Future Generation Computer Systems 32 (2014) 246–252

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Summarization of scientific documents by detecting common facts in citations

Jingqiang Chen, Hai Zhuge *

Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100190, Beijing, China
Nanjing University of Posts and Telecommunications, 210003, Nanjing, China
Aston University, Birmingham, B4 7ET, UK

HIGHLIGHTS

- A summarization system that expands citations through common fact.
- Explore common fact phenomenon in the scientific literature.
- An approach to expand terms to associated terms.

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form
18 July 2013
Accepted 31 July 2013
Available online 22 August 2013

Keywords:
Summarization
Semantic link network
Natural language processing

ABSTRACT

Reading scientific articles is more time-consuming than reading news because readers need to search and read many citations. This paper proposes a citation guided method for summarizing multiple scientific papers. A phenomenon we can observe is that citation sentences in one paragraph or section usually talk about a common fact, which is usually represented as a set of noun phrases co-occurring in citation texts and it is usually discussed from different aspects. We design a multi-document summarization system based on common fact detection. One challenge is that citations may not use the same terms to refer to a common fact. We thus use term association discovering algorithm to expand terms based on a large set of scientific article abstracts. Then, citations can be clustered based on common facts. The common fact is used as a salient term set to get relevant sentences from the corresponding cited articles to form a summary. Experiments show that our method outperforms three baseline methods by *ROUGE* metric.

1. Introduction

Researchers have to read many papers relevant to their research, especially for new comers of areas. Reading one paper, a reader needs to search and read a big number of cited papers.

A shortcut is to develop a system that automatically retrieves the most relevant content from the cited papers to complement the citations.

Previous studies either exploit single citation [1] or summarize the articles co-cited in one citation sentence [2]. We make a further progress by exploiting the information contained in multiple citations that co-occur in one paragraph or section.

Each citation may talk about some facts. For example in [3], "*It is generally agreed upon that manually written abstracts are good summaries of individual papers. More recently, Qazvinian and Radev (2008) argue that citation texts are useful in creating a summary of the important contributions of a research paper. The citation text of a target paper is the set of sentences in other technical papers that explicitly refer to it (Elkiss et al., 2008a). However, Teufel (2005) argues that using citation text directly is not suitable for document summarization.*" has three citation sentences, all of which mention the noun phrase of '*citation text*'. This can be regarded as the common fact of the three citation sentences. The three citations and their corresponding cited papers discuss the common fact from different aspects. We can use it to create a summary by retrieving relevant sentences from the cited papers to enrich the citation paragraph.

However, sometimes citation sentences may not use the same words explicitly to refer to a fact. Two citation sentences may not share terms with each other but they do talk about the same fact. In the following citation paragraph [4]: "*In (Elmacioglu and Lee, 2005), it was shown that the DBLP network resembles a small world network due to the presence of a high number of clusters with a small average distance between any two authors. This average distance is compared to (Milgram, 1967)'s six degrees of separation' experiments, resulting in the DBLP measure of average distance between two authors stabilizing at approximately six*", there are two citation

* Corresponding author at: Nanjing University of Posts and Telecommunications, 210003, Nanjing, China.
E-mail address: zhuge@ict.ac.cn (H. Zhuge).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.07.018

sentences, the first one talks about *small world network*, and the second talks about *six degrees of separation* phenomenon. They are totally different terms. However, the two citations are actually about the same fact. The noun phrase *small world network* is highly associated with the other noun phrase *six degrees of separation* in the domain of *complex network* as they co-occur frequently.

Therefore, it is necessary to expand the terms in a citation first to find the common fact. One way to expand a term is to find the ones that co-occur frequently with it according to corpus. After term expanding, the terms that co-occur in the citations are taken as the common fact. Here we can simply regard common fact as a set of terms with weights. The weight associated with a term reflects its significance in the common fact. Sometimes we need to cluster the citation sentences first for that despite they are in the same paragraph or section there may be different common facts existing in different subsets of the citations. Such common fact is then used as a query to get relevant sentences from the cited papers to form a summary.

Our main contribution is to exploit the common fact phenomenon to design a multi-document summarization system called *CFDSumm* to expand citations in an article. The first key technique is to expand terms in citations. We construct a term co-occurrence base based on 18 514 scientific abstracts in the domain of *Computational Linguistics*. The second key technique is to detect common facts in citations. We find out the common facts in the citations first and then cluster the citations based on the common facts. The third key technique is to find a subset of the most relevant sentences and form a summary. We treat common fact as a saliency term set where each member term is weighted and is used to score sentences. To evaluate our method, we create gold standard summaries by collecting 13 citation paragraphs from 11 papers manually. Experiments show that our method outperforms three baseline methods *MEAD* [5], *SciSumm* [2] and *CSIBS* [1] by *ROUGE*. The improvement significance is at p-value <0.05 and p-value <0.01.

2. Related work

Lots of work have been done on document summarization [6–14]. Analyzing citations in scientific articles is a feasible approach to summarize documents. Although there are many studies on treating citations as guidance for summarization, little work exploits common facts in citations.

The system *CSIBS* uses nouns in one citation as query to get more information from Ref. [1]. It is based on single citations rather than multiple citations. Another similar system *SciSumm* [2] aims to summarize documents co-cited within the same citation using surrounding text as query. *SciSumm* first applies *TextTiling* [15] technique to split the text into tiles which are then clustered. The clusters most relevant to the query are extracted to form a summary. Our work differs from *SciSumm* in that we work with multiple citations which may talk about the same fact and use such information to do summarization.

In recent years, citations are also used to create summary directly, called citation-based summarization [16,17,3,18–20]. Elkiss pointed out that citation summaries contain information that does not existing in abstracts and main contexts [16]. Hence, Qazvinian used citation sentences to create scientific paper summaries through citation summary networks and apply keyphrase extraction techniques to extract the key information contained in each citation which are then used to get high-quality citation-based summaries [19]. Meanwhile, Mohammad and Dorr think that citations can also be used to generate surveys of the scientific literature [3]. Our work is a reverse process of citation-based summarization.

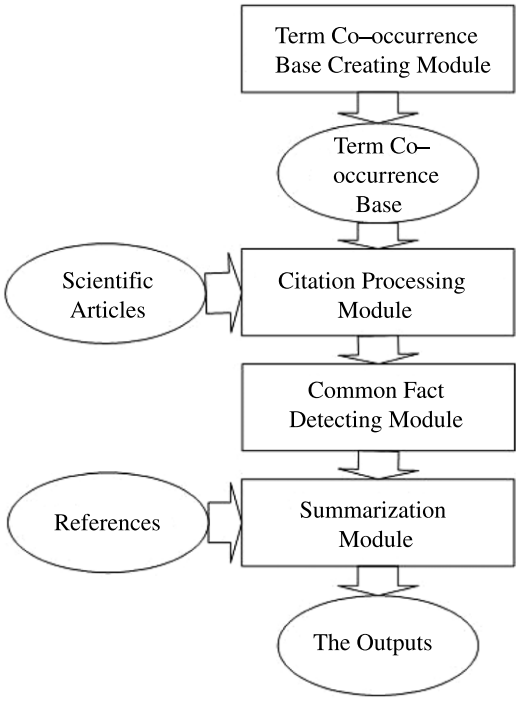


Fig. 1. The architecture of CFDSumm System.

Analyzing the function of citations of the scientific literature has also received a lot of attention in the past years. *Argumentative Zone Theory* is a rhetorical classification task that classify functions of citations into seven categories (Aim, Textual, Own, Background, Contrast, Basis and Other) [21,22]. Teufel pointed out that particularly important labels of Aim, Contrast and Basis are more suitable for creating extractive summaries [21]. In contrast, Nanba and Okumura classified citation functions into three main categories [23], i.e. type B, type C and type O. Here type B refers to references based on other researchers' theories or methods, type C refers to references that compare with related works or to point out their problems, and type O refers to references other than types B and C. They find that bibliographic coupling using citation type C is more accurate and efficient to classify scientific articles, which are then used to create review articles automatically.

A methodology for semantic linking through spaces for cyber-physical society was proposed [24]. A document is created by writers using language units according to semantic images in mental space. A probabilistic resource space model is proposed to manage resources in the cyber-physical society [25]. A text scanning mechanism simulating human reading process was proposed in [26].

3. The system

The architecture of the system is shown in Fig. 1, which is composed of the following four modules:

(1) The Term Co-occurrence Base Creating Module takes scientific abstracts, titles or even conclusions of a domain as input, preprocesses these resources and computes the frequently co-occurring terms to create the term co-occurrence base.
(2) The Citation Processing Module takes a scientific article as input, extracts citation sentences, parses the citation sentences

Table 1
Contingency table.

	A	$\overline{A}$
B	Sup(AB)	Sup($\overline{A}B$)
$\overline{B}$	Sup($A\overline{B}$)	Sup($\overline{A}\overline{B}$)

to get the noun phrases, from which we generate n-grams as terms and expands the terms based on the term co-occurrence base. An n-gram is a contiguous sequence of n items from a given sequence of text or speech. Here we only consider bigrams and trigrams.

(3) The Common Fact Detecting Module takes expanded terms set of citation sentences as input, computes the frequent term set as common facts, based on which citations are clustered.

(4) The Summarization Module takes a citation cluster and its corresponding references as input, retrieves these articles for relevant sentences according to the common fact of the cluster, eliminates the redundancy, reorders the sentences, and finally outputs the results.

3.1. Building the term co-occurrence base

Manually constructed lexical resources that specify sets of related terms and synonyms are expensive. Automatic approaches have been developed in text summarization: the use of thesaural resources for finding topically related terms was explored in [27], and resources such as *Wikipedia* have been mined as a similar resource of related terms in [28]. The former approach needs explicitly annotated sources which are usually uneasy to get and may need a great deal of human efforts. To the latter approach, topic-related terms are usually domain-specific, while *Wikipedia* resources do not have such property, so it is not suitable for our case.

We use the abstracts and titles of a collection of scientific articles in a specific domain as the sources, put the abstract and title of a scientific paper together and treat them as an item set. The information contained in the abstract and the title reflects the topic of a paper. As scientific papers are open to public, their titles and abstracts are available on the Web.

To get the training corpus, we collect 18 514 papers of *pdf* format in the domain of *Computational Linguistics* from the website of *ACL Anthology* (http://www.aclanthology.net). The documents are subsequently transformed into *txt* format. We process each article by the following steps:

(1) Extract the title and abstract of the article and put them together in a file which is treated as a transaction.

(2) Segment the text into sentences, parse the sentences and extract noun phrases of each sentence. At this step, *Stanford-CoreNLP* Tools are used.

(3) Generate bigrams and trigrams as terms from each noun phrases which are cleaned if containing stopwords; for unigrams we only use the nouns. At this step, the CMU statistical language model tool is used.

(4) Use *Porter Stemmer* to stem those n-grams which are then clustered together if having the same stems.

(5) Clean the un-frequent n-grams (i.e., n-grams that only occur in five or less transactions).

(6) Compute the highly associated terms for each term and build the Term Co-occurrence Base.

The reason of using n-grams of noun phrases as terms is that noun phrases contain the most important information of a sentence and n-grams conserve the common parts of noun phrases which may have unpredictable variations.

Table 2
Highly associated terms of *machine learning*.

	AB	$A\overline{B}$	$\overline{A}B$	$\overline{A}\overline{B}$	*Cosine*	χ^2
Learning technique	98	35	296	12 304	0.43	2489
Learning method	55	80	339	12 259	0.24	871
Supervised machine	34	19	360	12 320	0.24	997
Learning system	30	22	364	12 317	0.21	953
Classification	19	4	375	12 335	0.2	857

Table 3
Highly associated terms of *classification*.

	AB	$A\overline{B}$	$\overline{A}B$	$\overline{A}\overline{B}$	*Cosine*	χ^2
Sentiment	82	106	801	11 744	*0.27*	1397
Support vector	95	185	788	11 665	*0.25*	1122
Vector machine	93	182	790	11 668	*0.25*	1120
Multi way	19	4	864	11 846	*0.24*	1571
Verb lexical	16	2	867	11 848	*0.24*	1562

Table 4
Highly associated terms of *citation*.

	AB	$A\overline{B}$	$\overline{A}B$	$\overline{A}\overline{B}$	*Cosine*	χ^2
Experiment motivation	12	3	33	12 685	*0.46*	2734
Corpus machine	12	4	33	12 684	*0.45*	2563
Research paper	13	6	32	12 682	*0.44*	2528
Acl anthology	13	6	32	12 682	*0.44*	2528
Author topic	11	3	34	12 685	*0.44*	2465
Scientific article	11	3	34	12 685	*0.44*	2465

Then association discovery algorithms are applied. χ^2 and Cosine are two metrics to measure the degree of association between two terms [29]. They depend on a contingency table, as shown in Table 1. Sup(AB) denotes the number of term sets that term A and term B co-occur. Sup($\overline{A}B$) denotes the number of item sets that term A does not occur and term B occurs.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

$$\text{cosine}(A, B) = \frac{\sup(AB)}{\sqrt{\sup(A) \times \sup(B)}}. \tag{2}$$

In Eq. (1), n denotes the number of cells in the table, O_i denotes the observed frequency of each cell, and E_i denotes the expected frequency of each cell. Herein the *Cosine* measure is null-invariant while the χ^2 measure is not. A good strategy is to perform the cosine analysis first, and when the result shows that they are weakly positively/negatively correlated (the value is around 0.5), χ^2 analysis is performed to assist in obtaining a more complete picture. Tables 2–4 show some results of term association.

In Table 2, A denotes the occurrence of the term *machine learning*, and B denotes the occurrence of the term in the left column of the table. For example, the first number 98 in the left corner of Table 2 means there are 98 transactions containing term *machine learning* and term *learning technique* at the same time. The total number of transactions is 12 733 because the quality of the additional 5781 abstracts is low and they are eliminated. We can see that most of the *Cosine* values are weakly negatively correlated, while the corresponding χ^2 values are high and correct the property of the association. We can first order the related terms by *Cosine* and then by χ^2. Based on these two metrics, a weight function $W(t_i, t_j)$ is defined. $W(t_i, t_j)$ reflects the degree to which the term t_j is related to term t_i. $W(t_i, t_j)$ can be *Cosine* value, χ^2 value or even their combination. Here we let $W(t_i, t_i) = 1$, which means the relatedness of term t_i to itself is 1. For simplicity, we also denote $W(t_i, t_j)$ as w_{ij}.

In this paper, we use the *Cosine* value as w_{ij} for that most times *Cosine* value is positively correlated with χ^2 value. To cut off the lowly associated terms, a threshold is set. We have the following definition.

Definition 1. The expanded term set T_i of term t_i is calculated as:

$$T_i = \{t_j | W(t_i, t_j) > \text{ threshold } \wedge t_j \in T\} \tag{3}$$

where T is the universal set of all possible terms.

3.2. Processing citation

The goal of the citation processing module is to extract citation sentences from the running text of scientific articles. Prior works have studied the problem of citation parsing. Nanba and Okumura use cue phrases to extract citing areas in the paragraph [23]. They select 86 cue phrases and developed rules to extract citing areas automatically. Qazvinian and Radev applied *Belief Propagation*(BP) to discover implicit citation context of an explicit citation sentence [18]. In our case, we apply rule-based approach to extract citations through regular expressions.

When extracting citation sentences, the section number and paragraph number are also taken down simultaneously. Then noun phrases parsing and *n*-grams generating procedures are applied to citations as with abstracts processing. The resultant *n*-grams are then expanded according to term co-occurrence base. It is not necessary to create a term co-occurrence base in advance. We can expand each term of citation sentence in a lazy fashion. This can improve the efficiency of the system for that term co-occurrence procedure is rather time consuming when the number of terms increases to a certain degree.

We treat each citation sentence S_k as bag-of-words. Then, when its terms are expanded, S_k is also expanded to ES_k. We can see ES_k as a union of all its expanded terms whose weight is the summation of weights in the expanded term sets it occurs. The following is the formal definition of ES_k.

Definition 2. Let S_k be the original term set of citation sentence k; then the expanded term set ES_k is a two-tuple (EST_k, ESW_k), where EST_k is a term set and ESW_k is a weight function:

$$EST_k = \{t_i | \exists t_j \in S_k t_i \in T_j\} \tag{4}$$

$$ESW_k(t_i) = \begin{cases} \sum_{t_j \in S_k \wedge t_i \in T_j}^{W(t_i,t_j)} & t_i \in EST_k \\ 0 \quad \text{otherwise} \end{cases} \tag{5}$$

where W is the weight function defined in Section 3.1.

Another problem is how to get the reference articles of citations. Wan et al. deal with this issue by retrieving the cited scientific papers from the Internet directly [1], using *Embase* and *Science Direct* as their database. We automatically retrieve the reference papers through *Google Scholar*, which takes a title or author names as keywords and outputs a list of related hyperlinks. Usually the most relevant paper link is ranked first, so we simply get the *pdf* format of a reference paper through this link.

3.3. Detecting common fact

A common fact *CF* is a set of terms that co-occur in a set of citations. The weight w_i of term in *CF* can be defined as the minimum, maximum or average of the original weight of t_i in each citation S_i.

Definition 3. Let Common Fact be $CF, S_1, S_2, \ldots, S_n$ be n citations, and $ES_1, ES_2, \ldots, ES_n$ be corresponding expanded term sets; then the common fact *CF* of these citations is a two-tuple (*CFT*, *CFW*), where *CFT* is a term set and *CFW* is a weight function:

$$CFT = \left\{ t_i | t_i \in \bigcap_{k \geq 1 \wedge k \leq n} EST_k \right\} \tag{6}$$

$$CFW(t_i) = \begin{cases} \text{avg}_{k \leq 1 \wedge k \geq n} ESW_k(t_i) & t_i \in CFT \\ 0 \quad \text{otherwise.} \end{cases} \tag{7}$$

CF can be used as a set of salient terms to retrieve relevant sentences. Sometimes citations in the same paragraph do not talk about the same fact. So it is necessary to decide whether there are common facts in citations and then find out the common facts. To finish this task, the following two steps are taken.

Step 1: Cluster the given citation sentences.

Step 2: Find the common fact of each cluster.

We employ Frequent Term Based Clustering Algorithm [30], which treats each item set as a transaction. It utilizes frequent item set generating techniques (e.g. *Apriori*, *FPTree*) to generate frequent term sets. Usually the number of resultant term sets will be very large, so we only reserve those that contain most terms and are not subset of each other. Then, those frequent term sets are used as common facts. The weight of each term in the frequent term sets is defined as the average of its weights in original term sets.

We use *WEKA*, a machine learning tool set written in *Java* to generate the frequent term sets. The association rules creating module of *WEKA* utilize *Apriori* Algorithm, taking a set of transactions as input and output a set of frequent term sets which are then post-processed.

The final frequent term sets are treated as common facts, each of which corresponds to a citation cluster.

3.4. Summarization

The goal of the summarization module is to retrieve the most relevant sentences to complement the citation sentences based on the common facts. This module takes a citation cluster, its corresponding common fact and its reference article as input and outputs a summary. The following three steps are taken:

Step 1. Use the common fact as a set of salient terms to assign each sentence a score according to the count and weight of salient terms it contains, and choose *len* sentences with the highest scores, where *len* is the summary length.

Step 2. Use *MMR* [31] to eliminate the redundancy in the sentence set.

Step 3. Reorder the retrieved sentences to make it as coherent as possible.

Prior works have studied various methods for query-based summarization [27,28,9]. Lin presents a topic signature based approach to summarize multiple documents. This method first computes a set of terms that relate to a topic and then summarizes documents based on the computed term set. We treat the common fact in hand as a topic signature and compute the score of each sentence S_i as follows:

$$\text{score}_i = \sum_{t_j \in S_i \wedge t_j \in CFT} CFW(t_j). \tag{8}$$

Then, the redundancy exists in the candidate sentences should be taken into consideration. We use MMR to add a penalty factor to the computing of the score as shown in the following:

$$p(s_i) = \lambda \max_{s_j \in \text{summary}} \text{Sim}(s_i, s_j) \tag{9}$$

where λ is between 0 and 1 and can be tuned. *Sim* is the cosine similarity.

The last step is to reorder the sentences to make it as coherent as possible. We reorder the sentences based on the publishing date of the article and the order it appears in the original article.

teufel2005#00014 even though citeseer shows a text snippet around the physical location for searchers to peruse , there is no guarantee that the text snippet provides enough information for the searcher to infer the relation .

teufel2005#00019 in the citation map , the most important textual sentence about each citation can be displayed ; these sentences are extracted from the original text .

teufel2005#00023 citeseer makes the simplifying assumption that the most important information about a citation is always local to the physical citation , but this assumption does not hold .

teufel2005#00024 in the annotated corpus from teufel and moens (2002) , where sentences are marked up according to rhetorical context , we found that 69 % of the 600 evaluative contrast sentences and 21 % of the 246 basis sentences do not contain the corresponding citation ; the citation is found in preceding sentences instead .

elkiss2008#00026 recent work by nakov et al . (2004) has shown the utility of text near citations , which they neologized as "citances" and used to automatically learn paraphrases from biomedical papers.

qazvinian2008#00027 although there has been work on analyzing citation and collaboration networks (teufel et al . , 2006 ; newman , 2001) and scientific article summarization (teufel and moens , 2002) , to the knowledge of the author there is no previous work that study the text of the citation summaries to produce a summary .

qazvinian2008#00125 each fact in the citation summary of a paper is a summarization content unit (scu) (nenkova and passonneau , 2004) , and the fact distribution matrix , created by annotation , provides the information about the importance of each fact in the citation summary .

qazvinian2008#00044 The ACL Anthology is a collection of papers from the Computational Linguistics journal, and proceedings from ACL conferences and workshops and includes almost 11, 000 papers.

Fig. 2. The summaries for the first citation paragraph shown in Section 1. The sentences are from different articles.

4. Data

We use the *ACL Anthology Network* (*ANN*) which is collected and maintained by the *University of Michigan*. It includes 18 514 papers dating from 1960s to 2011 published in the most important venues in the domain of *Computational Linguistics*. *ANN* contains a metadata and a citation network. We get the titles of the papers from the metadata. The *txt* format of the papers provided by *ANN* is so rough-and-tumble (e.g. duplication of paragraphs, lack of sections, messy codes) that the abstracts extracted from them are low-quality and unusable. Therefore, we re-collect the *pdf* format of the papers from the ACL Anthology website and use *PDFBox* tools (downloaded at http://pdfbox.apache.org/) to transform them into *txt* format. Then, we extract the abstract from the papers, followed by an abstract cleaning procedure which cleans the low-quality abstracts out. Finally, we get 12 733 high-quality abstracts which are then used to create term co-occurrence base.

5. Evaluation

Both manual evaluation and automatic evaluation can be taken. Here we choose automatic evaluation metric *ROUGE*. We collect 11 papers in the domain of *Computational Linguistics* as shown in Fig. 2. We select 13 citation paragraphs. Each of the paragraphs contains at least two citations which may be about the same fact. Some paragraphs contain more citations and are about 2 or more different facts. Table 5 shows the number and titles of the papers where we select the paragraphs. We ask two specialists in this domain to read through the articles corresponding to each citation paragraph and manually create gold standard summaries of a fixed length (10 sentences ∼200 words). Both specialists read the citation paragraphs and the cited papers. They select the sentences they deem most relevant to the citation paragraphs and re-organize them to create summaries. We thus get two gold standard summaries for each of the 13 citation paragraphs.

Fig. 2 shows the automatically created summary of the citation paragraph exemplified in Section 1. This citation paragraph only has one citation cluster whose common fact is "citation text" and three papers are cited. We name the cited papers by their authors and publishing years. The sentences in the papers are numbered by their occurring order. For the limitation of space, we only show 8 most relevant sentences, each of which comes with its author, publishing year and sentence number. We can see that 8 sentences are from 3 different articles and are ordered by publishing year and sentence number.

Table 5
Papers where we select citation paragraphs and create gold standard summaries.

Paper no.	Title
P11-1051	Coherent citation-based summarization of scientific papers
N07-1040	Whose idea was this, and why does it matter? Attributing scientific work to citations
N07-1055	A unified local and global model for discourse coherence
N09-1066	Using citations to generate surveys of scientific paradigms
C08-1087	Scientific paper summarization using citation summary network
P10-1057	Identifying non-explicit citing sentences for citation-based summarization
C10-1101	Citation summarization through keyphrase extraction
W06-0804	How to find better index terms through citations
C10-2170	Sentence ordering with event-enriched semantics and two-layered clustering for multi-document news summarization
P11-3002	Sentence ordering driven by local and global coherence for summary generation
P10-1142	Supervised noun phrase coreference research: the first fifteen years

The widely used and freely available state-of-the-art *MEAD* [5] system is taken as our first baseline. We use the default configuration of *MEAD*. In the configuration, *MEAD* uses length, position and centroid for ranking each sentence. The second baseline is *SciSumm*. *SciSumm* is initially designed to summarize papers co-cited in the same citation using the surrounding text as query. To apply it to our case, we simply treat all the nouns in citation cluster as a query and then use it to get the appropriate content from the cited articles. The third baseline is the *CSIBS* system which is designed to expand each single citation sentence in a paper. For *CSIBS* method we first do summarization for each single citation and then put the resultant summary of each citation together to form a final summary.

We compute the *ROUGE* scores based on the $2 * 13$ gold standard summaries that are manually created by the two specialists as references. *ROUGE* has been traditionally used to evaluate the performance based on the *n*-gram overlap between the automatically created summaries and the target gold standard summaries. For our evaluation, we use two different versions of the *ROUGE* metric, *ROUGE-1* and *ROUGE-2*.

Table 6
The ROUGE scores.

Metric	CFDSumm	SciSumm	CSIBS	MEAD
ROUGE-1 F-measure	0.60992	0.56981	0.49616	0.50527
ROUGE-1 Recall	**0.66259**[b]	0.52688	0.48412	0.53616
ROUGE-1 Precision	0.57052	0.62816	0.51889	0.49818
ROUGE-2 F-measure	**0.43123**[a]	0.32615	0.26324	0.28990
ROUGE-2 Recall	**0.45137**[b]	0.31348	0.25708	0.28099
ROUGE-2 Precision	0.39125	0.37352	0.27495	0.29658

[a] Represent improvement significant at p-value <0.05.
[b] Represent improvement significant at p-value <0.01.

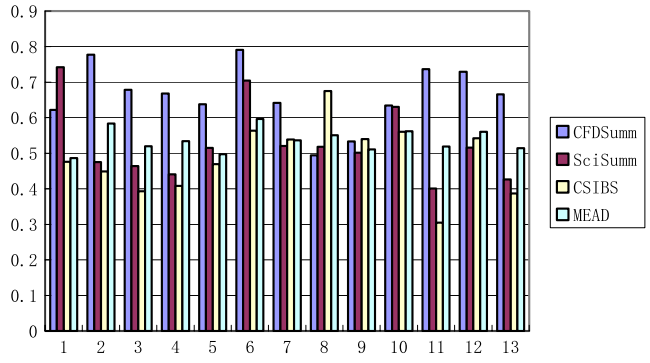


Fig. 3. ROUGE-1 Recall for the 13 citation paragraphs.

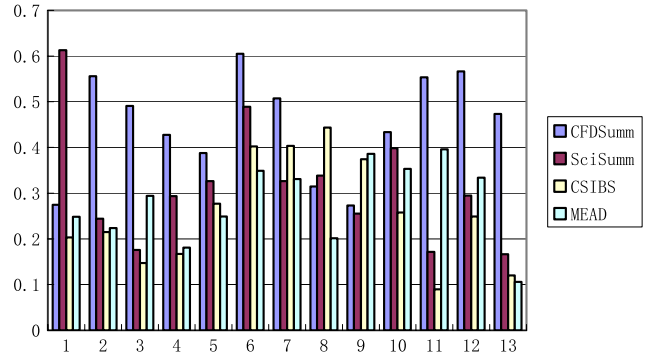


Fig. 4. ROUGE-2 Recall for the 13 citation paragraphs.

Table 6 shows the final comparative results. Figs. 3 and 4 show the *ROUGE-1 Recall* scores and *ROUGE-2 Recall* scores of each method for the 13 citation paragraphs, respectively. In Table 6, we use *Student T-Test* to measure the significance of improvement. The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. In our experiment, the null hypothesis is that the results generate by two systems are equivalent. The p-values generated by *Student T-Test* show our system performs significantly better than other three baseline systems by *ROUGE-1 Recall*, *ROUGE-2 F-measure* and *ROUGE-2 Recall*. The systems are evaluated by using $p < 0.05$ or $p < 0.01$ as criterion for statistical significance.

As shown in Table 6, our method performs better than *SciSumm* and *CSIBS*. The possible reason is that *SciSumm* and *CSIBS* do not focus on the common fact the multiple citations discussed. The two systems treat all the nouns in the citations equally and use them as query to retrieve sentences, so their resultant summaries are not as focused as that of *CFDSumm*, which first finds out the common fact in the citations and then uses it as a query to create summaries. This makes the generated summaries of *CFDSumm* more relevant to the common fact discussed in each citation. While *CFDSumm* may neglect some important information in the surrounding words of common facts. The term expanding step in our system completes our approach by considering the terms frequently co-occur with the terms in common facts.

6. Conclusion

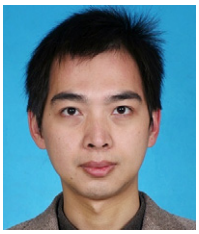
Humans have created an enormous document space by using languages according to semantic images in their mental spaces. Different documents are linked by various semantic links. Citations in scientific documents are such links given by writers. Citations in the same paragraph may be about a common fact which is depicted from different perspectives in different citations. A summary can be created on the common fact. Such summary is helpful for researchers who want a concise description of a group citation about a topic.

This paper introduced an approach to summarizing multiple scientific documents by detecting common facts. We expand the terms in citations first and then make use of frequent term based clustering algorithm to cluster the citations. Each citation cluster corresponds to a common fact. A saliency based summarization approach is applied to select sentences mostly related to the common fact. Experiments show that our method performs better than *MEAD*, *SciSumm*, and *CSIBS* methods. Our future work is to apply this approach to the citation sentences that are not in the same paragraph or even not in the same document. *Semantic Link Network* and *Resource Space Model* will be used to organize documents and summaries in the future [32,33].

References

[1] S. Wan, C. Paris, R. Dale, Whetting the appetite of scientists: producing summaries tailored to the citation context, in: JCDL2009, 2009, pp. 59–68.
[2] N. Agarwal, K. Gvr, R.S. Reddy, C.P. Ros, Scisumm: a multi-document summarization system for scientific artcles, in: ACL2011, 2011, pp. 115–120.
[3] S. Mohammad, B. Dorr, Using citations to generate surveys of scientific paradigms, in: NAACL2009, 2009, pp. 584–592.
[4] M.T. Joseph, D.R. Radev, Citation analysis, centrality, and the acl anthology, Technical Report CSE-TR-535-07, University of Michigan, 2007.
[5] D.R. Radev, T. Allison, S.B. Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, Z. Zhang, Mead—a platform for multi-document multilingual text summarization, in: LREC2004, Lisbon, Portugal, 2004.
[6] Y. Chali, S. Matwin, S. Szpakowicz, Statistics-based summarization łstep one: sentence compression, in: AAAI1999, 1999.
[7] G. Erkan, D.R. Radev, Lexrank: graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research 22 (2004) (2004) 457–479.
[8] K. Knight, D. Marcu, Summarization beyond sentence extraction: a probabilistic approach to sentence compression, Artificial Intelligence 139 (1) (2002) 91–107.
[9] D.R. Radev, H. Jingb, M. Stys, D. Tama, Centroid-based summarization of multiple documents, Information Processing & Management 40 (6) (2004) 919–938.
[10] D. Shen, J. Sun, H. Li, Q. Yang, Z. Chen, Document summarization using conditional random fields, in: IJCAI2007, 2007, pp. 2862–2867.
[11] T. Strzalkowski, J. Wang, B. Wise, A robust practical text summarization, in: AAAI1998, 1998.
[12] X. Wan, J. Yang, J. Xiao, Manifold-ranking based topic-focused multi-document summarization, in: IJCAI2007, 2007, pp. 2903–2908.
[13] W. Yih, J. Goodman, L. Vanderwende, H. Suzuki, Multi-document summarization by maximizing informative content-words, in: IJCAI2007, 2007, pp. 1776–1782.
[14] H. Daume, D. Marcu, Bayesian query-focused summarization, in: ACL2006, 2006, pp. 305–312.
[15] M.A. Hearst, TextTiling: segmenting text into multi-paragraph subtopic passages, in: Proceedings of LREC 2004, Lisbon, Portugal, May 2004.

[16] A. Elkiss, S. Shen, A. Fader, Blind men and elephants: what do citation summaries tell us about a research article? Journal of the American Society for Information Science and Technology Archive 59 (1) (1999) 51–62.
[17] A. Jbara, D.R. Radev, Coherent citation-based summarization of scientific papers, in: ACL2011, 2011, pp. 500–509.
[18] V. Qazvinian, D.R. Radev, Identifying non-explicit citing sentences for citation-based summarization, in: ACL2010, 2010, pp. 555–564.
[19] V. Qazvinian, D.R. Radev, Scientific paper summarization using citation summary networks, in: Proceedings of the 22nd International Conference on Computational Linguistics, Vol. 1, Association for Computational Linguistics, 2008, pp. 689–696.
[20] V. Qazvinian, D.R. Radev, A. Ozgur, Citation summarization through keyphrase extraction, in: COLING2010, 2010, pp. 895–903.
[21] S. Teufel, Summarizing scientific articles: experiments with relevance and rhetorical status, Computational Linguistics 28 (4) (2002) 409–445.
[22] S. Teufel, A. Siddharthan, D. Tidhar, Automatic classification of citation function, in: EMNLP2006, 2006, pp. 103–110.
[23] H. Nanba, M. Okumura, Towards multi-paper summarization using reference information, in: IJCAI1999, 1999, pp. 926–931.
[24] H. Zhuge, Semantic linking through spaces for cyber-physical-socio intelligence: a methodology, Artificial Intelligence 175 (2011) 988–1019.
[25] H. Zhuge, Y. Xing, Probabilistic resource space model for managing resources in cyber-physical society, IEEE Transactions on Services Computing 5 (3) (2012) 404–421.
[26] B. Xu, H. Zhuge, A text scanning mechanism simulating human reading process, in: IJCAI 2013.
[27] C. Lin, E. Hovy, The automated acquisition of topic signatures for text summarization, in: COLING2000, 2000, pp. 495–501.
[28] V. Nastase, Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation, in: EMNLP2008, 2008, pp. 763–772.
[29] J. Han, M. Kamber, Data mining: Concepts and Techniques, second ed., Elsevier Science & Technology, San Francisco, 2006.
[30] F. Beil, M. Ester, X. Xu, Frequent-term based text clustering, in: SIGKDD2002, 2002, pp. 436–442.
[31] J. Carbonell, J. Goldstein, The use of mmr, diversity-based re-ranking for reordering documents and producing summaries, in: SIGIR1998, 1998, pp. 335–336.
[32] H. Zhuge, The Knowledge Grid, first ed., World Scientific Publishing Co., 2004. 2012, second ed.
[33] H. Zhuge, The Web Resource Space Model, Springer, 2008.



Jingqiang Chen is a Ph.D. student at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His current research interests are text summarization, the theory and application of semantic link network.

Hai Zhuge is the pioneer of Cyber-Physical Society research (http://www.knowledgegrid.net/~h.zhuge/CPS.htm) and Knowledge Grid research. He is the author of The Knowledge Grid: Toward Cyber-Physical Society and the author of The Web Resource Space Model. Professor Zhuge invented a complex semantic space model by creating and integrating multi-dimensional classification space theory and self-organized semantic link network method. The model is based on normalized probabilistic multi-dimensional classifications, self-organized semantic interaction principles, rules for networking and reasoning. Professor Zhuge also created a set of models and methods for effectively sharing and managing knowledge in a self-organized scalable environment. His innovation includes the methodology of knowledge space, methods and techniques for analyzing and developing knowledge sharing and management environment, knowledge flow models for decentralized knowledge sharing, and a set of scalable high-performance peer-to-peer semantic networking mechanisms with innovative topologies for efficient knowledge sharing and management. Innovations significantly transform traditional centralized knowledge management methods and have influenced multiple areas. Professor Zhuge presented 15 keynotes at international conferences. He was ranked the top scholar in relevant area by journal assessment report. He received Wang Xuan Award of China Computer Federation for his fundamental theory of the Knowledge Grid. He was awarded a Distinguished Visiting Fellow of Royal Academy of Engineering. He is an ACM Distinguished Scientist and ACM Distinguished Speaker. He is serving as an associate editor of IEEE Intelligent Systems and steering the International Conference on Semantics, Knowledge and Grids (SKG, www.knowledgegrid.net). Email: zhuge@ict.ac.cn. Webpage: www.knowledgegrid.net/~h.zhuge.