Random) are the same.

As you can see, for comparison to the three above-mentioned algorithms, the Random method is added – compiling the abstract from the random sentences of the text (except for the first sentence – title).

The test results allow to summarize:

The proposed approaches lead to results, the quality of which is not lower, presented at the well-known conference on the analysis of texts TAC [12].

If the criterion of cosine measure of the proximity of the document and the abstract when taking into account the weight values of the individual hieroglyphs, the best results showed the method $\sum tf \cdot idf$ (which, of course, on the sum $TF \cdot IDF$ determined the weight of proposals, with the most significant included in the abstract), then by the same criteria, the proposed network method was the best way to take into account separate words of natural language.

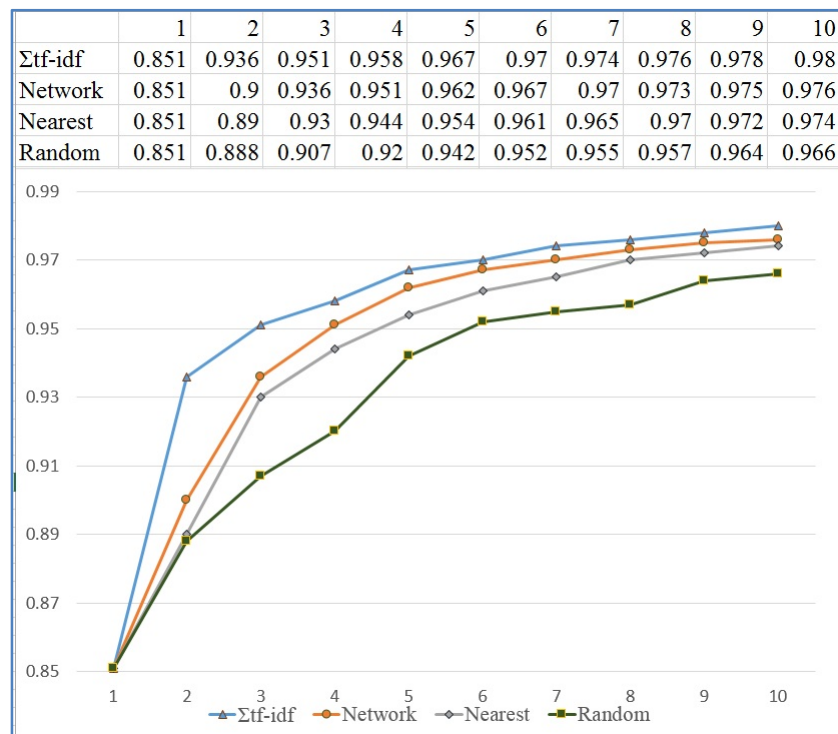


Fig. 4. A cosine measure of proximity of the text and abstract – accounting of the weight values of separate hieroglyphs.

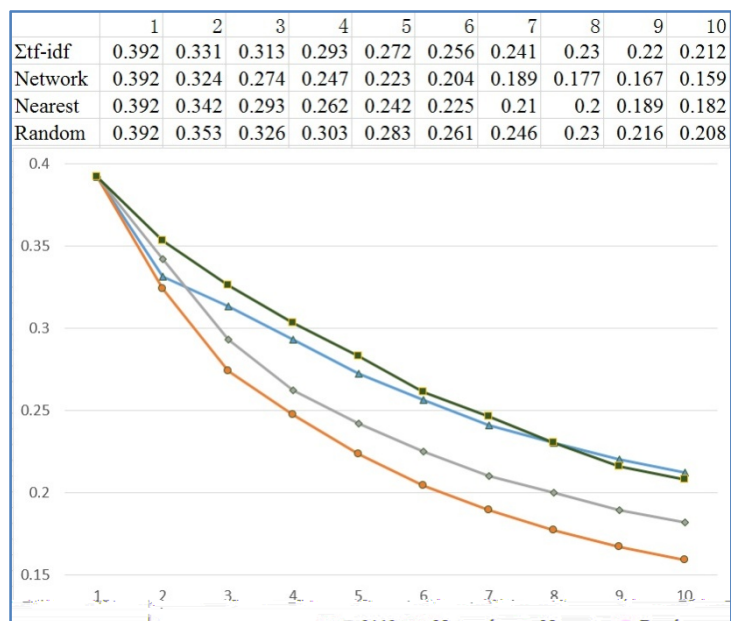


Fig. 5. Jensen-Shannon's divergence of loss of information when summarizing – accounting of weight of separate hieroglyphs.

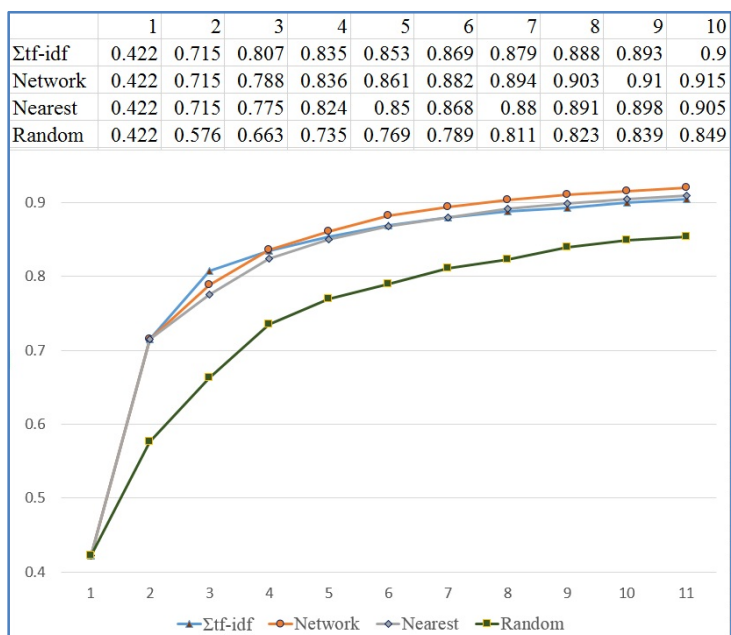


Fig. 6. A cosine measure of proximity of the text and abstract – accounting of the weight values of separate words.

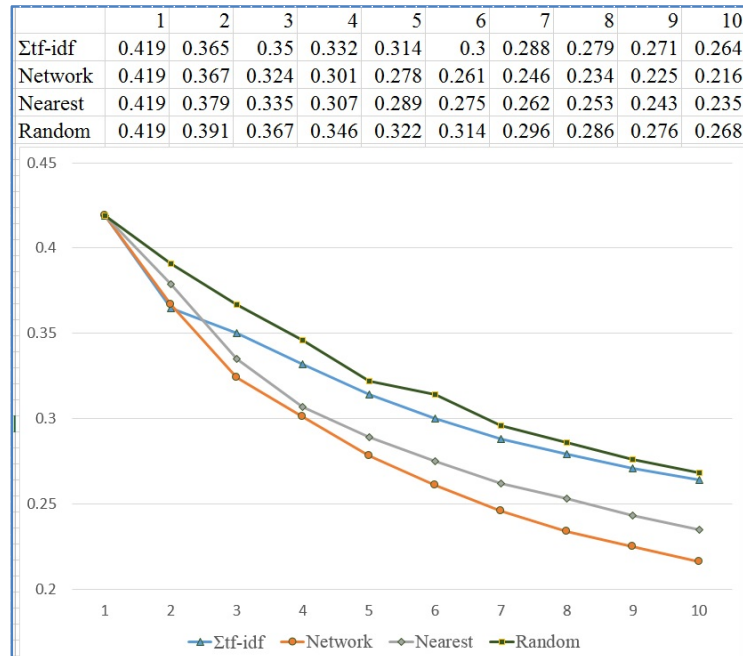


Fig. 7. Jensen-Shannon's divergence of loss of information when summarizing – accounting of weight of separate words.

7 Conclusions

We introduce a new hybrid method of automatic text summarization, covering statistical and marker methods, as well as taking into account the location of sentences in the text of the document. The offered model of the paper abstract reflects information need of customers during the work with legal information.

We brought the approach to determination of weights of separate hieroglyphs instead of segmented words in the text of documents. This technique avoids the expensive procedure of words segmentation required for other semantic methods of Chinese language processing.

Various methods of automatic text summarization are implemented and tested. Summarizing on the basis of the offered network model of the document was the best by criteria of a cosine measure and Jensen-Shannon's distances for papers which volume exceeds 2 sentences.

The offered approach, with minimal modifications, can be applied to texts on any subject of scientific, technical or news information.

References

1. *Luhn, Hans Peter* (1958). "The automatic creation of literature abstracts". *IBM Journal of research and development*, 2:159–165.
2. *Zhang, C.* (2008). "Automatic Keyword Extraction from Documents using Conditional Random Fields". *Journal of Computational Information Systems*, 4 (3): 1169–1180.
3. *Ramos, J.* (2003). "Using tf-idf to determine word relevance in document queries". *Proceedings of the first instructional conference on machine learning*, pp. 1-4.
4. *Bharti Santosh Kumar, Babu KorraSathya, Pradhan Anima* (2017). "Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles». *European Journal of Advances in Engineering and Technology*, 4 (6): 410-427.
5. *Chien, L.-F.* (1997). "Pat-tree-based keyword extraction for Chinese information retrieval". *ACM SIGIR Forum*. 31, ACM, pp. 50-58.
6. *Salton, G.; Buckley, C.* (1988). "Term-weighting approaches in automatic text retrieval". *Information Processing & Management*, 24(5): 513—523.
7. *Lande, D.V.; Snarskii, A. A.; Yagunova, E. V.; Pronoza E.* (2013). "The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text". *12th Mexican International Conference on Artificial Intelligence*. pp. 209-215. DOI: 10.1109/MICAI.2013.33
8. *Yatsko, V.A.* (2002). "Symmetric Summarization: Thematic Foundations and Methods". *Nauchno-Tekh. Inf., Ser. 2. – N. 5*: 18–28.
9. *Cherven, Ken* (2013). "Network Graph Analysis and Visualization with Gephi". *Packt Publishing*. ISBN: 9781783280131.
10. *Kleinberg, J.* (2000). "Navigation in a small world". *Nature*, 406 (6798): 845. DOI: 10.1038/35022643
11. *Langville, Amy N.; Meyer, Carl D.* (2011). "Google's PageRank and beyond: the science of search engine rankings". *Princeton university press*. ISBN: 978069115266
12. *Louis, Annie; Nenkova, Ani* (2008). "Automatic Summary Evaluation without Human Models". In *First Text Analysis Conference (TAC'08)*, Gaithersburg, MD, Etats-Unis, 17-19 November 2008.
13. *Kullback, S.; Leibler, R.A.* (1951). "On information and sufficiency". *Annals of Mathematical Statistics*, 22 (1): 79-86. DOI: 10.1214/aoms/1177729694.
14. *Kullback, S.* (1959), *Information Theory and Statistics*, *John Wiley & Sons*. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.
15. *Schütze, Hinrich; Manning, Christopher D.* (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass: MIT Press. p. 304. ISBN 0-262-13360-1
16. *Dagan, Ido; Lillian Lee; Fernando Pereira* (1997). "Similarity-Based Methods for Word Sense Disambiguation". *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*: 56–63. *ArXiv:cmp-lg/9708010*. DOI: 10.3115/979617.979625.
17. *Endres, D. M.; J. E. Schindelin* (2003). "A new metric for probability distributions". *IEEE Trans. Inf. Theory*. 49 (7): 1858–1860. DOI: 10.1109/TIT.2003.813506.
18. *Fuglede, Bent; Topsøe, Flemming* (2004). "Jensen-Shannon divergence and Hilbert space embedding". *Proceedings of International Symposium on Information Theory, ISIT 2004*, p. 31
19. *Lin, J.* (1991). "Divergence measures based on the Shannon entropy". *IEEE Transactions on Information Theory*, 37 (1): 145–151. DOI:10.1109/18.61115.

20. *Boris Berezin, Dmitry Lande, Oleh Pavlenko* Development, Evaluation and Usage of Word Segmentation Algorithm for National Internet Resources Monitoring Systems. CEUR Workshop Proceedings. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017).2067:16-22.