DE GRUYTER
OPEN

Research Paper

CitationAS: A Tool of Automatic Survey Generation Based on Citation Content*

Jie Wang[1,3], Chengzhi Zhang[2,3†], Mengying Zhang[1,3], Sanhong Deng[1,3]

[1]School of Information Management, Nanjing University, Nanjing 210023, China
[2]Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China
[3]Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing University, Nanjing 210023, China

Citation: Jie Wang, Chengzhi Zhang, Mengying Zhang, Sanhong Deng (2018). CitationAS: A Tool of Automatic Survey Generation Based on Citation Content.

Vol. 3 No. 2, 2018
pp 20–37
DOI: 10.2478/jdis-2018-0007

Received: Dec. 17, 2017
Revised: Feb. 2, 2018
Accepted: May 8, 2018

Abstract

Purpose: This study aims to build an automatic survey generation tool, named CitationAS, based on citation content as represented by the set of citing sentences in the original articles.

Design/methodology/approach: Firstly, we apply LDA to analyse topic distribution of citation content. Secondly, in CitationAS, we use bisecting K-means, Lingo and STC to cluster retrieved citation content. Then Word2Vec, WordNet and combination of them are applied to generate cluster labels. Next, we employ TF-IDF, MMR, as well as considering sentence location information, to extract important sentences, which are used to generate surveys. Finally, we adopt manual evaluation for the generated surveys.

Findings: In experiments, we choose 20 high-frequency phrases as search terms. Results show that Lingo-Word2Vec, STC-WordNet and bisecting K-means-Word2Vec have better clustering effects. In 5 points evaluation system, survey quality scores obtained by designing methods are close to 3, indicating surveys are within acceptable limits. When considering sentence location information, survey quality will be improved. Combination of Lingo, Word2Vec, TF-IDF or MMR can acquire higher survey quality.

Research limitations: The manual evaluation method may have a certain subjectivity. We use a simple linear function to combine Word2Vec and WordNet that may not bring out their strengths. The generated surveys may not contain some newly created knowledge of some articles which may concentrate on sentences with no citing.

Practical implications: CitationAS tool can automatically generate a comprehensive, detailed and accurate survey according to user's search terms. It can also help researchers learn about research status in a certain field.

JDIS
Journal of Data and Information Science

* This paper is an extended version of the article (http://ceur-ws.org/Vol-2004/paper2.pdf) in CLBib 2017 proceedings.

† Corresponding author: Chengzhi Zhang (Email: zhangcz@njust.edu.cn).

Originality/value: CitaitonAS tool is of practicability. It merges cluster labels from semantic level to improve clustering results. The tool also considers sentence location information when calculating sentence score by TF-IDF and MMR.

Keywords Automatic survey system; Citation content; Clustering algorithms; Label generation approaches; Sentence extraction methods

1 Introduction

Currently, quantity of electronic academic literatures has reached a massive level. Challenges have shown up when people want to investigate research status quo in a field (Liu, 2013): (1) When searching in academic databases (e.g., Web of Science[①]) or search engines (e.g., Google Scholar[②]), users are often given the relevant and ranked results which include many redundant information in themselves or among different platforms. (2) Although manual literature reviews can help researchers learn quickly about a new field, such surveys are in a small amount and their formation cycles are long which will lead to delay. Therefore, tools and systems are urgently needed to automatically generate a comprehensive, detailed and accurate survey according to the given topic words (Nenkova & McKeown, 2011). Here, survey means literature review. At the same time, such tools and systems should also help researchers retrieve relevant information in real time.

Ideally, automatic survey tools may deal with problems mentioned above. When applying such tools, how to choose data for survey generation is another challenge. Firstly, if all literature contents are used to generate a survey, system cost will be increased and unimportant and redundant contents may be added. Secondly, if we only use abstracts for survey generation, there will be information loss compared with using full text. Hence, citation content (represented by citing sentences) can be chosen as the source dataset and the main reasons include: (1) citation content is not only consistent with original abstract, but also can provide more concepts, such as entities and experimental methods (Divoli, Nakov & Hearst, 2012), and even retain some original information from cited articles. (2) since citation content reflects author's analysis and summarization of other articles, it has objectivity and diversity (Elkiss et al., 2008).

In this paper, we use citation content to conduct automatic documents surveys, and apply clustering algorithms to build an automatic survey generation tool, named CitationAS[③]. The main works include: (1) we build a demonstration website which can automatically generate surveys under a given topic; (2) we optimize a search

[①] http://isiknowledge.com/
[②] http://scholar.google.com/
[③] http://117.89.118.178:8001/CitationAS/

results clustering engine, Carrot2[®] (Osiński & Weiss, 2005a), in three aspects, similar cluster labels merging, important sentences extraction and survey generation; (3) we consider the location information of citation sentences when calculating sentence importance.

2 Related Works

Document summarization aims to solve the problem of information overload and help users better understand vast amount of text information. It can be divided into two categories (Marujo et al., 2015): (1) single document summarization. This technology applies a method to extract important content from one document and organizes them according to the order in original text. (2) multi-document summarization. It deals with document collection. This approach firstly needs to extract the relevant sentences from documents. Then, it sorts sentences according to redundancy and importance. Finally, the summary is generated based on the predefined length. Automatic surveys can use the technology of multi-document summarization.

Multi-document summarization technology includes the following aspects: (1) graph-based ranking method. This method usually works by performing over an affinity graph among sentences to extract important sentences for summarization. For example, Sarkar, Saraf and Ghosh (2015) used a hybrid similarity measure to improve the graph. Valizadeh and Brazdil (2015) designed a density-based graph model by adding density to LexRank and T-LexRank. (2) information extraction method. It first needs to design a summarization framework to extract information from documents. Then the summarization framework can be used to transform acquired information into semantically coherent summaries. Zhang, Li, Gao and Ouyang (2013) proposed a speech act-guided summarization approach, which could extract key terms with the recognized speech acts and generate template-based summaries. Jaidka, Khoo and Na (2013) developed a literature review framework by deconstructing human-written literature reviews and identified integrative and descriptive literature reviews. (3) sentence-feature-based method. This approach splits documents into sentences and classifies similar sentences into the same category. Then, it detects sentences from different categories to generate summary. Yang, Cai, Pan, Dai and Mu (2017) employed Nonnegative Matrix Tri-Factorization to cluster sentences using inter-type relationships and intra-type information.

Recent years, some researchers have applied citation content to generate surveys. For example, Qazvinian and Radev (2008) proposed to use citation networks to

[®] http://project.carrot2.org/

generate the summary. They constructed the weighted graph by using citation sentences as vertexes and similarity between sentences as edges. Then they applied clustering algorithms and calculated the weight of sentences in each cluster to select the summary sentences. Tandon and Jain (2012) generated the structured summary by classifying citation content into one or more classes. Cohan and Goharian (2015) grouped citation content and its context at first, and then ranked sentences within each group, finally sentences were selected for summary. Yang et al. (2016) utilized key phrases, spectral clustering and the ILP optimization framework to generate surveys. On the whole, the research is still at its beginning stage for automatic survey based on citation content, and it rarely achieves a demonstrable system. Therefore, this research reported here belongs to a relatively new research direction.

3 Dataset and Methodology

3.1 Dataset

In this paper, we collected about 110, 000 articles in XML format from PLOS One[5] between 2006 and 2015, covering subjects such as cell biology, mental health, computer science and so on. We identified citation sentences by rules whether a sentence containing reference marks (e.g., "[12]", "[1]-[3]"). And XML labels were removed. In our work, citation content refers to citation sentences. 4, 339, 217 citation sentences were extracted to be used as dataset for automatic survey generation. Table 1 displays citation sentences examples.

Table 1. Citation Sentences Examples.

No.	Citation sentence
1	Their transcription is dependent on mouse Cebpe and human CEBPE [12].
2	These changes may derive in a higher risk for type 2 diabetes development [8], [9].
3	It interacts with a variety of transcriptional factors and MLL proteins [9]–[12].
4	Most pathogens of humans, animals and plants are multi-host pathogens [1]–[3], [20].

Before the experiment of automatic survey generation, we analyse topic distribution of citation content dataset from an overall perspective. The aim is to understand the content of dataset and to provide a reference for acquiring search terms. Therefore, we apply LDA to analyse topics. LDA is a document topic generation model, which is also a three-level hierarchical Bayesian model (Blei, Ng & Jordan, 2003). In this model, firstly, it supposes that the word is generated by topic probability distribution. And each topic is a multinomial distribution on the vocabulary. Then, the document is assumed to be a mixture of potential topic

[5] http://journals.plos.org/plosone/

probability distribution. Finally, it generates topics from Dirichlet distribution for each document, and combines probability distribution of topics and words to generate each word in the document. There are some main parameters in LDA, which are α, β, K, d and z. Among them, α is the parameter of Dirichlet prior on per-document topic distribution. β is the parameter of Dirichlet prior on per-topic word distribution. K is the number of topics, d refers to the documents and z represents the topics. In this paper, we set α, β and K to 0.1, 0.01 and 10. We also run LDA for 2000 iterations.

3.2 Framework of CitationAS

Figure 1 shows the framework of CitationAS. Firstly, relevant citation sentences are retrieved from index files according to search terms from user interface. Then, we apply clustering algorithms to classify sentences into clusters which share the same or similar topic. After that, we merge clusters whose labels are more similar with each other. Finally, survey is generated based on important sentences extracted from each cluster. Results evaluation is carried out by volunteers.

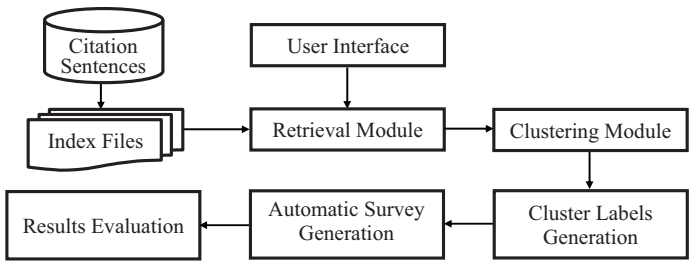

Figure 1. Framework of CitationAS.

3.3 Key Modules and Technology of CitationAS

3.3.1 Retrieval module

We use Lucene® to index and retrieve the dataset. When establishing index files, we add citation sentences and structure information including doi, cited count, positions of the sentence and its first word in the original article and paragraph. Since the final survey is based on user's search terms in CitationAS, we choose 20 high-frequency phrases from the dataset as search terms and used them for experiments. Phrases are shown in Table 2. Here, the frequency refers to the number of phrases presented in the citation content dataset. We divided them into ten high frequency 2-gram and 3-gram separately. When retrieved, our system gets relevant

® http://lucene.apache.org/

citation sentences based on inverted index and search terms. Then the results are ranked based on TF-IDF scoring mechanism. The higher score results in higher ranking.

Table 2. Top 20 Phrases According to High Frequency.

Phrase (Frequency)	Phrase (Frequency)
cell line (37507)	reactive oxygen species (5160)
gene expression (37001)	central nervous system (4418)
amino acid (35165)	smooth muscle cell (3439)
transcription factor (25626)	protein protein interaction (3286)
cancer cell (25605)	single nucleotide polymorphism (2535)
stem cell (22567)	tumor necrosis factor (2482)
growth factor (17531)	genome wide association (2386)
signaling pathway (16597)	case control study (2269)
cell proliferation (14203)	false discovery rate (2209)
meta analysis (12647)	innate immune response (2133)

3.3.2 Clustering module

In this module, we firstly apply VSM (Yang & Pedersen, 1997) to represent citation sentences and use TF-IDF (Salton & Yu, 1973) to calculate feature weights. In VSM, each citation sentence is equivalent to a document and expressed as $s_j = s_j(t_1,w_{1j};\ldots t_i,w_{ij}\ldots;t_m,w_{mj})$, where t_i is the i^{th} feature item, w_{ij} is feature weight of t_i in the j^{th} sentence, meanwhile, $1 \le i \le m$, $1 \le j \le N$, m and N are the number of feature items and citation sentences. The formula of TF-IDF is shown as (1).

$$w_{ij} = tf_{ij} * idf_i = tf_{ij} * \log\left(\frac{N}{n_i} + 0.01\right) \quad (1)$$

Where tf_{ij} is frequency of t_i in sentence s_j, and n_i represents the number of sentences where t_i is located.

Next, bisecting K-means, Lingo and STC, built-in Carrot2, are used to cluster the retrieved citation sentences respectively. The cluster results will be used for the next step of cluster labels generation. Since VSM represents documents in a high dimension, which can cost efficiency of clustering algorithms, we adopt NMF algorithm (Lee, 2000) to reduce dimensions. This algorithm obtains non-negative matrix after decomposing term-document matrix. It can be described as that for non-negative matrix A_{m*n}, we need to find non-negative matrix U_{m*r} and V_{r*n}, which should satisfy the following formula:

$$A_{m*n} \approx U_{m*r} \times V_{r*n} \quad (2)$$

Where U_{m*r} is the base matrix, V_{r*n} is the coefficient matrix, and r is the number of new feature items. When r is less than m, we can replace A_{m*n} with V_{r*n} to achieve dimensionality reduction.

In bisecting K-means, we will use coefficient matrix to calculate similarity between citation sentences and clustering centroids. Each sentence is assigned to the most similar cluster. Labels of each cluster are individual words which are three feature items with the greatest weight in the cluster. In this algorithm, three main parameters are set as follows: the number of new feature item r = 30, the number of clusters C = 25, the maximum number of K-means iterations *Iter* = 15. As long as reaching C or *Iter*, bisecting K-means will end.

Lingo algorithm (Osiński & Weiss, 2005b) firstly extracts key phrases by the suffix sorting array and the longest common prefix array. The number of words in each key phrase is between 2 and 8. Then it builds term-phrase matrix based on the key phrases, where feature weights are calculated by TF-IDF. Thirdly, it constructs base vectors according to the term-phrase matrix and the base matrix through NMF. Finally, each base vector gets corresponding words or phrases to form one cluster label, and sentence containing label's words will be assigned to the corresponding cluster. In Lingo, the number of clusters C is based on a very simple heuristic method, which is shown in formula (3). And in the base matrix, $r = C$.

$$C = \frac{dC}{10} * \sqrt{d} \tag{3}$$

Where dC denotes desired cluster count and d means the number of documents on input. Here, we set dC to 30.

STC algorithm (Stefanowski & Weiss, 2003) is based on Generalized Suffix Tree, not using NMF, which recognizes key words and phrases that occurred more than once in citation sentences. The maximum number of words in each key phrase is 4. Then each key words and phrases are used to come into being one base clusters. There may be many same citation sentences in two clusters, while the cluster labels are different. So, we merge these base clusters to form final clusters in order to decrease overlap rate of citation sentences between clusters. In STC, some parameters are set as follows: the maximum number of clusters C = 15, the maximum number of words and phrases in each cluster labels m = 3, the threshold of base clusters to merge t = 0.6.

3.3.3 Cluster labels generation

In clustering process, it is possible that some cluster labels are semantically similar to each other, for example, labels like "*data mining method*" and "*data mining approach*" for the search term "*data mining*". To improve experimental accuracy, similar cluster labels are merged when CitationAS uses clustering algorithms. We apply three methods to calculate semantic similarity between labels by using Word2Vec (Mikolov, Le & Sutskever, 2013) and WordNet (Fellbaum & Miller, 1998).

1) Similarity Computation Based on Word2Vec

Word2Vec is a statistical language model based on corpus. It applies neural network to get word vectors, which can be used to compute similarity between words. Given the phrase P, we assume that it is made up of word A, B and C. Then we can get the i^{th} dimensional representation in the phrase P, namely $\frac{1}{L}\sum_{i=1}^{Len}\left(a_i + b_i + c_i\right)$, where L means the number of words in P (Berry, Dumais & O'Brien, 1995). Finally, we use cosine measure to compute similarity between phrases. The formula is shown as (4).

$$sim\left(p_1, p_2\right) = \frac{\sum_{i=1}^{n} p_{1i} \times p_{2i}}{\sqrt{\sum_{i=1}^{n} p_{1i}^2} \times \sqrt{\sum_{i=1}^{n} p_{1i}^2}} \tag{4}$$

2) Similarity Computation Based on WordNet

WordNet is a semantic dictionary and organizes words in a classification tree, so semantic similarity between words can be calculated by path in the tree. The formula is shown as (5).

$$sim(w_1, w_2) = 1/distance(w_1, w_2) \tag{5}$$

Where $distance(w_1, w_2)$ denotes the shortest path between words in the tree (Rada, Mili, Bicknell & Blettner, 1989).

Then, similarity between phrases can use formula (6) to calculate.

$$sim\left(p_1, p_2\right) = \sum_{i=1}^{L_{p_1}} \sum_{j=1}^{L_{p_2}} \frac{sim(p_{1i}, p_{2j})}{L_{p_1} \times L_{p_2}} \tag{6}$$

Where p_1 and p_2 represent phrases, L_{p_1} and L_{p_2} mean the number of words in phrases, $sim(p_{1i}, p_{2j})$, calculated via formula (5), means the similarity between words in p_1 and p_2.

3) Similarity Computation Based on Combination of Word2Vec and WordNet

We linearly combine Word2Vec and WordNet to obtain a new similarity calculation method. The formula is shown as (7), where α is a weight and we set it to be 0.5.

$$sim(p_1, p_2) = \alpha sim_{word2vec}(p_1, p_2) + (1 - \alpha) sim_{wordnet}(p_1, p_2) \tag{7}$$

3.3.4 Automatic survey generation

In CitationAS, after several steps mentioned above, we can get many clusters. The last step is that clusters are sorted according to their sizes and each cluster is taken as a paragraph in the final survey. To choose important citation sentences

from each cluster, we design the following methods to calculate a sentence score. The higher score is, the higher ranking of sentence in the paragraph.

1) TF-IDF-based sentence extraction

Since each citation sentence is represented by the term-document matrix, we can obtain the sentence score (Zechner, 1996). For the sentence $s = s\ (t_1,w_1;\ldots t_i,w_i\ldots;\ t_m,w_m)$, its score is computed via the following formula (8):

$$w_{\mathrm{TF-IDF}} = \sum_{i=1}^{m} w_i \ / \ m \tag{8}$$

Thereby, we rank citation sentences in each cluster based on their scores. The sentences with higher scores will be used as survey sentences.

2) MMR-based sentence extraction

Maximal Marginal Relevance (MMR) (Carbonell, Jaime & Goldstein, 1998) method considers similarity of the selected sentence to search term and redundancy to sentences in survey. Calculation method of the sentence score is shown in formula (9).

$$w_{MMR} = \max_{s_i \in C-S} \left[\beta sim\left(s_i, q\right) - (1-\beta) \max_{s_j \in S} sim\left(s_i, s_j\right) \right] \tag{9}$$

Where C denotes the set of citation sentences in clusters, S denotes the set of survey sentences, so $s_i \in C - S$ denotes the set of those not selected as survey sentences. s_i means current citation sentence and q means search term. β is a parameter and generally set to be 0.7.

This method first selects the sentence with maximum MMR value as the survey sentence from the candidate sentence set, then it recalculates MMR values of the rest. It keeps selecting sentences until the set of candidate sentences is empty.

3) Sentence-location-based sentence extraction

Citation sentences have different importance and play diverse roles in different locations of citing papers. Maricic, Spaventi, Pavicic and Pifat-Mrzljak (1998) classifies the citation locations with respect to one of the sections of the citing paper: "*Introduction*", "*Method(ology)*", "*Results*", "*Discussion/Conclusions*", as IMRD. And they assigned the values of citation sentences in IMRD to 15, 30, 30, and 25, respectively. In our work, we adopt and modify their method. In our dataset, some citation sentences may not specify which section they belong to. So we put those into "*Others*" section. Finally, we set the weight of IMRDO as 0.15, 0.30, 0.30, 0.20, and 0.05, which means the corresponding location weight is given when citation sentence appears in a certain section.

In a word, when considering the citation sentence location, we have two other methods to calculate the sentence score. One is combination of sentence location and TF-IDF, which shown in formula (10), the other is combination of sentence location and MMR, as shown in formula (11).

$$w_{TF\text{–}IDF\_Location} = w_{Location} * w_{TF\text{–}IDF} \tag{10}$$

$$w_{MMR\_Location} = w_{Location} * w_{MMR} \tag{11}$$

4 Experiments and Results Analysis

4.1 Evaluation Method

In this paper, we invite 2 volunteers to make manual evaluation for the generated surveys. And we assign a score between 1–5 where 5 means the survey is very comprehensive, 1 means it is very poor and has no logical. The volunteers are fifth-year undergraduate students majoring in clinical medicine. The evaluation standards are described in Table 3.

Table 3. Evaluation Standards.

Score	Evaluation standards
5	Sentences are very smooth. Paragraphs and surveys are very comprehensive, exist very small redundancy and can fully reflect retrieval topics. The logical structure of survey is reasonable.
4	Sentences are relatively smooth. Paragraphs and surveys are relatively comprehensive, exist relatively small redundancy and can relatively reflect retrieval topics. The logical structure of survey is relatively reasonable.
3	Sentences are basically smooth. Paragraphs and surveys are basically comprehensive, exist certain redundancy and can basically reflect retrieval topics. The logical structure of survey is basically reasonable.
2	Sentences are not smooth enough. Paragraphs and surveys are not comprehensive, exist relatively high redundancy and cannot reflect retrieval topics enough. The logical structure of survey is confusing.
1	The smoothness of sentences becomes very poor. Paragraphs and surveys are far from comprehensive, exist very high redundancy and cannot fully reflect retrieval topics. There is no logical structure in the survey.

In the evaluation process, we give volunteers the surveys generated by CitationAS and the corresponding search terms for each survey, but we do not let them know the generated method behind each survey. Volunteers are demanded to evaluate each paragraph in the survey, thus we can get an average score of each survey. Since each survey is obtained by one of the methods, the average score of each method can be obtained as well.

4.2 User Interface of CitationAS

The user interface of CitationAS is shown in Figure 2. Users can input search terms and set parameters to get a survey. The parameters ("*Parameter setup*"

section) are about the number of citation sentences for clustering, clustering algorithms, cluster labels generation methods and important sentence extraction approaches. When users click "*search*" button, sub-topics, which are cluster labels and the number of sentences in each cluster, will appear in the left frame. Then users can click "*All Topics*", the survey will be presented on the right side, where the bold fonts are titles and others are content in survey's paragraph. Survey sentence's structure information will be displayed, when users put the mouse on it.

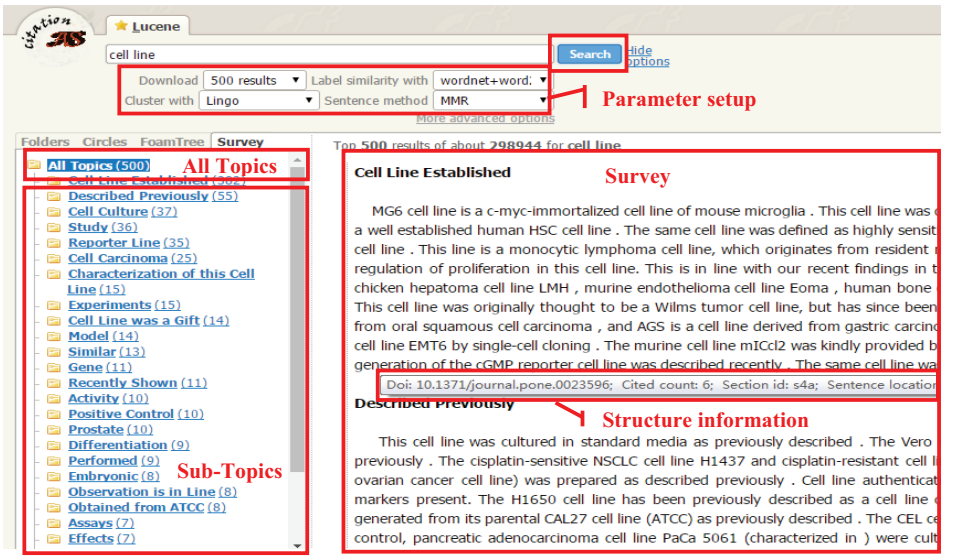

Figure 2. User Interface of CitationAS.

4.3 Results Analysis

4.3.1 Results of LDA

We applied LDA to extract topic distribution from citation content dataset, and the results were illustrated in Table 4. In each topic, topic words were sorted in descending order by their weights and we chose the top 10 topic words with the highest weight.

From Table 4, we could find that the dataset was mainly related to biomedical field, including protein and amino acid study, diseases study (for example, tolerance, chronic diseases and diabetes), bacteria and virus study (for example, HIV), gene sequence study, analytical method and models, and cell expression and activity. Through calculation, citation sentences on biology and life sciences accounted for 81.02%, indicating that the generated surveys would be about biomedicine. And the results obtained by LDA were consistent with the field reflected by search terms.

Table 4. Topic Distribution in Dataset.

Topic No.	Topic words
1	**protein**, domain, binding, structure, membrane, residue, acid, interaction, site, **amino**
2	disease, **patient**, increase, risk, study, disorder, **chronic**, factor, blood, clinical
3	**bacteria**, gene, strain, plant, resistance, species, report, found, host, pathogen
4	study, health, patient, year, **hiv**, treatment, country, population, report, clinical
5	**gene**, **sequence**, data, analysis, based, identified, study, expression, number, region
6	**model**, **method**, data, based, test, **analysis**, value, number, calculated, approach
7	**cell**, **expression**, tissue, mice, differentiation, development, human, stem, bone, mouse
8	acid, increase, level, activity, **glucose**, concentration, stress, enzyme, **insulin**, effect
9	study, process, task, response, visual, effect, memory, information, social, related
10	cell, signalling, pathway, **activation**, receptor, role, factor, protein, expression, **apoptosis**

4.3.2 Results of automatic survey generation

In this part, we show the results of label and survey generation. In the cluster labels generation test, we applied Davies-Bouldin (DB) and SC clustering index (Fahad et al., 2014) to find the best label generation method for each clustering algorithm. SC index is equal to the ratio of clusters' separation and compactness. If DB value is lower and SC value is higher, clusters are more compact and further from each other. The more number of search terms for consistency between DB and SC is, the better clustering results obtained by the method will be. Through experiments, we found combination of Lingo and Word2Vec had better clustering results with 8 search terms. When combining STC with WordNet, there were 6 search terms. If combining bisecting K-means with Word2Vec, we found a total of 9 search terms. However, combination of Word2Vec and WordNet did not have good performance compared with applying them separately on the three clustering algorithms. The quality of some cluster results based on this method was between WordNet and Word2Vec. Finally, we used Lingo with Word2Vec, STC with WordNet and bisecting K-means with Word2Vec to carry out the final automatic survey generation experiment.

In this paper, we chose 20 search terms and each of them generated surveys in 12 different approaches. Finally, 240 surveys were produced. Meanwhile, compression ratio was set to be 20%, meaning that the final survey length was equal to the number of retrieved citation sentences multiplied by 20%. In the experiments, we set the number of retrieved citation sentences to 500. Approach of *Evaluation Method* part was used to score the generated surveys. In order to describe concisely the selected survey generation approaches, we omitted Word2Vec and WordNet in Table 5, Table 6 and Figure 3. For example, method Lingo-Word2Vec-TF-IDF would be described as Lingo-TF-IDF. The following are the evaluation results of generated surveys.

Table 5. Six Methods Rankings based on Two Volunteers.

Ranking	Volunteer A	Volunteer B
1	STC-TF-IDF	Lingo-TF-IDF
2	Lingo-TF-IDF	Lingo-MMR
3	STC-MMR	STC-MMR
4	Lingo-MMR	STC-TF-IDF
5	bisecting K-means-MMR	bisecting K-means-MMR
6	bisecting K-means-TF-IDF	bisecting K-means-TF-IDF

Table 6. Six Methods Rankings based on Two Volunteers when Considering Sentence Location.

Ranking	Volunteer A	Volunteer B
1	Lingo-MMR	Lingo-MMR
2	Lingo-TF-IDF	Lingo-TF-IDF
3	STC-TF-IDF	STC-MMR
4	STC-MMR	STC-TF-IDF
5	bisecting K-means-TF-IDF	bisecting K-means-MMR
6	bisecting K-means-MMR	bisecting K-means-TF-IDF

Table 5 and Table 6 can be used to compare the similarities and differences between two volunteers' evaluation results. When not considering sentence location, we ranked six methods according to the average score of each method, which was shown in Table 5. We could find that rankings of STC-WordNet-MMR, bisecting K-means-Word2Vec-TF-IDF and bisecting K-means-Word2Vec-MMR were same in two volunteers' evaluations. They both identified the survey quality obtained by bisecting K-means algorithm was poor, especially the combination of bisecting K-means, Word2Vec and TF-IDF. Reasons of this phenomenon may be that bisecting K-means is hard clustering and each sentence must belong to one cluster. Some sentences in the same cluster may not have subjects aligned to the cluster's topic. And cluster labels may also not effectively reflect the topic of citation sentences in cluster. Volunteers gave different rankings for the rest of methods, which indicated each of these approaches had its own advantages and disadvantages.

When considering sentence location, rankings of six methods were shown in Table 6. As it showed, approaches of Lingo-Word2Vec-MMR and Lingo-Word2Vec-TF-IDF had the same high rankings in two volunteers' scores, which indicated that volunteers both agreed these two methods were better. In the third and fourth rankings were surveys generated by STC-WordNet methods. And the surveys using bisecting K-means and Word2Vec were in the fifth and sixth rankings. We could also find that, in the last four rankings, two volunteers had different ranking results on the TF-IDF and MMR methods, which were used to extract important sentences. But they had the same rankings on STC-WordNet and bisecting K-means-Word2Vec.

Comparing Table 5 and Table 6, we found that, when considering sentence location, rankings of six methods were varied, indicating that location information

affected quality of final surveys. Combination of bisecting K-means, Word2Vec, TF-IDF or MMR always had the lowest rankings, meaning that quality of final surveys generated by these methods were relatively poor. Seeing from two tables, sentence location had bigger influence on the surveys quality for volunteer A. However, it had a small effect for volunteer B.

To make a comprehensive analysis about 12 methods and the effect of sentence location information on survey quality, we averaged the scores of two volunteers. As illustrated in Figure 3, no matter whether sentence location was considered, scores obtained by these methods were close to 3, indicating the generated surveys were comprehensive and the surveys' logical structure was basically reasonable.

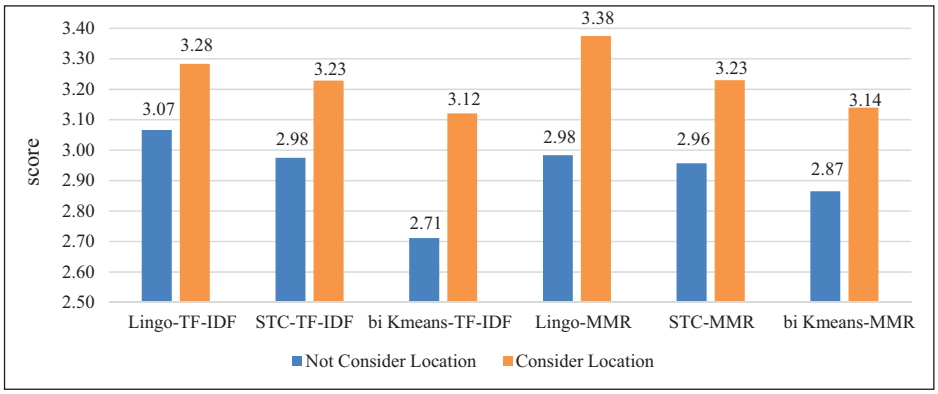

Figure 3. Average Scores of Different Methods.

First, we analysed the situation without considering sentence location. Combination of Lingo, Word2Vec and TF-IDF acquired the highest survey quality which was 3.07. When it came to TF-IDF or MMR, quality of survey generated based on combination of Lingo and Word2Vec was better. The reason may be that Lingo algorithm uses abstract matrix and the longest common prefix array when obtaining clustering labels, so that it can get more meaningful labels. In addition, the citation sentence is assigned to the cluster containing corresponding labels, instead of calculating similarity between sentence and cluster centroid. Compared to TF-IDF, we also found that survey quality was higher based on MMR after using combination of bisecting K-means and Word2Vec. Survey quality obtained by combination of STC, WordNet and TF-IDF or MMR was almost same, which indicated that sentences selection approaches did not have much impact on survey quality based on this clustering algorithm.

Next, when considering sentence location, we could find that the overall scores of survey quality had been improved. The survey quality obtained by Lingo-Word2Vec-MMR was the best, whose score was 3.38. Compared with the method

of no sentence location, survey scores had the same trend in TF-IDF or MMR method, which stated that sentence location information did not change the rankings of automatic survey generation methods in the comprehensiveness. The combination of bisecting K-means, Word2Vec, TF-IDF or MMR produced weaker results. And the survey quality obtained by combination of STC, WordNet, TF-IDF or MMR produced almost the same scores.

5 Discussion

According to the above results, we find that the survey quality can be improved after considering citation sentence location. In different sections, the importance of citation sentence is different, and by giving it different weights can change its score of being selected as a survey sentence, which will change the final survey quality. The sentence location weights set up here emphasize that sentences in *Method(ology)* and *Result* sections are the most significant, *Discussion/Conclusions* section comes second, and *Introduction* section is relatively unimportant. The experimental results reflect the rationality of this weight setting method.

In average scores of different methods, Lingo and STC belong to soft clustering, however, bisecting K-means is hard clustering. In soft clustering, a sentence may be assigned to a number of clusters. And the sentence may contain more than one topic. Therefore, soft clustering can better reflect the sentence's diverse topics. Then it can get a higher survey quality. Lingo algorithm uses non-negative matrix factorization to obtain latent semantic cluster labels, which makes the generated survey contains richer content. Consequently, compared with STC algorithm, the survey quality made by Lingo is better. However, hard clustering method assigns the sentence to only one cluster, which inevitably eliminates other topics within a sentence and affects the clustering results, reducing the survey quality.

For Lingo and STC algorithms, there is no definite conclusion whether or not using TF-IDF or MMR for the sentence importance. The reason may be that these two algorithms acquired meaningful cluster labels through their own approaches, and then assigned sentences to the clusters according to the labels. However, this procedure does not involve the similarity calculation between sentences. So, it is not clear which important sentence extraction method is better for these two clustering algorithms. Especially for STC algorithm, survey quality is almost the same obtained by TF-IDF and MMR whether or not considering sentence location. The possible reason is that STC identifies key words and phrases based on generalized suffix tree and frequency. After getting cluster results, citation sentences in each cluster contain some same words and phrases and the meaning of sentences may be very similar. Hence, when using TF-IDF and MMR to calculate sentence importance,

quality of generated survey may be same. For bisecting K-means algorithm and Word2Vec label generation method, results suggest positively using MMR method to calculate the sentence score. This is probably because bisecting K-means algorithm divides citation sentences according to similarity between cluster centroids and sentences. Meanwhile, MMR also considers similarity between citation sentences. However, TF-IDF ranks sentences only by their weights.

Finally, the experiments show that combination of Word2Vec and WordNet does not perform well. The reason may be that we only used linear function and set equal weights to combine them, which is too simple to bring out their strengths. In addition, we do not start from methods themselves, which means that we do not combine them based on their fundamentals.

6 Conclusion

CitationAS is built to automatically generate surveys. It mainly contains three components. The first is clustering algorithms including bisecting K-means, Lingo and STC. The second is cluster labels generation methods, Word2Vec, WordNet and the combination of them. The last one is automatic survey generation approaches based on TF-IDF and MMR. We also consider sentence locations when calculating sentence scores. Citation sentences are applied as the source data. We also use LDA to acquire topic distribution of dataset, which shows that dataset is mainly about biomedical field. Through experiments, we choose the best label generation approach for each clustering algorithm from semantic level, and then they are used in automatic survey generation. We also find that combination of Word2Vec and WordNet does not improve system performance compared with using them separately. Finally, automatic surveys obtained by 12 methods are positive, which means that sentences are basically smooth, survey content is basically comprehensive and reflects the retrieval topic, but some have redundancy. In addition, sentence location information can improve the generated survey quality. For soft clustering, such as Lingo and STC, survey quality may be better. We also note that the newly created knowledge of some article may concentrate on sentences with no citing. So those sentences may not be represented in the generated surveys which relying only on citing sentences.

In our future work, we will apply ontologies to calculate semantic similarity between labels and use deep learning to improve quality of generative survey. And citation frequency can be taken into account when calculating sentence importance. We will also select new approach to combine WordNet and Word2Vec in order to take their advantages. Besides, automatic evaluation can be made to avoid subjective or wrong judgements by human.

Acknowledgments

This work is supported by Major Projects of National Social Science Fund (No. 17ZDA291), Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201704) and Qing Lan Project.

Author Contributions

J. Wang (wangjie1342@qq.com) implemented the experiment, processed and analyzed data, designed system and drafted paper. C.Z. Zhang (zhangcz@njust.edu.cn, corresponding author) collected data, proposed the research idea and revised the paper. M.Y. Zhang (mf1714065@smail.nju.edu.cn) revised the paper. S.H. Deng (sanhong@nju.edu.cn) revised the paper.

References

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. Siam Review, 37(4), 573–595.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Carbonell, Jaime, & Goldstein. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 335–336.

Cohan, A., & Goharian, N. (2015). Scientific article summarization using citation-context and article's discourse structure. Proceedings of Conference on Empirical Methods in Natural Language Processing, 390–400.

Divoli, A., Nakov, P., & Hearst, M. A. (2012). Do peers see more in a paper than its authors? Advances in Bioinformatics, 2012(2012), 750214.

Elkiss, A., Shen, S., Fader, A., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? Journal of the American Society for Information Science and Technology, 59(1), 51–62.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: taxonomy and empirical analysis. Emerging Topics in Computing IEEE Transactions on, 2(3), 267–279.

Fellbaum, C., & Miller, G. (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Jaidka, K., Khoo, C., & Na, J. C. (2013). Deconstructing human literature reviews - A framework for multi-document summarization. The Workshop on European Natural Language Generation, 127, 125–135.

Lee, D. D. (2000). Algorithms for nonnegative matrix factorization. Advances in Neural Information Processing Systems, 13(6), 556–562.

Liu, X. (2013). Generating metadata for cyberlearning resources through information retrieval and meta-search. Journal of the American Society for Information Science and Technology, 64(4): 771–786.

Maricic, S., Spaventi, J., Pavicic, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. Journal of the Association for Information Science & Technology, 49(6), 530–540.

Marujo, L., Ribeiro, R., Matos, D. M. D., Joao P. Neto, Gershman, A., & Carbonell, J. (2015). Extending a single-document summarizer to multi-document: a hierarchical approach. Computer Science, 176–181.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. Computer Science, 1–10.

Nenkova, A., & McKeown, K. (2001). Automatic summarization. Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Student Research Workshop and Tutorial Abstracts, 5(3), 1–42.

Osiński, S., & Weiss, D. (2005a). Carrot^2: Design of a flexible and efficient web information retrieval framework. Proceedings of the Third International Atlantic Web Intelligence Conference, 439–444.

Osinski, S., & Weiss, D. (2005b). A concept-driven algorithm for clustering search results. IEEE Intelligent Systems, 20(3), 48–54.

Qazvinian, V., & Radev, D. R. (2008). Scientific paper summarization using citation summary networks. Proceedings of International Conference on Computational Linguistics, 689–696.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems Man & Cybernetics, 19(1), 17–30.

Salton, G., & Yu, C. T. (1973). On the construction of effective vocabularies for information retrieval. Acm Sigplan Notices, 9(3), 48–60.

Sarkar, K., Saraf, K., & Ghosh, A. (2015). Improving graph based multidocument text summarization using an enhanced sentence similarity measure. Proceedings of IEEE nternational Conference on Recent Trends in Information Systems, 359–365.

Stefanowski, J., & Weiss, D. (2003). Carrot^2 and language properties in web search results clustering. Proceedings of the First International Atlantic Web Intelligence Conference, 2663, 240–249.

Tandon, N., & Jain, A. (2012). Citation context sentiment analysis for structured summarization of research papers. Proceedings of 35th German Conference on Artificial Intelligence, 1–5.

Valizadeh, M., & Brazdil, P. (2015). Density-based graph model summarization: attaining better performance and efficiency. Intelligent Data Analysis, 19(3), 617–629.

Yang, L., Cai, X., Pan, S., Dai, H., & Mu, D. (2017). Multi-document summarization based on sentence cluster using non-negative matrix factorization. Journal of Intelligent & Fuzzy Systems, 33(1), 1–13.

Yang, S., Lu, W., Yang, D., Li, X., Wu, C., & Wei, B. (2016). KeyphraseDS: Automatic generation of survey by exploiting keyphrase information. Neurocomputing, 224, 58–70.

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th International Conference on Machine Learning, 4(3), 412–420.

Zhang, R., Li, W., Gao, D., & Ouyang, Y. (2013). Automatic twitter topic summarization with speech acts. IEEE Transactions on Audio Speech & Language Processing, 21(3), 649–658.

Zechner, K. (1996). Fast generation of abstracts from general domain text corpora by extracting relevant sentences. Proceedings of the 16th Conference on Computational linguistics, 2, 986–989.