The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0264-0473.htm

Exploring multiple diversification strategies for academic citation contexts recommendation

Received 28 February 2020
Revised 22 June 2020
6 August 2020
Accepted 8 August 2020

Haihua Chen
*Department of Information Science,
College of Information University of North Texas, Denton, Texas, USA*
Yunhan Yang
Dingxiangyuan Company, Hangzhou, China
Wei Lu
School of Information Management, Wuhan University, Wuhan, China, and
Jiangping Chen
Department of Information Science, University of North Texas, Denton, Texas, USA

Abstract

Purpose – Citation contexts have been found useful in many scenarios. However, existing context-based recommendations ignored the importance of diversity in reducing the redundant issues and thus cannot cover the broad range of user interests. To address this gap, the paper aims to propose a novelty task that can recommend a set of diverse citation contexts extracted from a list of citing articles. This will assist users in understanding how other scholars have cited an article and deciding which articles they should cite in their own writing.

Design/methodology/approach – This research combines three semantic distance algorithms and three diversification re-ranking algorithms for the diversifying recommendation based on the CiteSeer^X data set and then evaluates the generated citation context lists by applying a user case study on 30 articles.

Findings – Results show that a diversification strategy that combined “word2vec” and “Integer Linear Programming” leads to better reading experience for participants than other diversification strategies, such as CiteSeer^X using a list sorted by citation counts.

Practical implications – This diversifying recommendation task is valuable for developing better systems in information retrieval, automatic academic recommendations and summarization.

Originality/value – The originality of the research lies in the proposal of a novelty task that can recommend a diversification context list describing how other scholars cited an article, thereby making citing decisions easier. A novel mixed approach is explored to generate the most efficient diversifying strategy. Besides, rather than traditional information retrieval evaluation, a user evaluation framework is introduced to reflect user information needs more objectively.

Keywords Information retrieval, Diversity, Recommendations, Evaluation of retrieval results, Retrieval models and ranking, Text classification, Retrieval models, Retrieval ranking

Paper type Research paper

1. Introduction

The amount of scientific literature has greatly increased over the years. For example, in the computer science field, the number of publications indexed in the Web of Science database has grown from 396 in 1995 to 37,684 in 2017 (www.webofknowledge.com). Because of the massive growth of scientific literature, it has become more time-consuming and difficult for

The Electronic Library
Vol. 38 No. 4, 2020
pp. 821-842

0264-0473
DOI 10.1108/EL-02-2020-0046

users to read every related article, as well as decide which articles to cite and how to cite them in their own writing. Although different kinds of citation recommendation systems claim to be able to recommend articles for users to improve their writing efficiency, there are some well-known drawbacks, such as the existence of too many similar or redundant items in most traditional recommendation algorithms (Eskandanian *et al.*, 2017). Instead of a list of articles with meta-information, many users prefer recommendation systems that provide feedback on a set of knowledge fragments. A list of citation contexts can indicate how other users cited an article, usually defined as a sequence of words appearing around a citation placeholder. For example, "Horvitz identifies key landmarks in a user's history based on calendar activity [citation placeholder]" is a citation context in an article (Begg *et al.*, 2005). This citation context can be used to represent one piece of information within Horvitz's article. Beacuse Horvitz's article may be cited by different scientific papers or with different citation contexts in the same paper, the authors hypothesize that its main content can be represented by its citation contexts (Wan *et al.*, 2009). For researchers who want to cite an article, a list of citation contexts may be regarded as the comments about a cited article. This list then helps them to quickly comprehend the key points without spending much time reading its full text, thereby improving the effectiveness of their literature review process (Cohan and Goharian, 2017). Academic databases, such as CiteSeer^X and Semantic Scholar, display the citation contexts as a non-required field for a cited article by clicking "show context" (an example of CiteSeer^X is shown in Figure 1). Pre-investigation by interviewing researchers shows that ordinary users rarely pay attention to this feature and so may not make good use of it. Also, this citation context list ranked by citation count may be redundant; that is, the same information in the cited article could appear multiple times in

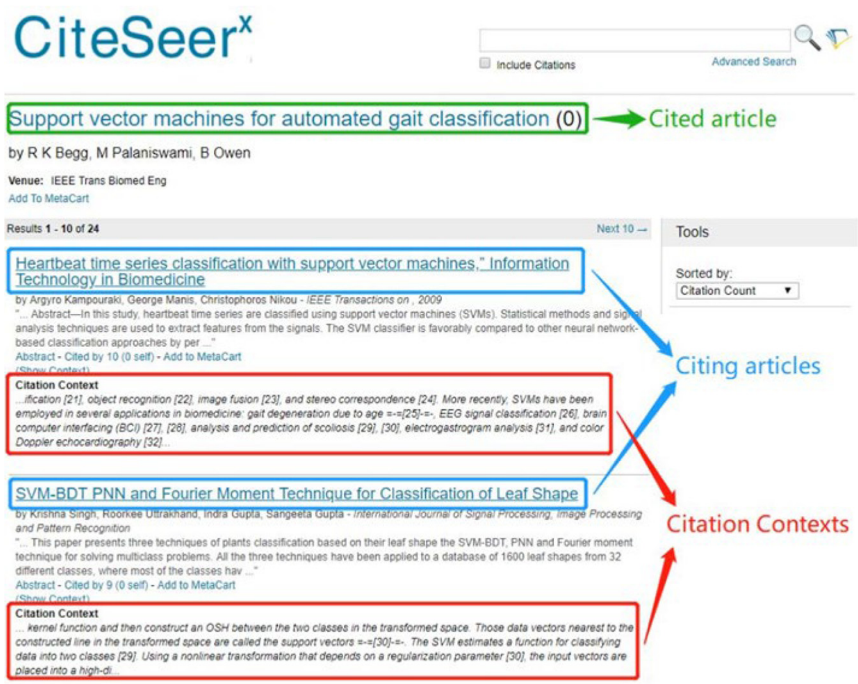

Figure 1.
Retrieval page of CiteSeer^X

the citation context list. Thus, it is necessary to re-rank the citation contexts to generate a diverse citation context list.

The three main contributions of this paper are as follows:

(1) The proposal of a novelty task that can recommend a diverse citation context list describing how other scholars cited an article, thereby making citing decisions easier.
(2) A novel mixed approach that combines different semantic distance calculation algorithms with different diversification re-ranking algorithms to explore the most efficient diversifying strategy for this task.
(3) The introduction of an evaluation approach with a user case study that reflects user information needs with a more objective evaluation compared to traditional information retrieval (IR) evaluation methods.

The remainder of this paper is structured as follows. First, Section 2 presents a brief overview of citation context-based applications and the current state-of-the-art re-ranking algorithms for diversification. Next, Section 3 gives a description of the research design and the re-ranking strategies used for diversification. Section 4 presents the experiment and results using the CiteSeer^X dataset, followed by deep analysis and discussion in Section 5. Finally, in Section 6, conclusions and discussion of future study areas are presented.

2. Related works

Related works can be divided into three main themes:

(1) citation context analysis;
(2) applications of citation contexts; and
(3) diversification for re-ranking in IR and recommendation.

In recent years, users have used citation contexts in academic literature for citation analysis (Doslu and Bingol, 2016), text summarization (Wu *et al.*, 2015) and IR and recommendation (Huang *et al.*, 2015; Tian and Zhuo, 2017), including many re-ranking algorithms for diversification.

2.1 Research on citation contexts analysis

Citation contexts analysis includes cited concepts analysis, citation sentiment detection, citation function identification and citation importance classification. Chang (2013) compared the difference of cited concepts and citation functions used between natural sciences and social sciences and humanities based on citation context analysis. Tandon and Jain (2012) proposed a method to analyse citation context sentiment for structured summarization of research papers. In other research (Teufel *et al.*, 2006), several functions of a citation and a group of manually crafted features (including cue phrases, verb tense and voice) were used for the automatic recognition of citation function. Färber *et al.* (2019) presented a system that allows researchers to search for papers and citation contexts. It displays indications about the so-called citation polarity; that is, whether the authors wrote about the cited publication in a positive, neutral or negative way. In addition, the citation context analysis presented an opportunity to use the wisdom of crowds for detecting important articles on a given topic (Doslu and Bingol, 2016). Recently, Semantic Scholar explored how to identify which of a paper's cited references were truly influential, rather than incidentally included for background or as a comparison; this breakthrough research also focused on citation context analysis (Bohannon, 2016; Cohan *et al.*, 2019). Existing research on citation contexts analysis has demonstrated the value of citation contexts. However, because the techniques of citation contexts identification still need to improve, there are not many high-quality public citation context data sets for further analysis.

2.2 Research on applications of citation contexts

The first important application is citation context-based summarization, which involves citation context extraction and selection. In detail, extracted citation contexts are classified into different clusters, a summary is generated by selecting citation contexts from each cluster (Ma *et al.*, 2020) and diversification should be considered to reduce duplications in the recommended list. For example, Cohan and Goharian (2017) extracted citation contexts from a reference article for each citation and extracted candidate sentences for the summary by using the discourse facets of the citations as well as the community structure of the citation contexts. The discourse model of the scientific article was also able to diversify the selection of citation contexts for the final summary, with a 34.6% mean ROUGE-L improvement for 100-word summaries and a 13.9% improvement for 250-word summaries, compared to the gold summaries. Qazvinian and Radev (2008) proposed using citation contexts to produce a summary of a single scientific article; they first clustered all the citation contexts based on different facts mentioned, then selected a sentence from each cluster. Later, they conducted citation context summarization based on key phrase extraction; however, this paper was not able to avoid the redundancy issue. In the citation context-based automatic survey generation system developed by Wang *et al.* (2018), recommending non-redundant citation contexts was the core step. Citation diversification can be a reasonable and implementable strategy to eliminate the redundancy issue in citation context-based summarization, as was also proven by Qazvinian and Radev (2008).

Citation contexts are also frequently applied to IR and recommendation systems in learning representations of scientific articles since they contain high concentrations of them (Liu *et al.*, 2014). Citation contexts can provide more semantic and accurate information for improving the IR performance: Doslu and Bingol (2016) showed that keywords extracted from citation contexts outperform a citing paper's own keywords in mining potential references, whereas Qazvinian *et al.* (2010) demonstrated that the window size of the citation contexts can significantly affect the IR performance. As for academic recommendation systems, citation contexts can either be used to model author research interests (Sugiyama and Kan, 2013) or be used to estimate the probability of citing a paper (Huang *et al.*, 2015), both of which were proven to enhance the accuracy of academic citation recommendations. Further, experiments also indicated that measuring the similarity of papers based on distributed representations learned from the citation context of papers could supplement the recommendation performance in the case of a lack of co-occurrence of items (Tian and Zhuo, 2017). In the citation recommendation process, diversification should be considered to ensure that a small result be diversified enough to contain more useful information (Küçüktunç *et al.*, 2013).

Although citation contexts have been proven useful in many situations, citation context applications still do not get the attention they deserve. Enlightened by the existing research, the authors noted that finding important but unique contributions of scientific articles based on their citation contexts will enhance user decision-making and the user's reading experience. Therefore, in this paper, the focus is on producing a high-quality and diverse citation context list which captures most of the valuable contributions of a citing article.

2.3 Research on diversification for re-ranking

The concept of diversity exists in many disciplines, such as sociology, business, biology and telecommunications. In this paper, its implementation in IR is discussed.

Drosou and Pitoura (2010) reviewed the major advances in the field of search result diversification (SRD) over the past few decades and organized existing approaches according to two complementary dimensions: aspect representation and diversification strategy. Aspect representation can be further divided into the two categories of implicit aspect representation and explicit aspect representation. Diversification strategy also includes coverage-based approaches, novelty-based approaches and hybrid approaches.

An implicit aspect representation relies on features belonging to a document to model different aspects. Maximal marginal relevance (MMR; Carbonell and Goldstein, 1998) is an earlier type of implicit diversified algorithm. There are many algorithms based on the idea of MMR that belong to the category of implicit diversified methods. Others include the quantum probability ranking principle (Zuccon and Azzopardi, 2010) and the diversified data fusion (Liang *et al.*, 2014) models.

By contrast, an explicit aspect representation seeks to directly approximate the possible information needs underlying a query, by relying on features derived from the query itself as candidate aspects, such as different query categories (Singh *et al.*, 2015) or query reformulations (Santos *et al.*, 2010a). A typical algorithm of this type is xQuADp (Santos *et al.*, 2010b). Liang *et al.* (2017) proposed a solution to the SRD task for short text streams. They developed a dynamic Dirichlet multinomial mixture topic model and a collapsed Gibbs sampling algorithm to infer latent topics. They then diversified search results based on a dynamic topic model.

In recent years, machine learning methods have been applied to SRD (Xia *et al.*, 2017). For instance, Wu *et al.* (2014) provided evidence that the learning to rank (LTR) method is an efficient way to retrieve biomedical information. This method is learned through a general ranking model (gLTR) and a diversity-biased model. The word2vec model has been proven to be effective in measuring the similarity between documents and therefore used in state-of-the-art SRD research (Shajalal *et al.*, 2018; Ullah *et al.*, 2016), which inspired the authors to compare the most popular algorithms in this paper.

Previous studies have shown that diversification in IR and recommendation usually positively influences user experience because it can cover a relatively full spectrum of user interests (Knijnenburg *et al.*, 2012), in turn reducing choice difficulty (Willemsen *et al.*, 2011). To achieve the goal stated before, the authors proposed a mixed approach in this study, which differs from existing methods in that it combines three semantic distance calculation algorithms with three diversification algorithms. Rather than automatic evaluation metrics, such as precision, recall or F-measure, normalized discounted cumulative gain is used in IR and recommendation, and a user study was conducted to evaluate the results.

3. Methodologies

The purpose of this study is to design a solution that allows users to read a diverse citation text with no duplicates when conducting research, especially at the literature review stage. To achieve this goal, several diversification strategies were investigated. In this section, the research design is described, as well as the three semantic distance algorithms and the three diversification re-ranking algorithms used.

3.1 Research design

This study consists of several steps, including data set construction, data pre-processing, semantic relevance calculation, the ranking of candidates by initial relevance score, re-ranking, diversification evaluation and result analysis. Figure 2 summarizes the research design.

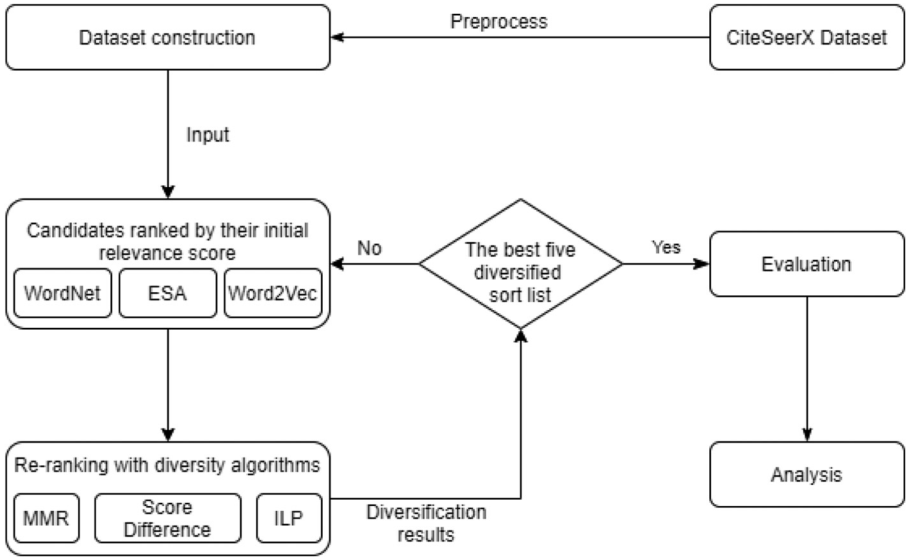

Figure 2.
Research design

The CiteSeer^X database was chosen as the source of text data for the experiments because it provides citation contexts. Specifically, articles were selected from the IR field that were cited between 11 and 100 times; that is, each had 11 to 100 citation contexts, and these citation contexts (Figure 1) were also included in the data set. In the CiteSeer^X database, the citation contexts were ranked by the citation count of the citing articles, which served as a baseline method in the study. Based on the CiteSeer^X citation context list, explicit semantic analysis (ESA), WordNet and word2vec were used to calculate the semantic distance between each pair of citation contexts and the semantic distance between the abstract of the cited article and each citation context. After that, three re-ranking algorithms [MMR, score difference algorithm (ScoreDiff), and integer linear programming (ILP)] were applied to generate a diverse citation context list based on the semantic distances. The final diverse citation context lists presented to users contained 10 items. As a part of the methods, the quality of the final diverse citation context lists was measured by the user's reading experience, obtained through a survey. Using the survey results, the authors were able to judge whether the citation contexts were better than the original list and which diversification strategy had performed the best.

3.2 Semantic distance calculation

In this study, a set of citation contexts of the cited article was defined as $C = C_1, C_2, \ldots C_n$. The semantic distance between them is calculated by their similarity sim(C_i, C_j) using three algorithms: WordNet, ESA and word2vec.

3.2.1 Wordnet similarity. WordNet is the most popular English lexical dictionary database. It has been widely used for studying semantics-related processes (Jain and Gaur, 2017). It contains more than 140,000 nouns, verbs, adjectives and adverbs. The synonym sets and various semantic relations among concepts are well-established, including synonymy, antonymy, hypernymy, hyponymy and meronymy. WordNet can be used to calculate the semantic similarity between different words. In this study, the context of two citations was first split into words and then the average semantic similarity between two words was used to express the semantic similarity between citation contexts. For example, if you want to calculate the semantic similarity between the two sentences "Hello world" and "Hello everyone", labeled C_1 and C_2, respectively, and Len (C_1) and Len (C_2) represent the number of words in C_1 and C_2. First, they are split into C_{11} (Hello), C_{12} (world), C_{21} (Hello) and C_{22} (everyone). Then, the semantic similarity of the two sentences is sim(C_1, C_2):

$$sim(C_1, C_2) = \frac{sim(C_{11}, C_{21}) + sim(C_{11}, C_{22}) + sim(C_{12}, C_{21}) + sim(C_{12}, C_{22})}{len(C_1) \times len(C_2)} \tag{1}$$

Among them, sim(x, y) is the semantic similarity between words calculated by WordNet. Specifically, the natural language toolkit (NLTK), a tool for natural language processing in Python, was used to provide a function called "path similarity" that outputs the semantic scores of two words (Loper and Bird, 2002).

3.2.2 Explicit semantic analysis. ESA is a knowledge repository and rule-based approach for the semantic relationship calculation of text. The method explicitly represents the text as a weighted vector of Wikipedia-based concepts with machine learning techniques (Gabrilovich and Markovitch, 2007). Because of the lack of semantic relations in semantic dictionaries, the effectiveness of the practical application is not significant, and the rule-based algorithm based on the knowledge base can make up for this deficiency. Compared with other semantic distance algorithms, using ESA resulted in substantial improvements in the correlation of computed relatedness scores with human judgments. Importantly, because of the use of natural concepts, the ESA model is easy to explain to a human.

ESA interprets a term vector **x** (with n dimensions) as a Wikipedia-based concept vector **v** (with m dimensions) by multiplying the index matrix $\mathbf{I}^{\top}$ by the term vector **x**. This multiplication represents a term vector of a text to a higher vector space that is considered to be a concept space. Each weight v_j of concept dimension v_j in vector **v** is defined as follows (Rahutomo and Aritsugi, 2014):

$$V_j = \sum_{k=1}^{n} x_k \cdot i_{jk} \tag{2}$$

where x_k is the dimension weight of term i in vector **x** and i_{jk} is the weight of concept j for term i. The weight of a term or a concept can be determined by its term frequency, collection frequency or normalization component.

If the ESA measures the semantic relatedness of two texts, then both texts are represented into two concept vectors **u** and **v**. The measurement of the vectors can then be accomplished in the vector space of the concept by a vector measurement such as cosine similarity (Rahutomo and Aritsugi, 2014):

$$\mathrm{sim}(u, v) = \frac{u \cdot v}{\|u\|\|v\|} \tag{3}$$

There are many implementations of ESA distance calculation. In this study, the ESAlib tool (Knoth *et al.*, 2011) was used which proved to be one of the most efficient methods to calculate semantic similarity.

3.2.3 Similarity based on word2vec. Word embedding constructed features for each word of a document and transformed the word into a low-dimensional vector representation, which is a distributed representation at the word level. Kong *et al.* (2018) took advantage of word embedding to represent text and citation networks to design an unsupervised feature. On this basis, they developed a scientific paper recommendation system with optimized performance. In 2013, two academic papers mentioned the word embedding toolkit, word2vec (Mikolov *et al.*, 2013a, 2013b). The citation count for their papers has reached more than 9,000, which reflects the huge influence of their studies. This tool could be effectively trained on a million-level dictionary or a million-level data set and also could conveniently calculate the similarity between different words. This study also chose this tool for semantic distance expression. The required word vector model is selected from the shared resources on GitHub (a project hosting platform). The model's training data is derived from Wikipedia and contains 300-word features and 174,015 words. The context window for word feature extraction was not fixed and depended on the syntax of the context. The result of the textual expression was the average score of the superposition of the vector expressions of all the words. Words that were not matched by the word vector model have been ignored.

In this paper, the sentence vector was constructed by averaging all word-embedding vectors, as shown in equation (4). Recently, the average of word-embedding vectors has proven to be a strong baseline in a multitude of tasks (Kenter and De Rijke, 2015):

$$S = \frac{1}{|W|}\sum_{i=0}^{|W|} v(W[i]) \tag{4}$$

where W is a word list. $|W|$ is the size of the word list. $v(\cdot)$ is a word2vec function. The cosine similarity between two vectors was computed by their dot product divided by the product of their norm, as shown in equation (3).

WordNet, ESA and word2vec are among the most popular algorithms in similarity calculation, representing a text sequence by using different information, and all of them have advantages and disadvantages. The WordNet-based method quantifies the similarity based on various relations among words, and heuristics and graph theories. However, the coverage is low. Furthermore, the WordNet-based methods require the mapping of words to concepts; that is, a word sense disambiguation step which could be extremely challenging to learn automatically. The ESA-based approach represents the meaning of words and measures their relationships, and the similarity score is calculated using ESA. However, ESA describes the word based on huge dimension Wikipedia concepts and brings unbearable computation. Compared with ESA, word2vec provides an efficient implementation of learning high-quality vector representations of words with the skip-gram model. Because of its simple model architectures and evidently lower computational complexity compared with other methods, it can compute accurate word vectors from a huge data set with billions of words. Word2vec has been proven one of the state-of-the-art methods in calculating the similarity between two text sequences.

3.3 Re-ranking for diversification

The diversification re-ranking algorithms were inspired by the research results of the SRD algorithm. Existing algorithms for SRD can be divided into two categories: explicit SRD and implicit SRD, which depends on whether the subtopics underlying a query are given beforehand or not (Yu *et al.*, 2017). This paper focused on the implicit SRD algorithms because of a lack of subtopics.

3.3.1 Maximal marginal relevance re-ranking for diversification. MMR was first introduced to accomplish the task of SRD (Carbonell and Goldstein, 1998). MMR was used as the re-ranking method for diversification in this study because it has been proven to be one of the most effective implicit SRD when proper parameters are set (Carbonell and Goldstein, 1998). Many other implicit SRD algorithms are based on the idea of MMR, such as modern portfolio theory (Wang and Zhu, 2009) and facility location analysis (Zuccon *et al.*, 2012). MMR uses a parameter to define the trade-off between relevance and diversity, which is an iterative greedy selection approach to rank the most diverse documents. The formula for it is as follows:

$$Score(C_i) = \underset{C_i \in C}{argmax}\left[\lambda\, sim_1(C_i, q) - (1-\lambda)^{max}_{C_j \in S} sim_2(C_i, C_j)\right] \tag{5}$$

where C_i is the citation context with the highest score in one round of iterative selection, Score(C_i) is the MMR score of C_i in one round of iterative selection. C is the list of candidate citation contexts, S is the re-ranked list, and λ is the coefficient $\lambda \in [0, 1]$. S is updated for every iteration, until completing the iteration. $sim_1(C_i, q)$ is the similarity degree between each citation context and the abstract of the cited article, and $sim_2(C_i, C_j)$ is the similarity degree between different citation contexts. Here, the same similarity degree formula sim (x, y) is set for $sim_1(C_i, q)$ and $sim_2(C_i, C_j)$.

3.3.2 Score difference-based re-ranking for diversification. The score difference (ScoreDiff) method for SRD was proposed in 2012 (Zuccon *et al.*, 2012). Different from MMR, there was no need for multiple iterative calculations in the ScoreDiff algorithm. It does not use any information apart from an initial retrieval run. The semantic similarity scores between a cited article and its citation contexts were used for this initial retrieval run. This approach was beneficial to effectiveness based on the competitive results with other implicit diversification algorithms (Kharazmi *et al.*, 2014). The pseudo algorithm is shown below:

Score difference algorithm

Algorithm 1 ScoreDif Algorithm

1: $C' \leftarrow$ ScoreDiff(C)
2: for $1 \leq i \leq |C|$ do
3: Score($C[i]$) $\leftarrow \frac{1|}{\text{Rank}(C[i])} + \frac{1|}{\text{Rank}(C'[i])}$.
4: end for
5: Sort C on Score($C[i]$)

ScoreDiff(C) represents the difference score between a citation context with the next one from the initial retrieval run, and it is calculated by the following equation:

$$ScoreDiff(C_i) = \frac{|sim(C_{i1}, q) - sim(C_i, q)|}{sim(C_i, q)} \tag{6}$$

Rank($C[i]$) represents the ranking position of C_i in the initial retrieval run. Rank($C[i]$) represents its difference-score ranking position in the ranking result of ScoreDiff(C_i). The final ranking position (Score($C[i]$)) of C_i is re-ranked by the combination of the two parts, which were Rank($C[i]$) and Rank($C[i]$).

3.3.3 Integer linear programming re-ranking for diversification. The shortcomings of classical MMR and MMR-based algorithms are that the first selected document plays an important role in the generation of the subsequent result list, but the choice of the first document was not guaranteed to be the optimal choice. In addition, a single weighting model, an initial retrieval run, and query types also had an important impact on the performance of the implicit SRD (Yu *et al.*, 2017). To solve the above problems, Zuccon *et al.* (2012) proposed the ILP method, which regarded SRD as the process of selecting and re-ranking documents by using ILP. It has proven to perform much better. This process can be expressed as the following formula:

$$\max_x \lambda \cdot (m-k) \cdot R'(x) + (1-\lambda) \cdot k \cdot D'(x) \tag{7}$$

$$R'(x) = \sum_{i=1}^{m} x_{ii} \cdot r(q, d_i) \tag{8}$$

$$D'(x) = \sum_{i=1}^{m} \sum_{j=1:j\neq i}^{m} x_{ij} \cdot s(d_i, d_j) \tag{9}$$

$$\text{s.t. } x_{ij} \in \{0,1\}, i \in \{1,\ldots,m\}, j \in \{1,\ldots,m\} \tag{10}$$

$$\sum_{i=1}^{m} x_{ii} = k \tag{11}$$

$$\sum_{j=1}^{m} x_{ij} = 1, i \in \{1,\ldots,m\} \tag{12}$$

$$x_{jj} - x_{ij} \geq 0, i \in \{1,\ldots,m\}, j \in \{1,\ldots,m\} \tag{13}$$

Specifically, q represents the query that is the abstract of the cited article and d_i and d_j represent the results retrieved by the query; that is, the citation contexts in this study. $r(q, d_i)$ denotes the cosine similarity score between the query and the results, and $R'(x)$ denotes the relevance part. $s(d_i, d_j)$ denotes the diverse score between the results, and $D'(x)$ denotes the diversity part; that is, the cosine similarity score between citation contexts. There are k numbers in the relevance part $R'(x)$ and m–k numbers in the diversity part $D'(x)$, and the coefficients k and m–k are added to avoid the skewness problem when $m \gg k$. The two parts are combined by the parameter, as shown in equation (7). Equation (11) guarantees that k documents are selected, and the restriction given by equation (12) means that each document must vote for another document. The constraint of equation (13) requires that if the document d_i selects d_j as its representation, then d_j will definitely be selected; that is, $x_{jj} = 1$. Once k documents are selected, they were ranked in the decreasing order of their respective contributions to the objective function given by equation (13). The implementation of the algorithm relies on the open-source Python algorithm library and the linear algebra module in the math toolkit SciPy called linprog.

4. Experiments and result analysis

4.1 Data set

CiteSeer^X, an automatic citation index database, contains more than seven million full-text data and its metadata, started from 1970 to the beginning of 2016. It has the unique advantage of providing citation contexts of a cited article for this study. So, Selenium in Python, a web scraping technique, was used to crawl the data from CiteSeer^X. A total of 4,599 articles and all the articles citing them (426,904 citations in total) were collected; the citation count distribution is shown in Table 1. The data was then pre-processed by transforming all the uppercase to lowercase, removing all the punctuations, special characters [1] and stop words [2] in the abstract of each cited article and their citation contexts, and extracting the stems of each word by using the Python NLTK tool. Throughout the rest of the paper, this data set is referred to as the CiteSeer^X data set.

4.2 Settings

4.2.1 Experimental setting. In this paper, the benchmark list was ranked by citation number in CiteSeer^X. Based on the pre-processed data in the last section, nine diversification strategies were set which pair-combined WordNet, ESA and word2vec as semantic distance calculating algorithms and MMR, ScoreDiff and ILP as re-ranking for diversification algorithms in the experiments. They are, respectively, ESA + MMR, WordNet + MMR, word2vec + MMR, ESA + ScoreDiff, WordNet + ScoreDiff, word2vec + ScoreDiff, ESA + ILP, WordNet + ILP and word2vec + ILP. So, for one cited article, there were ten citation context lists to be evaluated.

4.2.2 Experimental procedure. The experimental procedure is shown in the pseudo-code in the whole experiment procedure below. Generally speaking, this experiment was conducted in two phases: one was to carry out the semantic expression of the abstract in the cited article and citation contexts in citing papers and then calculate the semantic similarity between different contexts; the other was to rank citation contexts according to different diversity strategies, and directly or indirectly select the previous 10 items to recommend to users:

Algorithm 1 Rerank for Citation Contexts

```
Require: Abstract of a literature A as Aa; A citation-contexts set of A as
    Cn, n represents the number of citation contexts of Cn;
Ensure: Previous 10 items in re-rank list of Cn;
  for each i ∈ [1, n] do
     calculate the similarity between A and Ci with WordNet (label result as SWN(A, Ci)(1)), word2vec(SWV(A, Ci)(2)) and ESA(SESA(A, Ci)(3)).
    for each j ∈ [1, n] and j 6 i do
      calculate the similarity between Ci and Cj with  WordNet(SWN(Ci, Cj)(4)), word2vec(SWV(Ci, Cj)(5)) and  ESA(SESA(Ci, Cj)(6)).
    end for
  end for
  if MMR is the selected diverse algorithm, then
     the max similarity in (1)/(2)/(3) is the first item in the re-rank list.
    for each j ∈ [1, n − 1] do
```

Citation count	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100
Ratio	0.24	0.21	0.14	0.12	0.09	0.07	0.05	0.04	0.04

Table 1. Citation count distribution in the data set

add the minimal similarity item C_j in (4)/(5)/(6) into the re-rank list.
end for
select previous 10 items in the re-rank list as the result of diverse strategy.
end if
if Score Diference is the selected diverse algorithm, then
rank items with (1)/(2)/(3) and label the ranking number of each item as RCi.
calculate the ScoreDif(C_i) which is the difference degree between
$C_i - 1$ and C_i. Rank items with (1)/(2)/(3) and record the ranking number of each item as RCDi.
for each $i \epsilon$ [1, n] do
add the sum of 1/RCi and 1/RCDi as Score(C_i)
end for
rank Score(Ci) and select previous 10 items as the result.
end if
if ILP is the selected diverse algorithm, then
regard n citation contexts as a n by n matrix. This problem is to find
the items that maxmize the score of the linear programming equation
by adding some constraints to the matrix.
end if (1)/(2)/(3) and (4)/(5)/(6) are all used in the calculation.

4.3 Evaluation

To evaluate the above strategies for citation context diversification, several experiments were conducted on the CiteSeer^X data set and a subset for evaluation was constructed by using a user case study in which the best method was judged by respondents. There were two phases to conduct the evaluation process, which were pre-investigation and formal investigation.

4.3.1 Demographics of the respondents. The evaluation work in this study required respondents to read and compare professional context in the subject of information retrieval. So, there were some requirements for them on their research interests and English reading skill. Therefore, 20 PhD students were invited in related subjects to participate in this evaluation work. Among them, 75% of them were male. They all have at least two years of research experience and one of them had more than five years of research experience. Their research interests included "Information Retrieval", "User Information Behaviour", "Deep Learning", "Information Organization", "Text Mining", "Entity-based Text Analysis" and so on.

4.3.2 Evaluation data set. The performance of the 10 diversification strategies was evaluated by randomly selecting 30 articles from the CiteSeer^X dataset, which had citation counts that ranged from 11 to 78. Each article was evaluated by at least two respondents. The more frequent an article was cited, the more respondents would evaluate it. The authors received 116 questionnaires. Each questionnaire has five 10 citation context lists. So, there were 580 citation context lists to be evaluated. Not only do respondents need to be good at English but also they need to understand and compare the quantity and quality of professional information expressed by different re-rank lists. Therefore, the selection of articles and respondents increased the difficulty of the evaluation process.

A survey instrument for each article was developed which contained the re-ranking list with the top 10 citation context lists of the five strategies, as well as several questions following them (Table 2).

4.3.3 Pre-investigation procedure. In this study, two PhD students in the IR field were recruited from a state university in the USA to judge the performance of the ten diversification strategies in each article. They were asked to exclude five out of the ten strategies by judging how much duplicate content was in each list. To test the reliability of the survey, Cohen's kappa score was calculated for the two annotators before conducting the formal investigation. The formula is as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{14}$$

$P(A)$ is the observed agreement of the annotators, and $P(E)$ is the expected agreement. Cohen's kappa score is 0.752, which satisfies the requirements of reliability (Viera and Garrett, 2005). The five chosen strategies are "ESA + MMR", "ESA + ScoreDiff", "WordNet + MMR", "word2vec + ScoreDiff" and "word2vec + ILP".

4.3.4 Formal investigation procedure. Subsequently, 20 PhD students and faculty with a computer science background were recruited to evaluate each article. They were asked to evaluate the performance of each strategy in terms of the following three indicators:

(1) *Readability*: whether the citation context is easy to scan or understand (questions S1–S3 in Table 2).

(2) *Diversification*: whether each citation context in the list is different from the others (in terms of content or expression), and whether the list of citation contexts as a whole covers the main information of the cited article (questions S4–S6 in Table 2).

(3) *Usefulness*: whether the list of citation contexts will help users in academic writing, and whether users will use it or recommend others use it (questions S7 and S8 in Table 2).

Indicator	Question ID	Question description
Readability	S1	This sorted list is easy to read
	S2	This sorted list increases my reading interest in the top-ranked citing articles
	S3	I completely understand the contents of this sorted list
Diversity	S4	This sorted list can cover the main information of the cited article
	S5	The citation contexts in this sorted list do not duplicate each other and each of them provides new information
	S6	The number of citation contexts for this sorted list is appropriate. If fewer, there will be information missed; if more, it will create redundancy
Usefulness	S7	This sorted list is of high quality. Each citation context only covers a certain aspect of the cited article without redundant information
	S8	This sorted list can be useful for my future research activities (for example, rapid expansion of literature and easy understanding of the cited article)

Table 2.
Questions following in each strategy

Each indicator was represented via the questions in Table 2, which were measured on a two-point Likert scale ranging from agree (1) to disagree (0). They needed to answer the questions after reading through the five lists of citation contexts.

Afterwards, participants answered some open questions:

Q1. How can the citation context list be further improved? What criteria do you need to sort the citation contexts?

Q2. What can a diversification citation context list be used for?

Q3. Which citation context lists were designed to explore the potential improvements?

Q4. What are some applications for the citation context diversification task?

Data was collected to conduct statistical analysis on which strategy performed better from the user perspective, and it was compared to which strategy provided a better user experience in terms of the three indicators. All 116 questionnaires were recovered.

Table 3 shows the evaluation results of the different diversification strategies. The overall score of each re-ranking strategy was calculated by the average score of all three indicators obtained by the two-point scale on the 116 questionnaires, see Score(S-O) in equation (15). Score(R), Score(D) and Score(U) represent the score of the three indicators readability, diversity and usefulness, respectively. Number(Resp) represents the number of respondents who participated in the evaluation. The score of each indicator of a specific strategy was calculated by the average score of whose sub-questions, see Score(S-R), Score (S-D) or Score(S-U) in equations (16)–(18). Number(Rq), Number(Dq) or Number(Uq) represent the number of questions under each indicator. The results showed scores between 0 and 1 were obtained:

$$Score(S-O) = \frac{Score(R) + Score(D) + Score(U)+}{Number(Resp) \times 8} \tag{15}$$

$$Score(S-R) = \frac{Score(R)}{Number(Resp) \times Number(Rq)} \tag{16}$$

$$Score(S-D) = \frac{Score(D)}{Number(Resp) \times Number(Dq)} \tag{17}$$

Strategy	Readability	Diversity	Usefulness	Overall
Strategy 1	0.67	0.43	0.61	0.56*
Strategy 2	0.72***	0.47	0.63*	0.61***
Strategy 3	0.74***	0.45	0.66**	0.61***
Strategy 4	0.60	0.49	0.59	0.55
Strategy 5	*0.81****	*0.61****	*0.77****	*0.72****

Notes: Strategy 1 = ESA + MMR; Strategy 2 = ESA + ScoreDiff; Strategy 3 = WordNet + MMR; Strategy 4 = word2vec + ScoreDiff; Strategy 5 = word2vec + ILP; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 3.
Evaluation results

$$Score(S-U) = \frac{Score(U)}{Number(Resp) \times Number(Uq)} \tag{18}$$

They were indicative enough to compare the performance of the remaining five strategies.

4.4 Results and analysis

Since evaluating 30 articles with 10 rank lists for each manually is a very time-consuming process, the authors conducted a pre-investigation and comparison study of the ten strategies to exclude five of them. According to the results, it was found that the benchmark list which was ranked by citation number has been excluded. In other words, the strategies proposed in this paper (Strategies 1–5) have better performance than the benchmark list. As mentioned in the previous section, the strategies were evaluated from three aspects: readability, diversity and usefulness. To generate comparable results, the scores were averaged from all participants for the five strategies in terms of the three indicators and the scores were normalized to a value between 0 and 1. Table 3 shows the results.

As can be seen from the overall results, the diversifying citation recommendation was proven to be a better strategy compared to the citation count-based recommendation provided by CiteSeer^{X}. Meanwhile, the performance of the combination of different semantic distance algorithms and diversity algorithms varies.

5. Discussion

By furthering analysing the results from the following four essential aspects, it was found that the research results are valuable and can provide clear guidance for future applications:

(1) Which of the diversification strategies is the best for the diversifying citation recommendation task? As mentioned previously, compared with existing studies, this paper focused on the citation context list extracted from the citing articles rather than in the original paper. This task is valuable for those users who want to decide how to cite an article and authors who are willing to know how their articles have been cited quickly. Therefore, it is essential to know which strategy is the best because in this article 10 strategies were explored. Results show that Strategy 5, which combines "word2vec" and "ILP" achieved the best performance. It has an obviously better score than the other strategies under the independent sample *t*-test. This conclusion was supported by previous studies in that word2vec tends to uncover more of certain types of semantic relations than WordNet (Handler, 2014) and word2vec provides an efficient implementation of learning high-quality vector representations of words with the skip-gram model compared with ESA (Wang *et al.*, 2015). Therefore, word2vec has been a popular and widely used semantic distance algorithm in the past decade. As for the diversification re-ranking algorithms, ILP can maximize the resulting objective function. One major drawback of MMR and ScoreDiff is that they are non-optimal because the decision is made based on the scores at the current iteration (Kharazmi *et al.*, 2014).

Comparing Strategies 1 and 2, the diversification re-ranking algorithm ScoreDiff performed better than MMR when the similarity between contexts was calculated by the semantic distance algorithm ESA. This can be explained by the fact that ScoreDiff only uses the initial rank score which is dissimilar between the different citation contexts when expressed by the different extents between them, and this different extent is more accurate than the dissimilar score. From the evaluation results of Strategies 1 and 3, which were expressed by different semantic distance algorithms but were re-ranked by the same re-ranking algorithm, it was found that WordNet performed better than ESA; a semantic distance algorithm based on Wikipedia proved that it outperforms WordNet on some data sets (Ponzetto and Strube, 2007). It indicates that ESA and WordNet may have different performance on different data sets. Finally, comparing the results of Strategies 2 and 4, the semantic distance algorithm ESA performed better than word2vec. The reason was that the re-ranking algorithm ScoreDiff only needs the similarity between abstract and citation context, but the semantic distance algorithm word2vec did not reflect its advantages. Other semantic distance algorithms combining with ILP were discarded in the pre-investigation procedure. Based on the above analysis, it was concluded that Strategy 5 was the best strategy for this task:

(2) Which of the strategies performed better in terms of different evaluation indicators? Since different strategies for diversification have their advantages and disadvantages, it is worth knowing the performance on each indicator, because users might only care about one of the indicators among readability, diversity and usefulness. Strategy 5 also achieved the highest score on all three of the indicators. Independent sample *t*-tests were conducted on these indicators between the five strategies. The results show that Strategy 5 performed much better than the other strategies on these indicators. As for the indicator Readability, Strategies 2, 3 and 5 performed much better than the other two strategies. For the indicator Diversity, only Strategy 5 performed much better than the other strategies. For the indicator Usefulness, Strategy 5 performed much better than the other strategies, and Strategies 2 and 3 also performed well. Comparing Strategies 2 and 3, it was found that Strategy 3 has a higher score than Strategy 2 on the indicators readability and usefulness, but not on the indicator diversity. However, Strategy 3's overall score is lower than that of Strategy 2, indicating that the diversity of the citation contexts has an important impact on the overall score.

(3) What is the user's attitude towards this novel diversifying citation recommendation task and their expectations in the future? Since this paper proposed a new task which has never been studied before, user feedback will give potential directions for improving the task as well as the algorithms. On the one hand, the open survey affirmed the fact that the diversified citation context lists helped the respondents to understand the cited article better. The lists then assisted them in quickly deciding whether to cite, what to cite and how to cite this article. In other words, these citation context lists have improved the efficiency of literature reading and writing for the researchers. They envisioned that this method can help them to save time in writing the related work section, evaluate the cited paper in a new way, keep track of what cited article has been added to the citations, or serve as the first step for identifying new research directions in the future.

Second, the respondents also provided valuable comments about this study from two aspects. From the aspect of diverse content, they suggested that this citation context list can also be re-ranked by the timeline and impact of citing articles based on this list. They expected that this diverse list could show the reason why the article is cited, that is the citation function classification of citation context. Meng *et al.* (2017) improved the performance of citation function classification based on the large-scale academic text by mining the semantic features of citation context. In the future, the authors may focus on the citation classification of citation context by applying the research results of Meng *et al.* (2017). They also suggested that this experiment might have a better performance if these citation context lists were given for comparison from a single perspective (such as methods, data processing, evaluation or summarization), rather than on the citing paper as a whole. From the aspect of displaying the lists, they suggested the citation number and the level of journal or conference source information of each citing paper could also be provided on this list. The colourful label which locates the citation in the context is expected, especially for duplicated contexts. In the end, they said that they sometimes had difficulty in understanding the idea of an entire article through only reading the abstract. Indeed, how to express the content of an entire article in a concise way is a difficult problem for evaluation. Their comments will be considered when future research is conducted.

(4) How can the ideas and algorithms in this paper be used for both research and in a practical way? The innovative task could be very useful and promising in many scenarios. The first one will be finer-grained academic search and recommendation. Context-aware academic recommendation (Yang *et al.*, 2018) has been a popular topic recently, however, previous research only focused on recommending relevant documents based on citation texts, ignoring the user's knowledge level information needs. The ideas in this paper cannot only recommend a diverse citation context list for a given article but also recommend similar citation context sentences based on a given citation context; the open survey results also prove the significance of the applications. By comparison and experimental study on different strategies, the best solution was provided to implement the applications.

Another important application is the automatic abstract generation for a single article. Traditional automatic summarization is mainly based on extracting sentences from the article itself, without taking advantage of the citation contexts from the citing articles, which are informative in summarization. The task introduced in this paper not only provided a new perspective of automatic summarization by user investigation but also came up with a citation context diversification algorithm (word2vec as the semantic calculation algorithm and ILP as the re-ranking algorithm for diversification) which can implement these ideas.

In summary, when we talk about citation recommendation, we should consider user information needs from different perspectives. The first scenario is that when a user is working on a manuscript, they don't know what to cite in a specific context, and many existing studies have focused on this task. However, in this scenario, even when a paper is recommended to users, they still need to spend a lot of time in reading and deciding how to cite the paper. Suppose a set of diverse knowledge fragments (how others cited the paper) can be recommended, users might save time during this process since the citation contexts in the citing articles represent the most important contributions of the original article. Another scenario could be that when a researcher publishes a paper which becomes a highly cited paper, they will want to quickly figure out how the paper has been cited. The idea proposed in this article and the strategy explored to implement the proposed idea filled the gap in citation context analysis and applications.

6. Conclusion and future work

In this paper, different strategies were explored to generate diversified citation contexts and they were evaluated by conducting a user case study. Using the CiteSeer^X dataset, a mixed method experiment was designed which integrated content, novelty and semantic information of diversity to generate diverse citation context lists containing most of the core information of the cited paper. Specifically, WordNet, ESA and word2vec, respectively, were used as semantic distance algorithms to calculate the similarity between each citation context and the similarity between a citation context and the abstract of the cited article. Then, the citation context lists were re-ranked by applying MMR, ScoreDiff and ILP as diversifying algorithms to generate diverse citation context lists that each contained 10 items. In other words, 10 diversification strategies were designed and implemented. In the evaluation phase, five strategies were first excluded out of the ten possible strategies. Then, each generated citation context list of 30 IR articles were recommended to users and they were then asked questions related to the readability, diversity and usefulness of the list to evaluate the remaining five strategies. The experimental results showed that the proposed approach generated a more diverse citation context list with a better user reading experience than the original list presented by CiteSeer^X. Moreover, among the five diversification strategies which combined ESA, WordNet, word2vec and MMR, ScoreDiff, ILP, word2vec + ILP performed the best, allowing the improved removal of redundancy in the citation context list.

It is expected that a diverse set of citation contexts can be provided in an academic retrieval system. In the search results page for literature, the user can also see a diversified collection of citation context lists, not only the summary and citation information. The authors believe that to improve the efficacy of academic writing and the user's reading experience, recommending a citation context list is better than a list of relevant articles.

In the future, the authors will try to further improve the diversification performance and provide higher-quality citation context lists by exploring the usefulness of some other features, such as cite time and the impact factor of the journal where the citation is located. They may also explore deep learning methods to train models on large-scale data sets. To help make reading easier, they will discuss the display style in the next study, by including the label colour, the display location of different features of the citation context and so on. There are several scenarios that can make use of this list; for example, completing tasks, such as text retrieval, text summarization and citation recommendation. The citation context list could be a new way to evaluate cited articles. It may also prove to be a good context-resource to extract pairs of research questions as well as solving methods in the cited article to do some interesting work. To make the algorithm more robust, the authors will collect more articles and involve more users for evaluation.

Notes

1. https://github.com/danielmiessler/SecLists/blob/master/Fuzzing/special-chars.txthttps://github.com/danielmiessler/SecLists/blob/master/Fuzzing/special-chars.txt
2. https://gist.github.com/sebleier/554280https://gist.github.com/sebleier/554280

References

Begg, R.K., Palaniswami, M. and Owen, B. (2005), "Support vector machines for automated gait classification", *IEEE Transactions on Biomedical Engineering*, Vol. 52 No. 5, pp. 828-838.

Bohannon, J. (2016), "Who's the Michael Jordan of computer science? New tool ranks researchers' influence", Science Magazine, available at: www.sciencemag.org/news/2016/04/who-s-michael-jordan-computer-science-new-tool-ranks-researchers-influence (accessed 20 May 2020).

Carbonell, J. and Goldstein, J. (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*, Association for Computing Machinery, New York, NY, pp. 335-336.

Chang, Y.W. (2013), "A comparison of citation contexts between natural sciences and social sciences and humanities", *Scientometrics*, Vol. 96 No. 2, pp. 535-553.

Cohan, A. and Goharian, N. (2017), "Scientific article summarization using citation-context and article's discourse structure", arXiv preprint, 1704.06619.

Cohan, A., Ammar, W., van Zuylen, M. and Cady, F. (2019), "Structural Scaffolds for citation intent classification in scientific publications", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, Association for Computational Linguistics, Stroudsburg, PA, pp. 3586-3596.

Doslu, M. and Bingol, H.O. (2016), "Context sensitive article ranking with citation context analysis", *Scientometrics*, Vol. 108 No. 2, pp. 653-671.

Drosou, M. and Pitoura, E. (2010), "Search result diversification", *ACM SIGMOD Record*, Vol. 39 No. 1, pp. 41-47.

Eskandanian, F., Mobasher, B. and Burke, R. (2017), "A clustering approach for personalizing diversity in collaborative recommender systems", *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, New York, NY, pp. 280-284.

Färber, M., Sampath, A. and Jatowt, A. (2019), "PaperHunter: a system for exploring papers and citation contexts", *Proceedings of 41st European Conference on Information Retrieval*, Springer, Cham, pp. 246-250.

Gabrilovich, E. and Markovitch, S. (2007), "Computing semantic relatedness using Wikipedia-based explicit semantic analysis", *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, Inc., San Francisco, CA, pp. 1606-1611.

Handler, A. (2014), "An empirical study of semantic similarity in WordNet and word2vec", PhD dissertation, available at: https://scholarworks.uno.edu/td/1922/ (accessed 20 May 2020).

Huang, W., Wu, Z., Liang, C., Mitra, P. and Giles, C.L. (2015), "A neural probabilistic model for context-based citation recommendation", Proceedings 29th AAAI Conference on Artificial Intelligence, AAAI Press, pp. 2404-2410.

Jain, A. and Gaur, A. (2017), "Summarizing long historical documents using significance and utility calculation using WordNet", *Imperial Journal of Interdisciplinary Research*, Vol. 3 No. 3.

Kenter, T. and De Rijke, M. (2015), "Short text similarity with word embeddings", Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, pp. 1411-1420.

Kharazmi, S., Sanderson, M., Scholer, F. and Vallet, D. (2014), "Using score differences for search result diversification", *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, pp. 1143-1146.

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H. and Newell, C. (2012), "Explaining the user experience of recommender systems", *User Modeling and User-Adapted Interaction*, Vol. 22 Nos 4/5, pp. 441-504.

Knoth, P., Zilka, L. and Zdrahal, Z. (2011), "Using explicit semantic analysis for cross-lingual link discovery", *Proceedings of the Fifth International Workshop on Cross Lingual Information Access*, Asian Federation of Natural Language Processing, Chiang Mai, pp. 2-10.

Kong, X., Mao, M., Wang, W., Liu, J. and Xu, B. (2018), "VOPRec: vector representation learning of papers with text information and structural identity for recommendation", *IEEE Transactions on Emerging Topics in Computing*, doi: 10.1109/TETC.2018.2830698.

Küçüktunç, O., Saule, E., Kaya, K. and Çatalyürek, Ü.V. (2013), "Result diversification in automatic citation recommendation", Proceedings of the iConference Workshop on Computational Scientometrics: Theory and Applications, pp. 1-4.

Liang, S., Ren, Z. and De Rijke, M. (2014), "Fusion helps diversification", *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, pp. 303-312.

Liang, S., Yilmaz, E., Shen, H., Rijke, M.D. and Croft, W.B. (2017), "Search result diversification in short text streams", *ACM Transactions on Information Systems*, Vol. 36 No. 1, pp. 1-35.

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K. and Lin, Y. (2014), "Literature retrieval based on citation context", *Scientometrics*, Vol. 101 No. 2, pp. 1293-1307.

Loper, E. and Bird, S. (2002), "NLTK: the natural language toolkit", arXiv preprint, cs/0205028.

Ma, S., Zhang, C. and Liu, X. (2020), "A review of citation recommendation: from textual content to enriched context", *Scientometrics*, Vol. 122 No. 3, pp. 1445-1472.

Meng, R., Lu, W., Chi, Y. and Han, S. (2017), "Automatic classification of citation function by new linguistic features", iConference 2017 Proceedings, pp. 826-830.

Mikolov, T. Chen, K. Corrado, G. and Dean, J. (2013a), "Efficient estimation of word representations in vector space", arXiv preprint, 1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013b), "Distributed representations of words and phrases and their compositionality", *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, Curran Associates, Inc., Red Hook, New York, NY, pp. 3111-3119.

Ponzetto, S.P. and Strube, M. (2007), "Knowledge derived from Wikipedia for computing semantic relatedness", *Journal of Artificial Intelligence Research*, Vol. 30 No. 1, pp. 181-212.

Qazvinian, V. and Radev, D.R. (2008), "Scientific paper summarization using citation summary networks", arXiv preprint, 0807.1560.

Qazvinian, V., Radev, D. and Özgür, A. (2010), "Citation summarization through keyphrase extraction", *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, pp. 895-903.

Rahutomo, F. and Aritsugi, M. (2014), "Econo-ESA in semantic text similarity", *SpringerPlus*, Vol. 3 No. 1, p. 149.

Santos, R.L., Macdonald, C. and Ounis, I. (2010a), "Exploiting query reformulations for web search result diversification", *Proceedings of the 19th International Conference on World Wide Web*, Association for Computing Machinery, New York, NY, pp. 881-890.

Santos, R.L., Peng, J., Macdonald, C. and Ounis, I. (2010b), "Explicit search result diversification through sub-queries", Proceedings of European Conference on Information Retrieval, Springer, Cham, pp. 87-99.

Shajalal, M., Aono, M. and Azim, M.A. (2018), "Aspect-based query expansion for search results diversification", *2018 Joint 7th International Conference on Informatics, Electronics and Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision and Pattern Recognition (icIVPR)*, IEEE, pp. 147-152.

Singh, M., Patidar, V., Kumar, S., Chakraborty, T., Mukherjee, A. and Goyal, P. (2015), "The role of citation context in predicting long-term citation profiles: an experimental study based on a massive bibliographic text dataset", Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, pp. 1271-1280.

Sugiyama, K. and Kan, M.Y. (2013), "Exploiting potential citation papers in scholarly paper recommendation", *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 153-162.

Tandon, N. and Jain, A. (2012), "Citation context sentiment analysis for structured summarization of research papers", *Proceedings of the 35th German Conference on Artificial Intelligence*, Springer, Cham, pp. 24-27.

Teufel, S., Siddharthan, A. and Tidhar, D. (2006), "Automatic classification of citation function", *Proceedings of the 11th Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, pp. 103-110.

Tian, H. and Zhuo, H.H. (2017), "Paper2vec: citation-context based document distributed representation for scholar recommendation", arXiv preprint, 1703.06587.

Ullah, M.Z., Shajalal, M., Chy, A.N. and Aono, M. (2016), "Query subtopic mining exploiting word embedding for search result diversification", Asia Information Retrieval Symposium, Springer, Cham, pp. 308-314.

Viera, A.J. and Garrett, J.M. (2005), "Understanding interobserver agreement: the kappa statistic", *Family Medicine*, Vol. 37 No. 5, pp. 360-363.

Wan, S., Paris, C. and Dale, R. (2009), "Whetting the appetite of scientists: producing summaries tailored to the citation context", Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, Association for Computing Machinery, New York, NY, pp. 59-68.

Wang, X., Xiao, Y. and Wang, W. (2015), "Shortest path and word vector based relation representation and clustering", *Proceedings of the International Conference on Web-Age Information Management*, Springer, Cham, pp. 577-580.

Wang, J., Zhang, C., Zhang, M. and Deng, S. (2018), "CitationAS: a tool of automatic survey generation based on citation content", *Journal of Data and Information Science*, Vol. 3 No. 2, pp. 20-37.

Wang, J. and Zhu, J. (2009), "Portfolio theory of information retrieval", *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, pp. 115-122.

Willemsen, M.C., Knijnenburg, B.P., Graus, M.P., Velter-Bremmers, L.C. and Fu, K. (2011), "Using latent features diversification to reduce choice difficulty in recommendation lists", in Felfernig, A., Chen, L., Mandl, M., Willemsen, M., Bollen, D. and Ekstrand, M. (Eds), *The RecSys 2011 Workshops - Decisions@RecSys 2011 and UCERSTI-2*, pp. 14-20.

Wu, J., Huang, J.X. and Ye, Z. (2014), "Learning to rank diversified results for biomedical information retrieval from multiple features", *Biomedical Engineering Online*, Vol. 13 No. Suppl 2, p. S3.

Wu, J., Williams, K.M., Chen, H.H., Khabsa, M., Caragea, C., Tuarob, S., Ororbia, A.G., Jordan, D., Mitra, P. and Giles, C.L. (2015), "CiteSeer^{X}: AI in a digital library search engine", *AI Magazine*, Vol. 36 No. 3, pp. 35-48.

Xia, L., Xu, J., Lan, Y., Guo, J., Zeng, W. and Cheng, X. (2017), "Adapting Markov decision process for search result diversification", *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, pp. 535-544.

Yang, L., Zheng, Y., Cai, X., Dai, H., Mu, D., Guo, L. and Dai, T. (2018), "A LSTM based model for personalized context-aware citation recommendation", *IEEE Access*, Vol. 6, pp. 59618-59627.

Yu, H.T., Jatowt, A., Blanco, R., Joho, H., Jose, J., Chen, L. and Yuan, F. (2017), "A concise integer linear programming formulation for implicit search result diversification", *Proceedings of the 10th*

ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, NY, pp. 191-200.

Zuccon, G. and Azzopardi, L. (2010), "Using the quantum probability ranking principle to rank interdependent documents", *Proceedings of European Conference on Information Retrieval*, Springer, Cham, pp. 357-369.

Zuccon, G., Azzopardi, L., Zhang, D. and Wang, J. (2012), "Top-k retrieval using facility location analysis", *Proceedings of European Conference on Information Retrieval*, Springer, Cham, Switzerland, pp. 305-316.

Corresponding author
Wei Lu can be contacted at: weilu@whu.edu.cn