eISSN : 2287-4577
pISSN : 2287-9099
http://www.jistap.org
Vol. 9 No. 2 June 30, 2021

Journal of Information Science Theory and Practice

Journal of Information Science Theory and Practice

JISTaP

01 Word Embeddings-Based Pseudo Relevance Feedback Using Deep Averaging Networks for Arabic Document Retrieval
18 Social Media News in Crisis? Popularity Analysis of the Top Nine Facebook Pages of Bangladeshi News Media
33 Drive-thru Library Service in Korea
47 Topic Analysis of Scholarly Communication Research
66 Issues and Challenges in the Extraction and Mapping of Linked Open Data Resources with Recommender Systems Datasets
83 Influencing Factors of Research Collaboration Intention in Virtual Academic Communities in China

Vol. 9 I No. 2 I June 30, 2021

KiSTi
www.kisti.re.kr

Indexed/Covered by SCOPUS, LISA, DOAJ, and CrossRef

JISTaP
Journal of Information Science Theory and Practice

General Information

Aims and Scope

The Journal of Information Science Theory and Practice (*JISTaP*) is an international journal that aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics. JISTaP will be published quarterly, issued on the 30th of March, June, September, and December. JISTaP is indexed in the Scopus, Korea Science Citation Index (KSCI) and KoreaScience by the Korea Institute of Science and Technology Information (KISTI) as well as CrossRef. The full text of this journal is available on the website at http://www.jistap.org

Indexed/Covered by

Scopus LISA DOAJ DIRECTORY OF OPEN ACCESS JOURNALS Crossref Cited-by

Publisher

Korea Institute of Science and Technology Information
66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea
(T) +82-2-3299-6102
(F) +82-2-3299-6067
E-mail: jistap@kisti.re.kr
URL: http://www.jistap.org

Managing Editor: Ji-Young Kim, Eungi Kim

Copy Editor: Ken Eckert

Design & Printing Company: MEDrang Inc.

31st Floor, Prime Center, 85 Gwangnaru-ro 56-gil, Gwangjin-gu, Seoul 05116, Korea
(T) +82-2-325-2093
(F) +82-2-325-2095
E-mail: info@medrang.co.kr
Homepage: www.medrang.co.kr

Open Access and Creative Commons License Statement

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and au-thors, permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

This paper meets the requirements of KS X ISO 9706, ISO 9706-1994 and ANSI/NISO Z39.48-1992 (Permanence of Paper).

2021 Copyright © Korea Institute of Science and Technology Information

JISTaP
Journal of Information Science Theory and Practice

Editorial Board

Co-Editors-in-Chief

Gary Marchionini
University of North Carolina, USA

Dong-Geun Oh
Keimyung University, Korea

Co-Associate Editors

Kiduk Yang
Kyungpook National University, Korea

Taesul Seo
Korea Institute of Science and Technology Information, Korea

Co-Managing Editors

Ji-Young Kim
Korea Institute of Science and Technology Information, Korea

Eungi Kim
Keimyung University, Korea

Editorial Board

Dan Albertson
University at Buffalo, SUNY, USA

Daniel Martínez Ávila
San Paulo State University, Brazil

Beeraka Ramesh Babu
University of Madras, India

Pia Borlund
University of Copenhagen, Denmark

France Bouthillier
McGill University, Canada

Kathleen Burnett
Florida State University, USA

Ina Fourie
University of Pretoria, South Africa

Hamid R. Jamali
Charles Sturt University, Australia

Boryung Ju
Louisiana State University, USA

Thomas Mandl
Universiät Hildesheim, Germany

Lokman I. Meho
American University of Beirut, Lebanon

Jin Cheon Na
Nanyang Technological University, Singapore

Dan O'Connor
Rutgers University, USA

Helen Partidge
University of Southern Queensland, Australia

M. P. Satija
Guru Nanak Dev University, India

Ou Shiyan
Nanjing University, China

Paul Solomon
University of South Carolina, USA

Wolfgang G. Stock
Heinrich-Heine-Universität Düsseldorf, Germany

Kulthida Tuamsuk
Khon Kaen University, Thailand

Lihong Zhou
Wuhan University, China

Consulting Editors

Reysa Alenzuela
Oriental Institute, ASCR, Czech Republic

Wayne Buente
University of Hawaii, USA

Sujin Butdisuwan
Mahasarakham University, Thailand

Seon Heui Choi
Korea Institute of Science and Technology Information, Korea

M. Krishnamurthy
DRTC, Indian Statistical Institute, India

S.K. Asok Kumar
The Tamil Nadu Dr Ambedkar Law University, India

P. Rajendran
SRM University, India

Tsutomu Shihota
St. Andrews University, Japan

2021 Copyright © Korea Institute of Science and Technology Information

JISTaP
Journal of Information Science Theory and Practice

Table of Contents

Vol. 9 No. 2 June 30, 2021
Journal of Information Science Theory and Practice • http://www.jistap.org

Articles

01 Word Embeddings-Based Pseudo Relevance Feedback Using Deep Averaging Networks for Arabic Document Retrieval
Yasir Hadi Farhan, Shahrul Azman Mohd Noah, Masnizah Mohd, Jaffar Atwan

18 Social Media News in Crisis? Popularity Analysis of the Top Nine Facebook Pages of Bangladeshi News Media
Md. Sayeed Al-Zaman, Mridha Md. Shiblee Noman

33 Drive-thru Library Service in Korea
Seong-Kwan Lim

47 Topic Analysis of Scholarly Communication Research
Hyun Ji, Mikyeong Cha

66 Issues and Challenges in the Extraction and Mapping of Linked Open Data Resources with Recommender Systems Datasets
Rosmamalmi Mat Nawi, Shahrul Azman Mohd Noah, Lailatul Qadri Zakaria

83 Influencing Factors of Research Collaboration Intention in Virtual Academic Communities in China
Chunlai Yan, Hongxia Li

Call for Paper

Information for Authors

2021 Copyright © Korea Institute of Science and Technology Information

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(2): 1-17, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.2.1

Word Embeddings-Based Pseudo Relevance Feedback Using Deep Averaging Networks for Arabic Document Retrieval

Yasir Hadi Farhan*
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: yasir.hadi87@yahoo.com

Shahrul Azman Mohd Noah
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: shahrul@ukm.edu.my

Masnizah Mohd
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
E-mail: masnizah.mohd@ukm.edu.my

Jaffar Atwan
Prince Abdullah Bin Ghazi, Faculty of Information Technology, Al Balqa Applied University, Salt, Jordan
E-mail: jaffaratwan@bau.edu.jo

ABSTRACT

Pseudo relevance feedback (PRF) is a powerful query expansion (QE) technique that prepares queries using the top *k* pseudo-relevant documents and choosing expansion elements. Traditional PRF frameworks have robustly handled vocabulary mismatch corresponding to user queries and pertinent documents; nevertheless, expansion elements are chosen, disregarding similarity to the original query's elements. Word embedding (WE) schemes comprise techniques of significant interest concerning QE, that falls within the information retrieval domain. Deep averaging networks (DANs) defines a framework relying on average word presence passed through multiple linear layers. The complete query is understandably represented using the average vector comprising the query terms. The vector may be employed for determining expansion elements pertinent to the entire query. In this study, we suggest a DANs-based technique that augments PRF frameworks by integrating WE similarities to facilitate Arabic information retrieval. The technique is based on the fundamental that the top pseudo-relevant document set is assessed to determine candidate element distribution and select expansion terms appropriately, considering their similarity to the average vector representing the initial query elements. The Word2Vec model is selected for executing the experiments on a standard Arabic TREC 2001/2002 set. The majority of the evaluations indicate that the PRF implementation in the present study offers a significant performance improvement compared to that of the baseline PRF frameworks.

Keywords: automatic query expansion, information retrieval, word embedding, deep averaging networks, pseudo relevance feedback, Arabic document retrieval on TREC collection

Received: March 16, 2021 **Revised:** May 5, 2021
Accepted: May 10, 2021 **Published:** June 15, 2021

***Corresponding Author:** Yasir Hadi Farhan
https://orcid.org/0000-0002-2378-9177
E-mail: yasir.hadi87@yahoo.com

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© Yasir Hadi Farhan, Shahrul Azman Mohd Noah, Masnizah Mohd, Jaffar Atwan, 2021

1. INTRODUCTION

Search engines presently face the challenge of users' queries comprising two to three words, which are considered too little for a query (Berget & Sandnes, 2015). They are typically the consequence of users' challenges in expressing the requirement for information; these queries are not truly representative of users' information needs. The anomalous state of knowledge hypothesis by Belkin et al. (1982) suggests that query challenges are typically associated with users facing knowledge gaps. Automatic query expansion (AQE) is among the extensively-researched techniques to bridge that gap and facilitate better search queries by users (Azad & Deepak, 2019; Dalton et al., 2019).

AQE comprises global and local techniques, where the global techniques rely on a thesaurus to augment the original query instead of using retrieved results. Usually, WordNet is the most extensively used tool for identifying the new potential terms correlated with the terms of the initial query (Pal et al., 2014). On the other hand, local techniques rely on relevance-based information, where the first retrieved document are analysed and utilised to identify the most appropriate terms for augmenting the initial query (Miyanishi et al., 2013; Takeuchi et al., 2017). Pseudo relevance feedback (PRF) is a technique whereby it is presupposed that the *k* initial top documents yielded by the search are most relevant to identify new elements from the *k* documents; it is assumed to be a better expansion method since it facilitates automatic relevance feedback.

PRF is a potent query expansion (QE) technique to handle vocabulary mismatch between documents and user queries, which are then extended by identifying associated terms from the documents in the top results, also known as top pseudo-relevant documents (Farhan et al., 2020). Typically, PRF methods offer adequate performance (Carpineto et al., 2001; Clinchant & Gaussier, 2013); however, they rely on expansion term distribution pertaining to the pseudo-relevant document set, where the similarity shared by the original query elements and the expansion elements is not strongly considered. Fundamentally, in an ideal scenario, expansion elements should be identified considering their similarity to the query terms and their distribution pertaining to the pseudo-relevant document set. Several studies have proposed similar recommendations (Atwan et al., 2016; El Mahdaouy et al., 2019; Montazeralghaem et al., 2016).

Nevertheless, some of the discussed methodologies suffer from inherent challenges, such as the WordNet and other ontological databases returning keywords having several meanings; hence, QE should be performed only after word disambiguation (Farhan et al., 2020).

AQE research is of great interest to academicians who intend to utilise word embedding (WE) for semantic modelling to interpret text (ALMasri et al., 2016; Diaz et al., 2016; Roy et al., 2016) and handle the problems specified previously.

Under the natural language processing (NLP) domain, the umbrella term WE refers to modelling methods, where every term is denoted using a real-number vector, which means single-dimension mathematical WE is employed in a continuous vector space having substantially lower dimensions, where the embedding frameworks depend on proximity for corpus training. Word2Vec provides the skip-gram and continuous bag-of-words (CBOW) architectures for training the WE framework. CBOW uses proximal words for predicting the target, whereas skip-gram relies on the target to identify proximal words.

Nevertheless, considering the similar contexts shared by semantic and syntactically similar words, the objective of such training is to identify terms sharing syntactic and semantic similarity. Consequently, the *N* proximal terms may be combined, or neighbouring query-term techniques are employed for expansion in several AQE techniques (Roy et al., 2016).

1.1. The Need for Study

A majority of the AQE techniques that use WE rely on a single term to identify potentially identical terms, and fundamentally, the choice of candidate terms does not depend on the terms specific to the initial query. Nevertheless, we asserted that the need to enhance AQE effectiveness requires the query's semantic aspects to be designed entirely considering the potential vocabulary terms. Subsequently, the quality of the terms associated with queries can be enhanced; moreover, new terms having semantic correlation and high relevance can be added. Deep averaging networks (DANs) comprise a straightforward deep unordered framework that identifies the average WE in a piece of text, and subsequently, is classified using multiple linear stages.

Recently, WE has witnessed several developments for use in information retrieval (IR); however, representation concerning Arabic information retrieval (AIR) is being assessed. The extensive and complex structure of Arabic is among the extensively researched aspects concerning AIR frameworks, and despite its significance, Arabic lacks

 https://doi.org/10.1633/JISTaP.2021.9.2.1

sufficient NLP recognition (Alsmearat et al., 2014; Faqeeh et al., 2014). The fundamental challenge concerning Arabic is the scarcity of ontological or thesaurus knowledge-bases (Mohsen et al., 2018). Additionally, limited research has been conducted on the NLP side for diacritical Arabic compared to other languages such as English (Farghaly & Shaalan, 2009). To improve its efficacy, further studies are required due to weaknesses in conventional IR processes and techniques, particularly the Arabic system limitations.

Consequently, most AIR-specific research emphasises assessing or contrasting word-stemming methods (Abu El-Khair, 2007; Ben Guirat et al., 2016; Darwish & Mubarak, 2016; Larkey et al., 2002; Mustafa et al., 2008) that assess queries or documents to determine common-stemmed words used as the metric for ranking documents.

1.2. The Objective

Hence, we propose that the string-embedding method DANs be used to enhance AQE efficacy for Arabic, where the query vectors' average can be utilised with the PRF method to produce the potential expansion terms. Our objective is to enhance Arabic text retrieval by using the proposed DANs-based PRF technique, where the average vector of each query term's vectors has been used to find the candidate expansion vectors.

The present study uses the WE technique Word2Vec for training purposes, and the Okapi BM25 probabilistic framework has been used along with V2Q and EQE1 approaches for comparisons.

Furthermore, the proposed framework incorporates the BM25 probabilistic scheme along with EQE1 suggested by Zamani and Croft (2016) and V2Q formulated by Fernández-Reyes et al. (2018).

1.3. The Hypothesis

The present paper hypothesises that the DANs-based PRF technique, which utilises the similarity of the WE in generating the candidate expansion vectors, may be able to address the AIR term mismatch problem; consequently, AQE Arabic text retrieval performance can be enhanced. The *k* top documents identified during PRF are utilised to form the expansion vectors of the candidate terms.

2. RELATED WORKS

IR relates to storing, accessing, organising, and presenting information. IR systems are primarily used to offer users ease of access to information. Typically, users present their information requirements using queries that are processed by IR systems to provide the most relevant information.

There are several challenges during an interaction between IR systems and users; vocabulary mismatch is one of the significant ones (Carpineto & Romano, 2012; Farhan et al., 2020). Researchers have responded to this challenge by proposing several solutions including AQE. Schemes based on AQE augment the original query by automatically including new terms considering that the IR system's accuracy will be increased (Abbache et al., 2016). WE is among the leading AQE techniques that have gained widespread attention (ALMasri et al., 2016; Diaz et al., 2016; El Mahdaouy et al., 2019; Roy et al., 2016).

As specified previously, WE is an umbrella term from the NLP domain; it comprises feature learning methods and language modelling. Semantic parsing allows identifying the meaning of the text; consequently, the natural language can be interpreted. WE is one among the semantic vector space frameworks where real number vectors are employed to denote words in the corpus. The words are indicated using distributed or local representation. Local representation is also referred to as sparse representation. It uses the statistical presence of words while disregarding their potential associations. Distributed representation is based on the hypothesis that word representation in the WE vector space is similar for words sharing a context (Bengio, 2009; Kim et al., 2017; Turney & Pantel, 2010).

For instance, "ذكر" (male) and "انثى" (female) are words that share a similar context; consequently, there is a likelihood that the corresponding WE vector space distributions will be proximal. The works of researchers who have employed the WE vector space for word representation are discussed below.

Diaz et al. (2016) introduced an AQE technique by employing topic-guided local embeddings that were obtained from the documents retrieved after query submission. Document scoring requires using the Kullback-Leibler divergence (KLD) language framework. The extracted documents were re-ranked after computing the local embeddings corresponding to the queries. The results indicated that local embeddings provided superior results.

ALMasri et al. (2016) contrasted numerous expansion techniques using deep-learning derived vector representations. The resulting vectors comprise words such as "taxi" and "driver" that have a similar context (Mikolov et al., 2013b). Hence, the corpus comprises terms that are denoted using vectors of specific dimensions. Therefore, the similarity of the two terms is determined by normalising

the cosine similarity corresponding to the two vectors. The skip-gram technique was used to conduct assessments on four CLEF sets. It was established that deep-learning-based high-quality vector characterisation leads to a significant performance improvement compared to baseline language frameworks that lack expansion. Vector-based models were also superior to mutual information and PRF models (Manning et al., 2008).

Roy et al. (2016) suggested a relevance feedback framework that integrates semantic associations by utilising the terms compositionality and matching them with word vectors. Rather than implementing an ad-hoc method for determining the co-occurrence between the query terms and those present in the documents, kernel density estimation (KDE) is employed to methodically ascertain the co-occurrence. The objective is to formulate a systematic relevance feedback framework that offers a straightforward technique to integrate term-specific compositionality and semantic associations. The suggested technique considers query WE as data elements that are surrounded by kernel functions so that a density estimator can be formulated. This estimator attempts to compute the probability density expression that produces the detected query word vectors. Subsequently, KLD may be employed corresponding to the density function for document ranking determination. Further, this technique facilitates the use of term composition during QE. The proposed feedback scheme is assessed using several typical IR test sets, namely Web track test and TREC ad-hoc collections. The analysis indicates that the suggested KDE-based relevance feedback technique provides a statistically significant increase in relevance compared to model-based techniques that rely on quantitative co-occurrence of words and query elements in the top-ranked results.

Aklouche et al. (2018) proposed a QE technique that used NLP along with the Word2Vec toolkit (Mikolov et al., 2013a). In this technique, the TREC Washington Post Corpus is employed to train a CBOW and skip-gram to obtain fresh semantic elements associated with the initial query. Further, the reweighting technique and term selection are used to determine their effects on the retrieval performance. Vector similarity comparisons have been drawn using Euclidean distance. The vector selection is performed using the complete query or its constituent terms. Experimental observations indicated that optimal results were obtained using query reweighting, which was the best technique compared to the others. The retrieval performance is impacted when all the expanded query terms are set to the same weight.

El Mahdaouy et al. (2019) suggested a QE technique that uses the WE semantic similarities and integrates it with IR frameworks such as the Log-Logistic distribution Model, Language Model (LM), Okapi BM25, and smoothed power-law Model (SPL) to enhance Arabic text-retrieval efficacy. The present work proposes CBOW, skip-gram, and the Global Vector (GloVe) model, which are neural WE methods. The present scheme intends to handle term discrepancy by employing a low-dimensional vector space to position words. The suggested method integrates similarity and scoring expression. The initial query is reformulated using resembling terms identified during translation. These resembling terms comprise the most relevant words obtained using the vocabulary corpus. Fang and Zhai (2006) formulated semantic term matching constraints (STMCs) to assess the suggested augmentation. STMC comprises a scheme that facilitates integrating semantic comparability and regulating the weights corresponding to the initial query's words and the matching words. The present study proposes two schemes: The first comprises an assessment of the retrieved documents to identify potentially similar words. The second scheme identifies similar words for the entire corpus. The outcomes indicate that the suggested extension schemes offer a significant performance improvement over the baseline bag-of-words techniques. Further, the suggested method is superior to the three IR language frameworks founded on WE, namely the neural translation language model formulated by Zuccon et al. (2015), the LM-based WE, and the generalised language model (GLM) suggested by Ganguly et al. (2015). Hence, word vector indication is an appropriate beginning point for determining the semantic similarity shared by query terms and potential expansion candidates. Presently used QE methods proceed assuming that all query terms can identify the most useful expansion candidates (Esposito et al., 2020).

Nevertheless, considering that the query context is disregarded when an expansion is conducted, the critical factor is to create a context applicable for retrieving the relevant terms. Several studies such as that of Roy et al. (2016) have tried to address this question; however, the issue has not been fully resolved. Zamani and Croft (2016) and Fernández-Reyes et al. (2018) are two leading works that have attempted to address this challenge. Fernández-Reyes et al. (2018) suggested that terms be identified if they share a strong resemblance to the original query terms. It was hypothesised that there is a likelihood of all query terms not being adequate for expansion; several terms might impact expansion adversely. Hence, the au-

 https://doi.org/10.1633/JISTaP.2021.9.2.1

thors formulated the query-supplemented AQE technique based on the V2Q approach. The scheme comprises filters that disregard unnecessary words from the original query.

Zamani and Croft (2016) suggested that the PRF words which are embedded be based on embedding-specific relevance. This framework builds upon the relevance model technique. The model was evaluated using the TREC test set; the results highlighted a noteworthy performance increase compared to the baseline relevance framework. This technique evaluates the semantic similarity of the terms identified from WE vector similarity comparison. It is assumed that query terms are conditionally independent, thereby indicating that the terms comprising the expanded query should be related to the terms identified for augmenting the initial query. Considering that the present study uses both techniques in the baseline form, these techniques are discussed in detail.

El Mahdaouy et al. (2019) suggested building WE resemblance into the PRF frameworks for AIR. The concept is to choose expansion terms from PRF documents such that there is an alignment considering similarity and distribution with the initial query terms. AIR specific to the PRF model can use WE, because similar words can be included in a set at one side where they are proximal in the vector space. The primary objective is to enhance the weight and semantic-similarity to the terms of the original query. The present study assesses three WE frameworks: CBOW, skip-gram, and GloVe. Further, WE resemblance is integrated into four PRF frameworks, namely the KLD (Carpineto et al., 2001), Log-Logistic, Bo2 of the Family of Divergence from Randomness (Amati & Van Rijsbergen, 2002), and SPL of the information-based family of PRF models (Clinchant & Gaussier, 2013). Assessments were conducted on the TREC 2001/2002 Arabic test set using three neural WE frameworks. The study objective was to comprehend ways to integrate WE in the PRF methods for AIR. The outcomes indicate that the PRF extensions used in the present study offer a significant performance increase over baseline models. Further, an increase of 22% was observed for the elemental IR mean average precision (MAP) model; the robustness index was enhanced by 68%. Moreover, the performance differential of the three WE frameworks, namely skip-gram, CBOW, and GloVe was not statistically significant.

Fernández-Reyes et al. (2018) suggested a global AQE technique where the entire query is considered. This technique gathers pertinent information concerning potential expansion terms and uses it to reduce disambiguation issues and enhance precision and recall metrics. Word2Vec-based representation was initially suggested by Mikolov et al. (2013b); the present study uses this representation form for AQE (ALMasri et al., 2016) because of its computational efficiency. Fernández-Reyes et al. (2018) proposed two AQE techniques: a prospect-guided association scheme (V2Q) and a query-guided association scheme (Q2V). The Q2V technique identifies potential candidates using query terms. Conventional collection methods are used for this scheme. The top N closest candidates are identified and used as pre-selection elements; subsequently, they are subjected to refinement that produces the list of final expansion words. Consequently, this scheme produces $Rank_q$ for every query term q. This list comprises data pairs having words and their respective votes. The Q2V technique produces a candidate term list that is used for choosing expansion terms. After the ranking is completed, the list is sorted so that the expansion term selection methods may be used. Typically, $Rank_q$ is used to identify and select the top terms. Often, potential terms retrieved using the original query match the meaning of the original words; consequently, Q2V may lose the query context. In contrast, V2Q uses an opposite approach compared to the Q2V scheme. This scheme hypothesises that the potential expansion terms must provide votes for every query term; hence, candidates are likely to be selected. Several query terms might affect the expansion adversely; put differently, not all query elements can be expected to enhance expansion. Hence, the suggested V2Q scheme identifies words having a relatively higher semantic similarity to the query terms. Consequently, V2Q scores well over Q2V since the latter employs all query elements to identify candidate terms. There is a likelihood of disambiguation in the query. The suggested techniques enhance the precision and recall without relevance-specific feedback information. Experiments indicated that the outcomes were superior compared to classic IR frameworks such as Vector Space Model, LM, PRF model, and Okapi BM25.

Zamani and Croft (2016) assessed WE use to enhance the query language framework performance for ad-hoc retrieval tasks. An AQE-embedding relevance framework was suggested based on the work of Lavrenko and Croft (2017). The technique relied on semantic similarity at the WE vector level and suggested two different query language framework estimations. The first technique works under the presupposition that the query terms share conditional dependence. On the other hand, the second technique presupposes that two semantically similar words do not depend on the query. Words having weights close to

semantically synonymous query terms but not included in the query are assimilated to augment the original query. These techniques are used for the PRF relevance framework and termed EQE1 and EQE2. Three standard TREC sets, namely Robust (TRE C Robust Track 2004 collection), GOV2 (TREC Terabyte Track 2004-2006 collection), and AP (Associated Press 1988-1989), were used for the experiments. The experiments relied on the GloVe technique of Pennington et al. (2014) to extract WE from six billion tokens contained in Giga-words five and Wikipedia Dump 2014. The feedback document count was 10 for the PRF scenario. The assessment comprised four baselines: (1) standard maximum likelihood estimation for the query framework, (2) a WE (VEXP) heuristic QE technique (ALMasri et al., 2016), (3) document language model smoothing technique based on embedding (GLM) (Ganguly et al., 2015), and (4) a QE technique that used vocabulary term vectors' similarity along with the average embedding vector specific to all query terms. Cosine similarity specific to WE vectors was used for determining semantic similarity between words. The experiments indicated that embedding techniques have a significantly better performance than competitive baselines in most cases, considering MAP and average precision (AP) metrics. V2Q, EQE1, and previously-discussed techniques have a constraint considering the individual query elements.

Nevertheless, research indicates that individually-used query terms employed for QE might cause query drift by deflecting the query context that could be different from those individual query terms (Crimp & Trotman, 2018). Hence, the present study endeavours to assess the effects of AQE using WE on the full query instead of considering the query terms separately. Explicitly, this research focuses on AQE using DANs sentence embedding vectors.

3. METHOD

To augment AIR performance, we suggest using the DANs technique to formulate a WE-specific PRF scheme that integrates WE similarities into present PRF frameworks. The fundamental idea of the proposed technique is that performance enhancement can be obtained if expansion element similarity and distribution in pseudo-relevant documents are attached to the original query vectors using the average vector representation. Expansion information will be gathered using pseudo-relevant documents. Average vectors corresponding to DANs will be sorted, and the top results will be used to augment the original query.

To achieve the desired objective, we used DANs on the present bag-of-words IR framework. Probability-based Okapi BM25 framework is used along with two representative WE techniques, namely, EQE1 and V2Q for AQE. These techniques were formulated by Zamani and Croft (2016) and Fernández-Reyes et al. (2018), respectively.

We selected BM25 owing to its excellent performance in TREC retrieval tasks; furthermore, it has also influenced search engine ranking algorithms (Croft et al., 2010). BM25 is set as the baseline framework for non-AQE output. In contrast, V2Q and EQE1 schemes are used for comparing the output of DANs-based AQE with the presently used AQE schemes suggested by both techniques. Hence, we can evaluate whether DANs can output expansion terms having better relevance during PRF-based document retrieval.

3.1. Word2Vec and DANs

Word2Vec is a neural network-facilitated technique that outputs word vector representations; it was proposed by Mikolov et al. (2013b). It processes text input and provides the corresponding word vectors. The system is trained using training text, after which it learns how to represent words as vectors (Xue et al., 2014). The approach suggested in the present study relies on Word2Vec, that lists two WE model training frameworks: Skip-gram and CBOW. The CBOW framework estimates the middle word using distributed context-specific representations. On the other hand, the skip-gram framework estimates the context using distributed input word representation.

Most predictive frameworks for NLP activities suffer from performance degradation when unknown words that were not part of the training set due to low-frequency or other issues are used for experiments. Such words, referred to as Out-Of-Vocabulary (OOV), can affect NLP performance because of the inability of a representation framework to provide appropriate representation. To address OOV words, Arabic Wikipedia was used to augment the dataset. Hence, words not present previously were made available during training by dataset augmentation.

DANs is an average WE-based sentence embedding technique designed to convert text to vector format (Collobert et al., 2011). It is feasible to train DANs in little time with data (such as Arabic) with high syntactic variance. Embedding assumes a more abstract form as text length increases (Franco-Salvador et al., 2015). DANs combine the high speed and accuracy of unordered (such as Neural Bag-of-Word [NBOW]) and syntactic functions (such as Recursive Neural Networks [RecNN]), respectively.

 https://doi.org/10.1633/JISTaP.2021.9.2.1

Consequently, this approach provides superior results compared to the previously-used bag-of-words scheme. DANs implementation, specific to AQE, is described: the original query is processed using the WE framework, and the vectors are fed as input to DANs; consequently, the average vector is obtained as output. All inputs to DANs are passed across two linear layers. The average vector corresponding to every query term is computed to obtain a single vector or tensor (Iyyer et al., 2015). Fig. 1 depicts this process. Initially, a composition layer computes the average embedding. The intermediate data is fed through a hidden layer set comprising one or more layers; the network transforms the average. Lastly, a soft-max layer processes the intermediate data for prediction. In the case of using DANs for AQE, it follows three simple steps:

Step 1. The average of the embedding is taken, which is associated with the input sequence of tokens (in this case, the query terms).
Step 2. This average will pass through one or more feedforward layers.
Step 3. Implement a linear classification on the final representation of the layer.

The goal of this research is to investigate the impact of expanding a complete query sentence by using the DANs method based on the WE-based PRF technique on AQE. Thus, in the following sections, we illustrate how the DANs are being implemented in the three models, the basic BM25 IR model and the AQE approaches of EQE1 and V2Q. The BM25 is considered the basic approach of applying QE using DANs. We trained the Word2Vec CBOW model using the top-retrieved document of PRF technique on which search is performed. The dataset is the Arabic TREC 2001/2002 collections consisting of news about the Middle East and including articles from May 13, 1994, to December 20, 2000. Search is also performed on the same corpus. Initially, the dataset was divided into three groups: TREC 2001 contains 25 queries, TREC 2002 contains 50 queries, and TREC 2001/2002 contains 75 queries. The queries in TREC 2001/2002 are the combination of both TREC 2001 and TREC 2002 queries. Table 1 provides statistical information on the dataset.

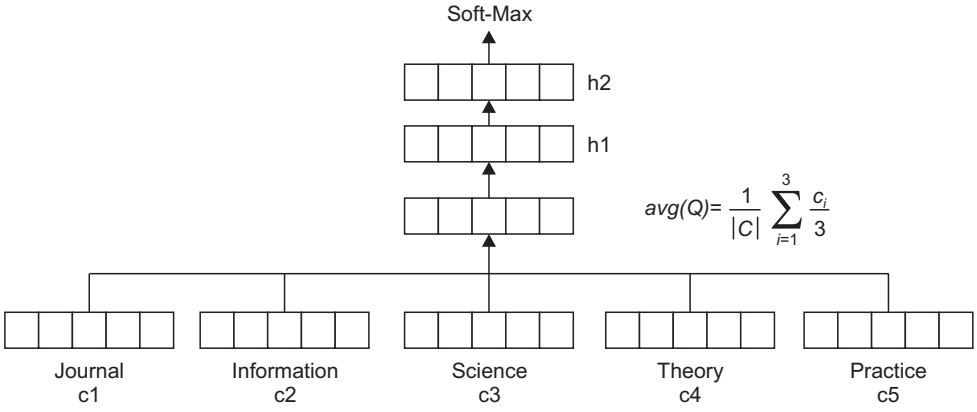


Fig. 1. Deep averaging networks model. Adapted from Iyyer et al., Association for Com putational Linguistics 2015.

Table 1. Statistics for Arabic TREC collections

Collections	TREC 2001	TREC 2002	TREC 2001/2002
Number of queries	25	50	75
The average number of words/queries	4.88	3.28	4.08
Number of documents		383,872	
Number of tokens		76 million	
Number of unique words		666,094	
Size (compressed)		209 MB	
Size (uncompressed)		869 MB	

3.2. Query Expansion Based-PRF using DANs (BM25+DANs-PRF)

The Okapi BM25 (Robertson et al., 1995) model is considered as one of the best-known term-weighting schemes derived from the probabilistic model. It considers three components, namely, the term frequency, inverse document frequency, and the length of the document (Robertson et al., 1995; Trotman et al., 2014; Vaidyanathan et al., 2015). The most common BM25 scoring function is shown in Equations 1 and 2:

$$\sum_{i\in Q} log \frac{\frac{(r_i+0.5)}{R-r_i+0.5}}{\frac{(n_i-r_i+0.5)}{(N-n_i-R+r_i+0.5)}} \cdot \frac{(k_1+1)\mathrm{f}_i}{K+\mathrm{f}_i} \cdot \frac{(k_2+1)q\mathrm{f}_i}{k_2+q\mathrm{f}_i} \quad (1)$$

$$K = k_1\left((1-b)+b.\frac{dl}{avdl}\right) \quad (2)$$

Where k_1, k_2, and K are parameters whose values set empirically. qf_i refers to how many times the same term repeated in the same query. R is the number of relevant documents in the query, and r_i is the number of relevant documents containing term i. The document length is dl. A typical value for k_1 is 1.2, and k_2 varies from 0 to 1,000 and b=0.75. $avdl$ is the average length of a document in the collection.

The application of QE using DANs in the BM25 model is a straightforward process, whereby the input is the entire query terms, and based on the WE model, vectors for each individual query terms will be extracted, and the average vector will be generated using DANs. The top k retrieved documents of PRF will be trained using CBOW model, where each term represented in the WE corpus as a vector of real number. Then, the most similar vectors in this corpus to the average vector of DANs are identified by using the cosine similarity measures, as illustrated in Equation 3. The most similar vectors are the potential candidate expansion vectors, and the candidate expansion terms related to the extracted vectors are identified using WE corpus. The top n candidates are used as expansion terms to the original query.

$$Cosine(A,B) = \frac{A\cdot B}{\|A\|\times\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\times\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

$$\frac{1}{|C|}\sum_{t=1}^{|C|}\log[\log[P(wt \setminus wt-c,\ldots,wt-1,wt+1,..,wt+c)]] \quad (4)$$

Equation 4 is the CBOW architecture of the Word2Vec model to learn the vector representation of the words, where it uses the surrounding words to predicts the target word, where $|C|$ is the number of words in the corpus, and c is the size of the dynamic context of wt.

$$avg(Q) = \frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3} \quad (5)$$

Equation 5 is used to find the average vector of the original query term vectors, where v_i is the vectors for each term i in query Q, $avg\ (Q)$ is the average vector of Q, and n the number terms in query Q.

The following steps illustrate the application of DANs-based PRF for QE on the BM25 model:

Step 1. Select a set of users' query Q consisting of n terms, whereby $Q = \{t_1, t_2, \ldots, t_n\}$;

Step 2. For each term t_i in Q, using the created WE model, we get the respective vector v_i from the WE corpus;

Step 3. For each vector v_i in query Q, we compute the average vector of this query Q using $s=avg(Q)=\frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3}$, where $avg\ (Q)$ is the average vector of Q, v_i is the vectors for each term i in query Q, and n the number terms in query Q;

Step 4. Select set $E = \{e_1, e_2, \ldots, e_n\}$, where E (consisting of terms e) is the top k document of PRF;

Step 5. For each term e in set E, using the created WE model, we get the respective vector ev_i from the WE corpus;

Step 6. For each vector ev_i, we compute the WE similarities to the $avg\ (Q)$ of DANs;

Step 7. Select the top k similar vectors to the $avg\ (Q)$ from set E (consists vectors ev_i) and save the selected vectors in set W, where set $W = \{w_1, w_2, \ldots w_k\}$ has the most similar vectors in set E to the $avg\ (Q)$ of DANs;

Step 8. Based on the created WE model, get the words $tw_1, tw_2, \ldots tw_k$ respective to the final candidate vectors $w_1, w_2, \ldots, w_k$ from the WE corpus;

Step 9. Add the new words $tw_1, tw_2, \ldots tw_k$ to the original query Q to get new query Q';

Step 10. Retrieve documents using the new query Q'.

The present study relied on the PRF method to identify potential expansion term vectors using top documents obtained during retrieval. QE implemented through DANs is easily integrated with the BM25 framework-

 https://doi.org/10.1633/JISTaP.2021.9.2.1

based PRF. The set of query terms is formed using the WE framework and forms the input. Vectors corresponding to every query element are computed; subsequently, DANs are employed to determine the average vector. Training is implemented using the CBOW scheme, where the top *k* documents comprising PRF are used. Real number vectors are used for denoting the elements of the WE corpus. Cosine similarity rules are employed to determine the vectors most similar to the DANs average vector; this process is specified using Equation 5. The vectors identified during the previous step are potential expansion element vectors whose corresponding word representations are potential query augmentation elements. The first *n* words are used for augmenting the query.

3.3. Embedding-Based Query Expansion Method using DANs (EQE1+DANs-PRF)

As indicated before, the EQE1 scheme suggested by Zamani and Croft (2016) suggested that query terms are presupposed to have conditional independence. This technique begins by identifying from the WE corpus a list of potential expansion candidates using similarity metrics. Subsequently, the terms identified in the previous step are compared with all query terms. Threshold candidates are those bearing high similarity to all query terms; these are chosen for QE. The EQE1 technique provides acceptable results when query terms have a substantial similarity. Nevertheless, most queries do not have a substantial similarity between all constituent words.

While implementing DANs using EQE1-based PRF, rather than identifying very similar candidates, we intended to identify the degree of similarity between the vectors corresponding to the top documents retrieved during PRF and the *average vector* corresponding to the initial query. Here, query-term vectors are identified using the WE corpus; these vectors are used to identify similar vectors from top PRF documents using the WE framework.

Once the vectors corresponding to the original query terms are obtained, cosine similarity is used to identify vectors from the top retrieved PRF documents resembling the initial query vectors. Furthermore, we used DANs on the initial query terms to determine the corresponding average vector. These two steps yielded the candidate and average vectors. Subsequently, the candidate and average DANs vectors were compared. Candidates having equal or greater than ≥ 0.7 similarity to average DANs vectors were selected.

Furthermore, the WE framework was used to reference the WE corpus to identify words using corresponding vectors. Lastly, new words are used for QE. Experiments indicated that 0.7 was an appropriate threshold for selecting similar words.

We proceeded under the assumption that the average query-term vector indicates a significant semantic similarity to the query. The top PRF documents retrieved during the process comprise valid words for QE. Furthermore, potential word vectors present in top PRF-retrieved documents having maximum similarity to the average vector are understood to be semantically analogous or associated with the original query's meaning because that vector denotes the query's centroid. The process of applying DANs in EQE1 can be summarized into the following steps:

Step 1. Select a set of users' query Q consisting of n terms, whereby $Q = \{t_1, t_2, ..., t_n\}$;
Step 2. For each term t_i in Q, using the created WE model, we get the respective vector v_i from the WE corpus;
Step 3. Compute the similarity (using cosine similarity) between v_i and all other vectors in the WE corpus;
Step 4. Select the top k similar vectors to v_i from the WE corpus and save the selected vectors in set E, where set $W = \{w_1, w_2, ..., w_k\}$ has the most similar vectors to v_i in the WE corpus;
Step 5. For each vector v_i in query Q, we compute the average vector of this query Q using $DANs = avg(Q) = \frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3}$, where $avg\ (Q)$ is the average vector of Q, v_i is the vectors for each term i in query Q, and n the number terms in query Q;
Step 6. Compute the similarity between $avg\ (Q)$ and set W;
Step 7. Select the vectors $\{vw_1, vw_2, ... vw_k\}$ from set W which have similarity ≥ *0.7* to $avg\ (Q)$;
Step 8. Based on the created WE model, get the words $tw_1, tw_2, ... tw_k$ respective to the final candidate vectors $vw_1, vw_2, ..., vw_k$ from the WE corpus;
Step 9. Add the new words $tw_1, tw_2, ... tw_k$ to the original query Q to get new query Q';
Step 10. Retrieve documents using the new query Q'.

3.4. Prospect-Guided Query Expansion Strategy Based on DANs (V2Q+DANs-PRF)

Fernández-Reyes et al. (2018) proposed V2Q and declared that not all query terms are necessarily practical for the expansion process; some terms might impact retrieval quality adversely. Hence, they suggested using a filtering stage to isolate the original terms before starting the ex-

pansion process. Potential expansion words are selected using filtered query words. Nevertheless, disregarding even a single word might lead to loss of query context and change in its meaning.

However, we have suggested using DANs to use the average vector corresponding to initial query terms rather than eliminating one or more query terms for candidate list creation. The vector is then used to prepare the potential candidate list. This technique allows keeping the query context unchanged. The candidate set will be created using the average DANs vector supposedly indicative of the original query context. This technique is understood as an augmented variant of V2Q.

Hence, we suggest that the average DANs vector be used instead of skimming the WE corpus for similar vectors. The top PRF documents will be searched for similar vectors. We propose to use V2Q+DANs-based PRF (V2Q+DANs-PRF). This technique comprises three primary steps: firstly, the initial candidate expansion set is created; this set is named W. It comprises vectors in the top PRF documents that closely resemble each query term vector. Consequently, D is the resulting candidate expansion vector list. The next step is to assess the W∪D set and calculate the cosine similarity of the vectors and compare it to the average vector *avg* (*Q*); all similarity values ≥ 0.7 are accepted, and the corresponding expansion vectors are selected. The terms corresponding to these vectors are used for QE. The following steps illustrate the process of applying DANs in the V2Q approach of QE.

Step 1. Select a set of user query Q consisting of n terms, whereby $Q = \{t_1, t_2, ..., t_n\}$;

Step 2. For each term t_i in Q, using the created WE model, we get the respective vector v_i from the WE corpus;

Step 3. Compute the similarity (using cosine similarity) between v_i and all other vectors in the WE corpus;

Step 4. Select the top k similar vectors to v_i from the WE corpus and save the selected vectors in set W, where set $W = \{w_1, w_2, ..., w_k\}$ has the most similar vectors in the WE corpus to v_i;

Step 5. For each vector v_i in query Q, we compute the average vector of this query Q using *DANs* = *avg* (*Q*) $= \frac{1}{|C|}\sum_{i=1}^{3}\frac{c_i}{3}$, where *avg* (*Q*) is the average vector of Q, v_i is the vectors for each term i in query Q, and n the number terms in query Q;

Step 6. Compute the similarity between *avg* (*Q*) and other vectors in the WE corpus;

Step 7. Select the top k similar vectors to *avg* (*Q*) from the WE corpus and save the selected vectors in set D, where set $D = \{d_1, d_2, ..., d_k\}$ has the most similar vectors in the WE corpus to *avg* (*Q*);

Step 8. Select the vectors from set W∪D which have similarity ≥ *0.7* to *avg* (*Q*) and save the selected vectors in set S, where set S $\{s_1, s_2, ..., s_n\}$ has the vectors from set W∪D that have similarity ≥ *0.7* to *avg* (*Q*);

Step 9. Based on the created WE model, get the words $ts_1, ts_2, ... ts_k$ respective to the final candidate vectors $s_1, s_2, ..., s_n$ from the WE corpus;

Step 10. Add the new words $ts_1, ts_2, ... ts_k$ to the original query Q to get new query Q';

Step 11. Retrieve documents using the new query Q'.

4. EXPERIMENTS

4.1. Experimental Setup

To assess the effects of the suggested sentence-embedding method DANs for PRF founded on QE, we evaluated three techniques, namely, basic BM25 framework, EQE1, and V2Q approaches. The previous section discusses the implementation of DANs as proposed. We contrast the outcomes of all DANs-based PRF techniques with the original techniques as formulated by researchers.

The TREC 2001/2002 Arabic newswire dataset was used for assessing the framework. Several researchers who worked on retrieval of Arabic text used the same dataset (Darwish & Ali, 2012; El Mahdaouy et al., 2019). This dataset comprises 2,117 Middle East-specific news documents dated between May 13, 1994, and December 20, 2000. The Word2Vec technique (Mikolov et al., 2013b) was used to build the WE corpus for experiments that used 383,872 Arabic news articles from Agence France Presse. The Linguistic Data Consortium disseminates these articles. The documents are encoded using UTF-8 and occupy about one gigabyte.

TREC 2001/2002 Arabic newswire contains three TREC sets: TREC 2001, TREC 2002, and TREC 2001/2002, comprising 25, 50, and 75 queries. TREC 2001/2002 comprises a combination of the queries of the constituent sets. Stemmer was used during experimentation because it has demonstrated a substantial impact on Arabic text retrieval performance. Farasa (Darwish & Mubarak, 2016) is considered the most efficacious stemmer for Arabic and was used in the present works of El Mahdaouy et al. (2019). Fig. 2 shows a comparison of precision and recall performance for 25 queries for TREC 2001, while Fig. 3 and Fig. 4 showing a comparison of

 https://doi.org/10.1633/JISTaP.2021.9.2.1

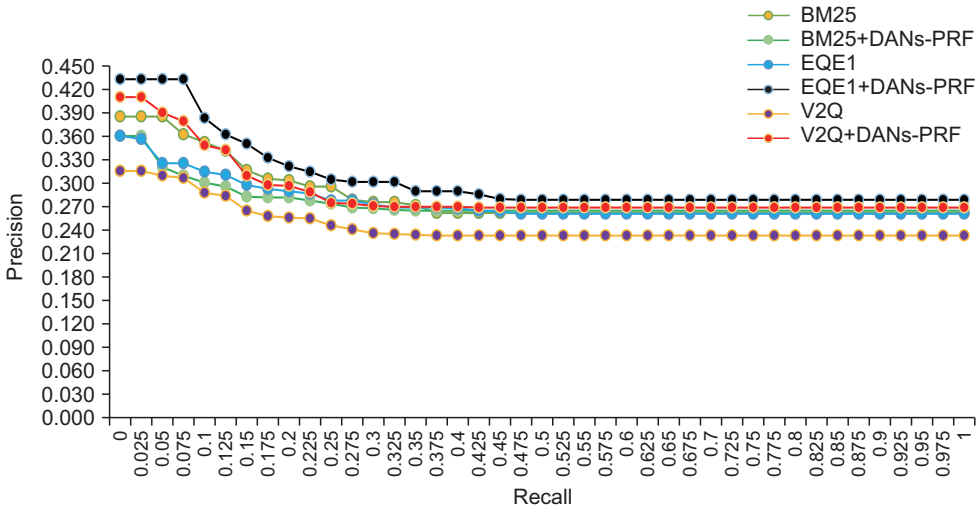


Fig. 2. Comparison of performance by precision and recall for 25 queries for TREC 2001 of all the models. DANs, deep averaging networks; PRF, pseudo relevance feedback.

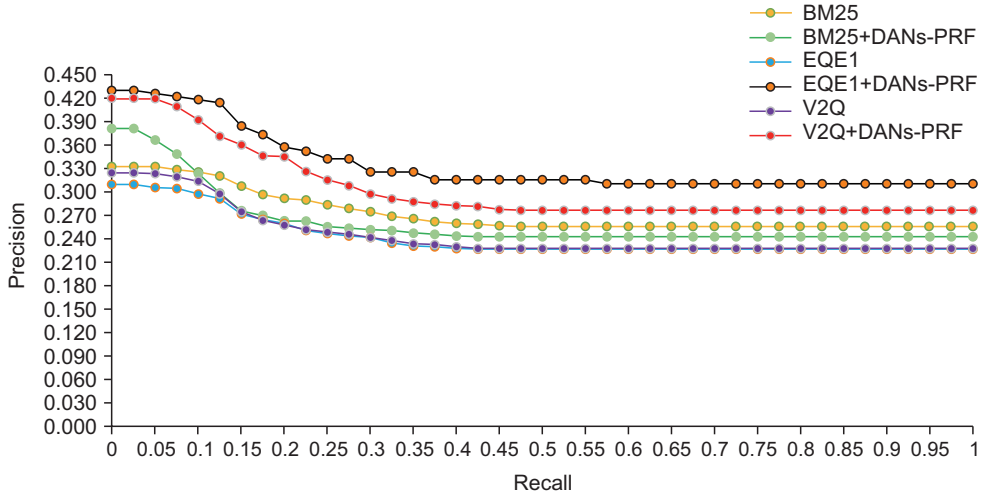


Fig. 3. Comparison of performance by precision and recall for 50 queries for TREC 2002 of all the models. DANs, deep averaging networks; PRF, pseudo relevance feedback.

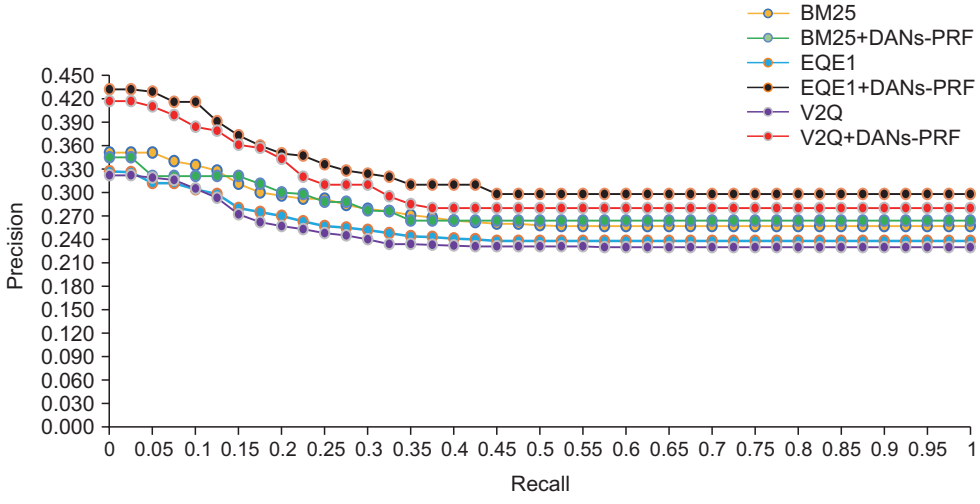


Fig. 4. Comparison of performance by precision and recall for 75 queries for TREC 2001/2002 of all the models. DANs, deep averaging networks; PRF, pseudo relevance feedback.

precision and recall performance for 50 queries for TREC 2002, and 75 queries for TREC 2001/2002 respectively, for all the baseline approaches and the proposed methods.

The system was set to retrieve 100 documents. The assessment comprised three baselines: (1) Okapi BM25 probabilistic framework used without expansion, (2) Embedding-based Query Expansion (EQE1), and (3) Prospect-Guided QE strategy (V2Q). Whoosh is a Python-based search engine library (Mukherjee & Kumar, 2019), which was used for all experiments. WE vector cosine similarity measure was used to calculate the semantic similarity between the terms.

4.2. Evaluation Metrics

The DANs QE-based PRF is benchmarked against six baseline techniques using MAP, corresponding to the top 100 documents for determining retrieval efficacy. Furthermore, precision for the top ten documents (P@10) was also used for performance measurement. The expressions for the performance metrics are specified below. Equations 6 and 7, respectively, express the recall and precision metrics, while Equation 8 is for the MAP, and the AP metric is presented in Equation 9.

$$Recall = \frac{|Rel| \cap |Ret|}{|Rel|} \tag{6}$$

$$Precision = \frac{|Rel| \cap |Ret|}{|Rel|} \tag{7}$$

Where:

Ret is the total number of retrieved documents, and *Rel* is the total number of relevant documents in the dataset.

$$MAP = \frac{1}{N}\sum_{n} AP_n \tag{8}$$

Where:

AP is the average precision value for a given query from the evaluation set of *n* queries, defined as:

$$AP = \frac{\sum_n P@r}{R} \tag{9}$$

Where:

r being the rank of each relevant document, *R* is the total number of the relevant documents, and *P@r* is the precision of the top-r retrieved documents.

5. RESULTS

The results of the DANs-based PRF used in this study are presented in this section. Conventional DANs-based QE is applied to the BM25 framework (BM25+DANs). DANs-based EQE1 and V2Q approaches were assessed for identification and selection performance pertaining to potential expansion terms. This study did not use the entire WE corpus, unlike the traditional AQE techniques. Instead, we used the top PRF documents for QE. These augmented techniques were designated BM25+DANs-PRF, EQE1+DANs-PRF, and V2Q+DANs-PRF, respectively. It is hypothesised that DANs can provide better expansion terms using top-retrieved PRF documents instead of the entire WE corpus. Consequently, AIR system performance can be enhanced.

Table 2 presents experimental results of all schemes, considering MAP and top-10 precision level (P@10). Ta-

Table 2. MAP and P@10 for all the models

Techniques	Collections					
	TREC 2001		TREC 2002		TREC 2001/2002	
	MAP	P@10	MAP	P@10	MAP	P@10
BM25	31.30	42.10	28.70	35.80	29.50	37.90
BM25+DANs-PRF	25.60	22.90	29.30[a)]	37.60[a)]	31.10[a)]	34.50
EQE1	30.70	42.90	25.70	33.50	26.90	30.40
EQE1+DANs-PRF	31.30[a)]	43.00[a)]	30.20[a)]	34.30[a)]	32.30[a)]	34.20[a)]
V2Q	27.00	35.20	26.20	33.30	26.50	33.90
V2Q+DANs-PRF	28.30[a)]	35.70[a)]	31.20[a)]	34.60[a)]	30.70[a)]	34.10[a)]

MAP, mean average precision; DANs, deep averaging networks; PRF, pseudo relevance feedback.
[a)]These values indicate that the corresponding scheme produced superior results than the baseline DANs-based PRF.

 https://doi.org/10.1633/JISTaP.2021.9.2.1

Table 3. MAP values for topic-by-topic analysis for all the models

Topics	BM25	BM25+DANs-PRF	EQE1	EQE1+DANs-PRF	V2Q	V2Q+DANs-PRF
1. Politics	0.169	0.295[a)]	0.217	0.216	0.244	0.261[a)]
2. Arts	0.235	0.330[a)]	0.277	0.303[a)]	0.273	0.283[a)]
3. Technology	0.093	0.148[a)]	0.098	0.118[a)]	0.127	0.185[a)]
4. Diseases	0.190	0.371[a)]	0.327	0.337[a)]	0.346	0.358[a)]
5. Waters	0.095	0.144[a)]	0.125	0.155[a)]	0.117	0.119[a)]
6. Environment	0.106	0.168[a)]	0.130	0.204[a)]	0.119	0.174[a)]
7. Tourism	0.101	0.161	0.078	0.082[a)]	0.112	0.220[a)]

MAP, mean average precision; DANs, deep averaging networks; PRF, pseudo relevance feedback.
[a)]These values indicate that the corresponding scheme produced superior results than the baseline DANs-based PRF.

ble 3 presents the MAP values for topic-by-topic analysis of all the proposed techniques in addition to the baseline approaches, respectively.

6. TOPIC-BY-TOPIC ANALYSIS

As mentioned earlier, the dataset used in this study is Arabic newswire TREC 2001/2002, which has two collections, namely TREC 2001 and TREC 2002. The first collection consists of 25 queries and the second 50 queries, and the total is 75 queries from various topics. However, to carry out a topic-by-topic analysis, these queries have been grouped into seven domains: politics (34 queries), arts (11 queries), technology (11 queries), diseases (seven queries), waters (six queries), environment (four queries), and tourism (two queries).

7. DISCUSSION

Empirical data presented in Table 2 suggests that the augmented BM25+DANs-PRF technique is superior to the BM25 probabilistic framework for TREC 2002 and TREC 2001/2002 datasets when evaluated using MAP. Considering P@10, the technique proposed in this present article yielded better results than the traditional BM25 executed only for the TREC 2002 dataset. For TREC 2001, the proposed BM25+DANs-PRF technique failed to pass the baseline model BM25 in terms of MAP and P@10.

Moreover, the proposed EQE1+DANs-PRF approach was superior to the traditional EQE1 technique when measured using MAP and P@10 for all TREC datasets. At the same time, the proposed V2Q+DANs-PRF approach used in the present study is superior to the traditional V2Q approach when evaluated using P@10 and MAP for all TREC datasets.

Furthermore, based on the results of topic-by-topic which were presented in Table 3, we can note that, in terms of MAP, the proposed technique BM25+DANs-PRF has outperformed their baseline BM25 (without expansion) for all the topics, where it significantly outperformed their baseline model for the topics "Politics" and "Diseases." The proposed EQE1+DANs-PRF technique has outperformed its baseline EQE1 approach in terms of MAP for all the topics, except for the "Politics" topic, where it has almost similar values. In addition, the proposed technique V2Q+DANs-PRF has outperformed its baseline V2Q approach for all the topics in terms of MAP.

8. CONCLUSION AND FUTURE WORKS

This research has evaluated the effects of DANs sentence-embedding methods on PRF-based AQE used for AIR. The intent is to identify the most appropriate terms from the top PRF documents. The vectors representing the terms are compared to the average vector of the initial query. The most-similar vectors are used to identify terms for AQE.

The present study is based on the Word2Vec WE model. Three baselines, namely, Okapi BM25, EQE1, and V2Q were used for performance benchmarking. The present study contributes by using a different approach: Rather than identifying the similarity between average DANs vectors and WE corpus, we attempted to determine the degree of similarity between the average DANs vector and top PRF documents.

The candidate vectors were generated based on the average vector of DANs in the enhanced BM25+DANs-PRF technique from the top retrieved documents of PRF. Meanwhile, in the other enhanced techniques EQE1+DANs-PRF and V2Q+DANs-PRF, the candidate

vectors are generated based on the original query term vectors in addition to the average vector of DANs, with the help of the top-retrieved documents of PRF.

The augmented BM25+DANs-PRF technique relies on the top-retrieved PRF documents and identifies potential vectors using DANs average vector. At the same time, other augmented methods like V2Q+DANs-PRF and EQE1+DANs-PRF rely on the vectors of the initial query terms and DANs average vectors to identify terms using top PRF documents.

Overall, the experiment results illustrated in Table 2 showed that the proposed DANs-based PRF, when incorporated into the probabilistic model BM25, are unable to pass their baseline standard BM25 (without expansion) in most cases in terms of MAP and P@10. On the other hand, the proposed DANs-based PRF when incorporated into the EQE1 and V2Q approaches has proven its success, where it outperformed the standard baseline approaches EQE1 and V2Q; however, in most of the cases, there is no significant difference between the results of the baselines approaches and the proposed techniques.

In addition, based on the topic-by-topic analysis, we can notice that the proposed techniques BM25+DANs-PRF, V2Q+DANs-PRF, and EQE1+DANs-PRF have significantly outperformed their baselines in most cases in terms of MAP. Finally, we can conclude that the proposed DANs technique is performing better with the specific topics than with the global topics.

This study's objective was to address the vocabulary mismatch challenge using the top-retrieved PRF documents that have strong similarity to the average vector of the query terms. We hypothesise that the top-retrieved PRF documents might have the required terms that demonstrate a high degree of similarity to the average vector of the initial query terms. Consequently, Arabic text retrieval performance can be enhanced.

ACKNOWLEDGMENTS

We want to thank the LDC for providing us with the LDC2001T55 Arabic Newswire Part 1 at no cost, and for awarding us with the Fall 2012 LDC Data Scholarship. This study is partially supported by the Universiti Kebangsaan Malaysia, grant: DCP-2017-007/4.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Abbache, A., Meziane, F., Belalem, G., & Belkredim, F. Z. (2016). Arabic query expansion using WordNet and association rules. *International Journal of Intelligent Information Technologies*, 12(3), 51-64. http://doi.org/10.4018/IJIIT.2016070104.

Abu El-Khair, I. (2007). Arabic information retrieval. *Annual Review of Information Science and Technology*, 41(1), 505-533. https://doi.org/10.1002/aris.2007.1440410118.

Aklouche, B., Bounhas, I., & Slimani, Y. (2018, November 14-16). *Query expansion based on NLP and word embeddings*. Paper presented at the TREC 2018, Gaithersburg, MD, USA.

ALMasri, M., Berrut, C., & Chevallet, J.-P. (2016, March 20-23). A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, & G. Silvello (Eds.), *Proceedings of the 38th European Conference on IR Research* (pp. 709-715). Springer. https://doi.org/10.1007/978-3-319-30671-1_57.

Alsmearat, K., Al-Ayyoub, M., & Al-Shalabi, R. (2014, November 10-13). An extensive study of the Bag-of-Words approach for gender identification of Arabic articles. In A. Bouras, Z. Tari, A. Erradi, & S. Abdelwahed (Eds.), *Proceedings of the 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications* (pp. 601-608). IEEE. https://doi.org/10.1109/AICCSA.2014.7073254.

Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4), 357-389. https://doi.org/10.1145/582415.582416.

Atwan, J., Mohd, M., Rashaideh, H., & Kanaan, G. (2016). Semantically enhanced pseudo relevance feedback for Arabic information retrieval. *Journal of Information Science*, 42(2), 246-260. https://doi.org/10.1177%2F0165551515594722.

Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, 56(5), 1698-1735. https://doi.org/10.1016/j.ipm.2019.05.009.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). Ask for information retrieval: Part II. Results of a design study. *Journal of Documentation*, 38(3), 145-164. https://doi.org/10.1108/eb026726.

Ben Guirat, S., Bounhas, I., & Slimani, Y. (2016). Combining indexing units for Arabic information retrieval. *International Journal of Software Innovation*, 4(4), 1-14. https://

doi.org/10.4018/IJSI.2016100101.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1-127. https://doi.org/10.1561/2200000006.

Berget, G., & Sandnes, F. E. (2015). Searching databases without query-building aids: Implications for dyslexic users. *Information Research: An International Electronic Journal*, 20(4), 689.

Carpineto, C., De Mori, R., Romano, G., & Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1), 1-27. https://doi.org/10.1145/366836.366860.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1), 1. https://doi.org/10.1145/2071389.2071390.

Clinchant, S., & Gaussier, E. (2013, September 29-October 2). A theoretical analysis of pseudo-relevance feedback models. In O. Kurland, D. Metzler, C. Lioma, B. Larsen, & P. Ingwersen (Eds.), *Proceedings of the ICTIR '13: International Conference on the Theory of Information Retrieval* (pp. 6-13). Association for Computing Machinery. https://doi.org/10.1145/2499178.2499179.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76), 2493-2537.

Crimp, R., & Trotman, A. (2018, December 11-12). Refining query expansion terms using query context. In B. Koopman, A. Trotman, & P. Thomas (Eds.), *Proceedings of the ADCS '18: 23rd Australasian Document Computing Symposium* (article no.: 12). Association for Computing Machinery. https://doi.org/10.1145/3291992.3292000.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Addison-Wesley.

Dalton, J., Naseri, S., Dietz, L., & Allan, J. (2019, April 14-18). Local and global query expansion for hierarchical complex topics. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, & D. Hiemstra (Eds.), *Proceedings of the 41st European Conference on IR Research*, *ECIR 2019* (pp. 290-303). Springer. https://doi.org/10.1007/978-3-030-15712-8_19.

Darwish, K., & Ali, A. (2012, July 8-14). Arabic retrieval revisited: Morphological hole filling. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 218-222). ACL.

Darwish, K., & Mubarak, H. (2016, May 23-28). Farasa: A new fast and accurate Arabic word segmenter. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.) *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1070-1074). European Language Resources Association.

Diaz, F., Mitra, B., & Craswell, N. (2016, August 7-12). Query expansion with locally-trained word embeddings. In K. Erk, & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 367-377). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1035.

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2019). Word-embedding-based pseudo-relevance feedback for Arabic information retrieval. *Journal of Information Science*, 45(4), 429-442. https://doi.org/10.1177%2F0165551518792210.

Esposito, M., Damiano, E., Minutolo, A., De Pietro, G., & Fujita, H. (2020). Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Information Sciences*, 514, 88-105. https://doi.org/10.1016/j.ins.2019.12.002.

Fang, H., & Zhai, C. (2006, August 6-11). Semantic term matching in axiomatic approaches to information retrieval. In S. Dumais, E. N. Efthimiadis, D. Hawking, & K. Järvelin (Eds.), *Proceedings of the SIGIR '06: 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115-122). Association for Computing Machinery. https://doi.org/10.1145/1148170.1148193.

Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., & Quwaider, M. (2014, August 27-29). Cross-lingual short-text document classification for Facebook comments. In M. Younas, I. Awan, & A. Pescape (Eds.), *Proceedings of the FiCloud 2014: 2nd International Conference on Future Internet of Things and Cloud* (pp. 573-578). IEEE. https://doi.org/10.1109/FiCloud.2014.99.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8(4), 14. https://doi.org/10.1145/1644879.1644881.

Farhan, Y. H., Noah, S. A. M., & Mohd, M. (2020). Survey of automatic query expansion for arabic text retrieval. *Journal of Information Science Theory and Practice*, 8(4), 67-86. https://doi.org/10.1633/JISTaP.2020.8.4.6.

Fernández-Reyes, F. C., Hermosillo-Valadez, J., & Montes-y-Gómez, M. (2018). A prospect-guided global query expansion strategy using word embeddings. *Information Processing & Management*, 54(1), 1-13. https://doi.org/10.1016/j.ipm.2017.09.001.

Franco-Salvador, M., Rangel, F., Rosso, P., Taulé, M., & Martít, M. A. (2015, September 8-11). Language variety identification using distributed representations of words and docu-

ments. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, & N. Ferro (Eds.), *Proceedings of the 6th International Conference of the CLEF Association*, *CLEF'15* (pp. 28-40). Springer. https://doi.org/10.1007/978-3-319-24027-5_3.

Ganguly, D., Roy, D., Mitra, M., & Jones, G. J. F. (2015, August 9-13). Word embedding based generalized language model for information retrieval. In R. González-Ibáñez, & N. Hidalgo (Eds.), *Proceedings of the SIGIR '15: 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 795-798). Association for Computing Machinery. https://doi.org/10.1145/2766462.2767780.

Iyyer, M., Manjunatha, V., & Daumé, H., III. (2015, July 26-31). Deep unordered composition rivals syntactic methods for text classification. In C. Zong, & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 1681-1691). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1162.

Kim, H. K., Kim, H., & Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266, 336-352. https://doi.org/10.1016/j.neucom.2017.05.046.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002, August 11-15). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In K. Järvelin, R. Baeza-Yates, & S. H. Myaeng (Eds.), *Proceedings of the SIGIR '02: 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275-282). Association for Computing Machinery. https://doi.org/10.1145/564376.564425.

Lavrenko, V., & Croft, W. B. (2017). Relevance-based language models. *ACM SIGIR Forum*, 51(2), 260-267. https://doi.org/10.1145/3130348.3130376.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). *Efficient estimation of word representations in vector space*. https://arxiv.org/abs/1301.3781v3.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b, December 5-10). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the NIPS'13: 26th International Conference on Neural Information Processing Systems* (pp. 3111-3119). Curran Associates.

Miyanishi, T., Seki, K., & Uehara, K. (2013, October 27-November 1). Improving pseudo-relevance feedback via tweet selection. In Q. He, A. Iyengar, W. Nejdl, J. Pei, & R. Rastogi (Eds.), *Proceedings of the CIKM '13: 22nd ACM international conference on Information & Knowledge Management* (pp. 439-448). Association for Computing Machinery. https://doi.org/10.1145/2505515.2505701.

Mohsen, G., Al-Ayyoub, M., Hmeidi, I., & Al-Aiad, A. (2018, April 3-5). On the automatic construction of an Arabic thesaurus. In M. Quwaider (Ed.), *Proceedings of the 2018 9th International Conference on Information and Communication Systems* (pp. 243-247). IEEE. https://doi.org/10.1109/IACS.2018.8355431.

Montazeralghaem, A., Zamani, H., & Shakery, A. (2016, July 17-21). Axiomatic analysis for improving the log-logistic feedback model. In R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, & J. Zobel (Eds.), *Proceedings of the SIGIR '16: 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 765-768). Association for Computing Machinery. https://doi.org/10.1145/2911451.2914768.

Mukherjee, S., & Kumar, N. S. (2019, December 9-11). Duplicate question management and answer verification system. In M. Chang, R. Rajendran, Kinshuk, S. Murthy, & V. Kamat (Eds.), *Proceedings of the 2019 IEEE Tenth International Conference on Technology for Education* (pp. 266-267). IEEE. https://doi.org/10.1109/T4E.2019.00067.

Mustafa, M., AbdAlla, H., & Suleman, H. (2008, December 2-5). Current approaches in Arabic IR: A survey. In G. Buchanan, M. Masoodian, & S. J. Cunningham (Eds.), *Proceedings of the 11th International Conference on Asian Digital Libraries*, *ICADL 2008* (pp. 406-407). Springer. https://doi.org/10.1007/978-3-540-89533-6_57.

Pal, D., Mitra, M., & Datta, K. (2014). Improving query expansion using WordNet. *Journal of the Association for Information Science and Technology*, 65(12), 2469-2478. https://doi.org/10.1002/asi.23143.

Pennington, J., Socher, R., & Manning, C. (2014, October 25-29). GloVe: global vectors for word representation. In Y. Marton (Ed.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). *Okapi at TREC-3*. Paper presented at the 3rd Text REtrieval Conference (TREC-3), Gaithersburg, MD, USA.

Roy, D., Paul, D., Mitra M., & Garain, U. (2016). *Using word embeddings for automatic query expansion*. Paper presented at the Neu-IR '16 SIGIR Workshop on Neural Informa-

tion Retrieval, Pisa, Italy.

Takeuchi, S., Sugiura, K., Akahoshi, Y., & Zettsu, K. (2017). Spatio-temporal pseudo relevance feedback for scientific data retrieval. *IEEJ Transactions on Electrical and Electronic Engineering*, 12(1), 124-131. https://doi.org/10.1002/tee.22352.

Trotman, A., Puurula, A., & Burgess, B. (2014, November 27-28). Improvements to BM25 and language models examined. In J. Culpepper, L. Park, & G. Zuccon (Eds.), *Proceedings of the ADCS '14: 2014 Australasian Document Computing Symposium* (pp. 58-65). Association for Computing Machinery. https://doi.org/10.1145/2682862.2682863.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141-188.

Vaidyanathan, R., Das, S., & Srivastava, N. (2015). A study on retrieval models and query expansion using PRF. *International Journal of Scientific & Engineering Research*, 6(2), 13-18.

Xue, B., Fu, C., & Shaobin, Z. (2014, June 27-July 2). A study on sentiment computing and classification of Sina Weibo with Word2vec. In P. Chen, & H. Jain (Eds.), *Proceedings of the 2014 IEEE International Congress on Big Data* (pp. 358-363). IEEE. https://doi.org/10.1109/BigData.Congress.2014.59.

Zamani, H., & Croft, W. B. (2016, September 12-16). Embedding-based query language models. In B. Carterette, & H. Fang (Eds.), *Proceedings of the ICTIR '16: 2016 ACM International Conference on the Theory of Information Retrieval* (pp. 147-156). Association for Computing Machinery. https://doi.org/10.1145/2970398.2970405.

Zuccon, G., Koopman, B., Bruza, P., & Azzopardi, L. (2015, December 8-9). Integrating and evaluating neural word embeddings in information retrieval. In L. A. F. Park, & S. Karimi (Eds.), *Proceedings of the ADCS '15: 20th Australasian Document Computing Symposium* (article no.: 12). Association for Computing Machinery. https://doi.org/10.1145/2838931.2838936.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(2): 18-32, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.2.2

Social Media News in Crisis? Popularity Analysis of the Top Nine Facebook Pages of Bangladeshi News Media

Md. Sayeed Al-Zaman*
Department of Journalism and Media Studies, Jahangirnagar University, Savar, Dhaka, Bangladesh
E-mail: msalzaman@juniv.edu

Mridha Md. Shiblee Noman
Department of Journalism and Media Studies, Jahangirnagar University, Savar, Dhaka, Bangladesh
E-mail: shibleenoman@juniv.edu

ABSTRACT

Social media has become a popular source of information around the world. Previous studies explored different trends of social media news consumption. However, no studies have focused on Bangladesh to date, where social media penetration is very high in recent years. To fill this gap, this research aimed to understand its popularity trends during the period. For that reason, this work analyzes 97.67 million page likes and 3.48 billion interaction data collected from nine Bangladeshi news media's Facebook pages between December 2016 to November 2020. The analysis shows that the growth rates of page likes and interaction rates declined during this period. It suggests that the media's Facebook pages are gradually losing their popularity among Facebook users, which may have two more interpretations: Facebook's aggregate appeal as a news source is decreasing to users, or Bangladeshi media's appeal is eroding to Facebook users. These findings challenge the previous results, i.e., Facebook's demand as a news source is increasing with time. We offer four explanations of the decreased popularity of Facebook's news: information overload, exposure to incidental news, users' selective exposure and different aims of using Facebook, and conflict between media agendas and users' interests. Some theoretical and practical significance of the results has been discussed as well.

Keywords: media organization, Facebook page, social media news, trend analysis, information, Bangladesh

Received: May 21, 2021 **Revised:** May 30, 2021
Accepted: May 31, 2021 **Published:** June 30, 2021

***Corresponding Author:** Md. Sayeed Al-Zaman
https://orcid.org/0000-0003-1433-7387
E-mail: msalzaman@juniv.edu

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© Md. Sayeed Al-Zaman, Mridha Md. Shiblee Noman, 2021

1. INTRODUCTION

This study analyzed the popularity of news media outlets' Facebook pages based on the historical trends of page likes and public engagement. The Internet has changed the existing ecology of news sharing and consumption: Social media platforms, such as Facebook and Twitter, are advancing it to a greater degree (Pentina & Tarafdar, 2014). This global shift in people's information-seeking behavior has also brought a new challenge for the traditional media organizations (Bergström & Belfrage, 2018), because the increasing online news consumption is making them more dependent on social media (Hong, 2012; Lysak et al., 2012; Paulussen & Harder, 2014; Stassen, 2010). However, instead of undermining traditional media as some researchers believe, social media is reinforcing its capacity, extending its reach, and enabling multiplatform interaction (Chadwick, 2011; Hermida et al., 2012; Tewksbury & Rittenberg, 2012). In other words, both types of media are now converging and becoming interdependent (Dominick, 2009). In Bangladesh like elsewhere, mainstream media outlets are also relying on social media to produce and share their news items (Fahad, 2014). Thus, the increasing social media use is promoting the country's news industry in some ways (Alam, 2020; Saiful & Park, 2019).

The social media penetration rate in Bangladesh has been high in recent years, and most of the social media users use Facebook (StatCounter, 2020). Facebook is a good source of news for many and its preference is increasing incessantly (Bene, 2017; Lampe et al., 2012; Shearer & Gottfried, 2017; Silver & Matthews, 2017). For that reason, most of the national media outlets in Bangladesh maintain their official Facebook page to promote news items and attract more audiences; it is an effective marketing strategy of the media outlets as well (Pletikosa Cvijikj & Michahelles, 2013). A study shows that a Facebook page of a media outlet can generate 33.9% to 58% of that media's total engagement (Welbers & Opgenhaffen, 2018). Perceiving the benefits, media organizations usually share regular news updates, news links, breaking and attractive news stories, polls, and other interactive posts on their Facebook pages. In most cases, they share news stories that have social significance and public interest and stories that match the organization's ideology and agenda (Al-Rawi, 2017). In this way, they often try to socialize with their followers as well.

Due to the lack of proper academic endeavor, however, it is still unknown what the popularity patterns of media Facebook pages are, and how their interaction trends change over time in Bangladesh. In practice, the news consumption pattern of Facebook users is still an emerging area of research worldwide; therefore it remains vastly unexplored (Pentina & Tarafdar, 2014). Also, previous studies only focused on Facebook's popularity as an information source (Bene, 2017; Shearer & Gottfried, 2017; Silver & Matthews, 2017), the shift in people's information-seeking behavior from offline to online (Hong, 2012; Lysak et al., 2012; Paulussen & Harder, 2014; Stassen, 2010), and changes in social media users' news consumption and sharing behavior (Pentina & Tarafdar, 2014). These studies barely investigated and acknowledged the contributions of news media's Facebook pages, and their public engagement and popularity trends. Moreover, the context of Bangladesh is largely ignored in the previous literature. Therefore, this research attempted to bridge the existing gap a little, contributing some insights to media and communication scholarship. Based on a research question, the present study seeks to understand the popularity growths of the nine selected news pages and the engagement growths of their posts. In the next section, some relevant concepts are discussed.

2. CONCEPTUAL FRAMEWORK

Information behavior is a relatively old field of research, which is now becoming more interested in exploring the information-seeking tendencies of online communities. In such explorations, researchers use a wide range of concepts; some of them are old, derived from the previous literature, while some are novel and yet to have a wide acceptance. We have discussed a few concepts below that are relevant to the present study.

2.1. Information Behavior Models

In information behavior research, some dominant models and concepts discussed why and how people seek, rely on, consume, and react to different information and information sources in different circumstances. The Information Seeking Behavior model proposed by James Kirkelas in 1983 was probably the first model of its kind that dealt with people's information behavior (Rather & Ganaie, 2018). According to this model, people's information behavior follows four steps: perceiving a need for information, searching for the available information to fulfill the need, finding the information, and using the information (Rather & Ganaie, 2018, p. 1415). Information needs can be divided into five categories: to meet the need for new

information, to elucidate the information held, to confirm the information held, to elucidate beliefs and values held, and to confirm beliefs and values held (Wilson, 1997). Wilson in his information behavior model indicated three types of barriers in the information-seeking (searching and finding) process: personal barriers (emotional, educational, and demographic), social barriers (interpersonal and group relations), and environmental barriers (economic, source characteristics). These barriers may impact information utilization as well.

Interestingly, utilization of new information can satisfy or dissatisfy an information seeker, which may lead to further consequences. If a source can satisfy an information seeker with proper information, one assumption can be that the information seeker would prefer that source in future information-seeking; otherwise, not. In other words, if the source's preference of information matches a seeker's preference, their relationship should continue. It is important to note that every information source, such as newspapers and television, involves some specific preferences and intentions to achieve some goals (e.g., shaping opinion): We may call it an *agenda*. A source's (media) primary agenda is to influence the information seeker (audience) to cause some changes (McCombs, 2013). On the other hand, an information seeker may also have some agendas. When agendas from both parties have more commonalities, as we have already suggested, an information seeker starts depending on that source for information. However, when both parties' agendas contradict, an information seeker may move to some other available information sources. Studies found that information seekers' and information sources' agendas often differ (Schmidt et al., 2017). This can explain to some extent why some people prefer some certain information sources over others and why this behavior changes over time.

2.2. Information Overload & Selective Exposure

Information overload has been an important topic of interest to communication and information science scholars since the mid-twentieth century. However, the outset of social media and user-generated content has recently urged more research in this area, focusing mainly on online communication. Its main reason is the profuse production of digital information around the world every consecutive day. Information overload discusses a unique phenomenon: the dilemma of more information with limited information-processing capacities (Schick et al., 1990). Put simply, information overload is the situation in which an individual finds himself/herself with abundant information with a comparatively limited capacity to process and interact with this large amount of information. If we consider the case of social media, on one hand, every social media user is empowered to be information producers, who produce a large amount of information. On the other hand, users are the consumers of their self-produced information, making their cognitive information processing capacity exposed to such a large amount of information (Rodriguez et al., 2014). It creates an information overload situation for them.

In most cases, information overload leads to negative consequences, such as poor decision-making, exhaustion, change in information-seeking behavior, stress and tension, and so on (Fu et al., 2020; Schick et al., 1990). For example, Fu et al. (2020) in their study exploring three overloads (i.e., information overload, system failure overload, and social overload) argued that social media users become overloaded with information when automated information processing systems fail. Such information can be messages from friends, advertisements, or other third parties. When this information exceeds the information processing ability of the users, their feeling of social media exhaustion increases. Schick et al. (1990) argued that information overload in the organizational setup adversely impacts decision-making, management, and finance.

Rodriguez et al. (2014) modeled social media data to understand how and to what extent information overload impacts social media users. They found that the most active users suffer the most from information overload. In practice, social media users' ability to process information is limited and relies on the rate at which they receive information (Rodriguez et al., 2014, p. 179). They also found that "an exposure to a piece of information, be it an idea, a convention or a product is much less effective for users that receive information at higher rates, meaning they need more exposures to adopt a particular contagion" (Rodriguez et al., 2014, p. 170). There is a relationship between the overloaded users and their source preferences. Users' limited attention is divided among information, which is subject to change. When a user is overloaded, (s)he tends to prioritize a smaller number of information sources, mainly based on their reliability, ideological proximity, imitation, and so on (Rodriguez et al., 2014). Therefore, in terms of news sources, it is normal for users to prioritize a few sources over others, and their tastes may change over time based on several factors.

Other studies investigated the relationship between different demographic variables with news overload, and its impact on them. For example, Holton and Chyi (2012)

 https://doi.org/10.1633/JISTaP.2021.9.2.2

noted that gender is positively correlated with news overload and females are more vulnerable to such overload. They also found that news interest is negatively associated with overload, meaning news overload is less prevalent when users choose news according to their personal preferences (Holton & Chyi, 2012). Put another way, when users face the situation of information overload, they tend to prioritize their personal choice of news and source selection. A study with similar findings explored the phenomena that when users perceive themselves as overloaded with information, they selectively expose themselves to limited and certain information sources based on their trust and the sources' reliability (Lee et al., 2017). In contrast, users with a lower level of information overload frequently access different sources of information. The overloaded users also temporarily avoid some news sources to resolve their problem with processing large amounts of information. This could be a reason for some news sources' fluctuating popularity and interaction trends over time.

Studies also found that users' preferences of news sources are intentional and selective, producing a community structure among a few media outlets (Schmidt et al., 2017). They often purposefully filter the news sources and items that do not match their interests (Lee et al., 2017). Also, they do not interact with many sources for a long time; rather, they change their preferences often: They only interact for a long time with very few sources dominated by their selective exposure. This further explains the fluctuations of information sources' popularity.

2.3. Incidental News

Consumption of incidental news is commonplace in social media and for social media users. Algorithms of platforms, advertisements, and promotions could be its main reasons. However, incidental news is a pre-digital phenomenon that gains new momentum in contemporary times. It simply denotes the idea of an individual's unintentional exposure to information or news. Around 74% of users aged under 29 years encounter incidental news when they surf social media, a survey found; another survey found 29% of users encounter incidental news while using social media for some other reason (Boczkowski et al., 2018). These findings indicate that for social media users unwarranted and incidental news is not uncommon even when it is unwanted. However, in many cases, users become interested in the news and pay a visit to it, which increases the reach and interaction of that news item (Boczkowski et al., 2018). On the contrary, it may also prevent many users from visiting the sources of unwanted news. For example, in a study conducted by Boczkowski et al. (2018) in Argentina, an interviewee stated that he does not consume incidental news because it will be popping up often afterward, perhaps due to algorithms. These two contrasting actions and perspectives may explain some information sources' interaction fluctuations from time to time. In relation to these findings, another study on incidental news claimed that social media users constantly get incidental news if they spend significant time on the Internet (Boczkowski et al., 2017). Sometimes it becomes indistinguishable from other usual contents, so many users access that information sporadically and engage with it superficially.

3. MATERIALS & METHODS

3.1. Research Question

The previous research is limited in at least three ways. First, it did not focus on the context of Bangladesh or consider its unique social media environment and its users' information-seeking behavior. Second, historical trends in information seeking from social media pages did not receive proper attention. Third, multivariable popularity analysis of Facebook information sources (e.g., Facebook pages) is absent. Considering these limitations, this study endeavored to understand the popularity of social media news in Bangladesh based on historical trends of the selected media organizations' official Facebook pages. The following research question guided the data collection and analysis of this study.

RQ: What are the popularity trends of news media's Facebook pages?

Since there was hardly any relevant study in this specific area to guide our analysis and/or study framework, we had to determine our own way of dealing with the question. Regarding the research question, we measured the popularity of a Facebook page based on two variables: the page's number of followers, and engagement of the posts. Put simply, the first variable indicates how popular a page is, and the second variable indicates how popular the page's posts are.

3.2. Data Source

We selected Facebook for our analysis because of its unparalleled popularity (95%) in Bangladesh compared to other social media platforms, such as Twitter (0.35%) and YouTube (2.73%) (StatCounter, 2020). We selected the

top nine Facebook pages based on their numbers of followers on 17 December 2020, based on the insights from Socialbakers (http://socialbakers.com), an international social media marketing company. Our selected pages were Prothom Alo, BBC News Bangla, NTV, RTV, Somoyer Kanthosor, Daily Naya Diganta, BDnews24.com, Independent24.tv, and Kaler Kantho (Table 1). Of the selected media outlets, three were newspapers, three were television channels, and three were online news portals. We set the data collection period from 1 December 2016 to 30 November 2020, a total of four years with one-month intervals.

3.3. Data Type & Measurement

Two types of data were collected based on the research question: page popularity data and post engagement data. The first type of data included total page likes during the period, page likes after a regular interval, and page-like growth rates over the period. The second type of data included total interactions of the pages during the period, interactions after a regular interval, and interaction growth rates over the period. The first type of data represents the *pages*' reach and popularity to its audiences, and it is measured based on the number of likes of that particular Facebook page. The second one represents the *posts*' reach and popularity to their audiences, and it is measured based on the number of comments, reactions, and shares each post generates. These two types of variables are the indicators of a page's popularity among users (Schmidt et al., 2017). Six categories of posts were included in the interaction analysis: photos, links, statuses, Facebook videos, Facebook likes, and embedded (mostly YouTube) videos shared on Facebook. Interaction types were measured based on reactions, comments, and shares. Note that we excluded branded content from the count because of its financial endeavor.

3.4. Data Collection & Analysis

We extracted the data in CSV format using CrowdTangle, a public insights tool owned and operated by Facebook. While collecting the data based on the discussed parameters, we also used the following built-in filters in CrowdTangle: We set the language to *Bengali*, country of the page to *Bangladesh*, the data source to *Facebook Page*, and post type to *All*. After data extraction, we converted the CSV file into Microsoft Excel and cleaned and prepared the data for the final analysis. The combined likes of the nine pages during the selected period were 97.67 million (*mean [M]*=10.85, *standard deviation [SD]*=2.58) and interactions were 3.48 billion (B) (*M*=386.87, *SD*=232.78). Also, the total number of posts was 1.51 million (*M*=0.17, *SD*=0.16), with monthly 3,472 posts on average and 2,321 interactions per post. We used line graphs for a simple time-series analysis of page likes and their growth, and interactions and their growth over the period to understand the selected pages' popularity trends throughout the time. The statistical analysis was conducted in IBM SPSS Statistics 25 (IBM Co., Armonk, NY, USA), and the visualizations were produced using Tableau 2020.4 (Tableau Software, Seattle, WA, USA), a professional data visualization software.

4. RESULTS

4.1. Page Like Analysis

During the period, the Prothom Alo Facebook page

Table 1. Basic descriptions of the media samples

Page name	Media type	Website link	Page like[a)]	Page link (fb.com/)
Prothom Alo	Newspaper	http://prothomalo.com	15.12	/DailyProthomAlo
BBC News Bangla	Online news portal	http:// bbc.com/bengali	13.08	/BBCBengaliService
NTV	Television channel	http://ntvbd.com	12.88	/ntvdigital
RTV	Television channel	http://rtvonline.com	11.95	/rtvonline
Somoyer Kanthosor	Online news portal	http://somoyerkonthosor.com	10.35	/somoyerkonthosor
Daily Naya Diganta	Newspaper	https://dailynayadiganta.com	9.81	/nayadiganta
BDnews24.com	Online news portal	http://bdnews24.com	9.05	/bdnews24
Independent24.tv	Television channel	http://independent24.com	8.08	/IndependentTVNews
Kaler Kantho	Newspaper	http://kalerkantho.com	7.35	/kalerkantho

[a)]The data (in millions) were collected from Socialbakers (http://socialbakers.com) on 17 December 2020.

 https://doi.org/10.1633/JISTaP.2021.9.2.2

had the highest number of followers (*n*=15.13M; 15.49%), followed by BBC News Bangla (*n*=13.08M; 13.39%) and NTV (*n*=12.87M; 13.18%); Kaler Kantho (*n*=7.35M; 7.53%) was on the bottom of the list, followed by Independent24.tv (*n*=8.08M; 8.27%) and BDnews24.com (*n*=9.05M; 9.27%) (Table 2). Although NTV had the third-highest number of followers, it earned the highest page likes (*n*=5.45M; 73.39%), followed by Prothom Alo (*n*=4.25M; 39%) and RTV (*n*=3.29M; 38%).

Fig. 1 shows the changes in likes of individual pages during the period. Prothom Alo remained in the highest position with 10.89M likes in December 2016, followed by BBC News Bangla (*n*=9.77M) and Somoyer Kanthosor (*n*=8.74M); RTV (*n*=9.77M) overtook Somoyer Kanthosor (*n*=9.75M) in May 2017. The combined page likes increased from 74.06M in December 2016 to 97.68M in November 2020. However, the line of the page like in Fig. 2 shows an uneven growth: The line was steady from July 2017 to January 2020. The combined page likes again started to increase after a fall in February 2020. Overall, the combined page likes increased by 31.62% (*n*=23.61M) from December 2016 to November 2020.

Table 2. Indicators of page popularity

Page name	Total likes[a)]	Percentage	Like growth[a)]	Like growth (%)
Prothom Alo	15.13	15.49	4.25	39.00
BBC News Bangla	13.08	13.39	3.31	33.82
NTV	12.87	13.18	5.45	73.39
RTV	11.95	12.24	3.29	38.00
Somoyer Kanthosor	10.35	10.60	1.62	18.53
Daily Naya Diganta	9.81	10.04	1.26	14.72
BDnews24.com	9.05	9.27	1.47	19.39
Independent24.tv	8.08	8.27	1.59	24.60
Kaler Kantho	7.35	7.53	1.38	23.16

[a)]Values are in the millions.

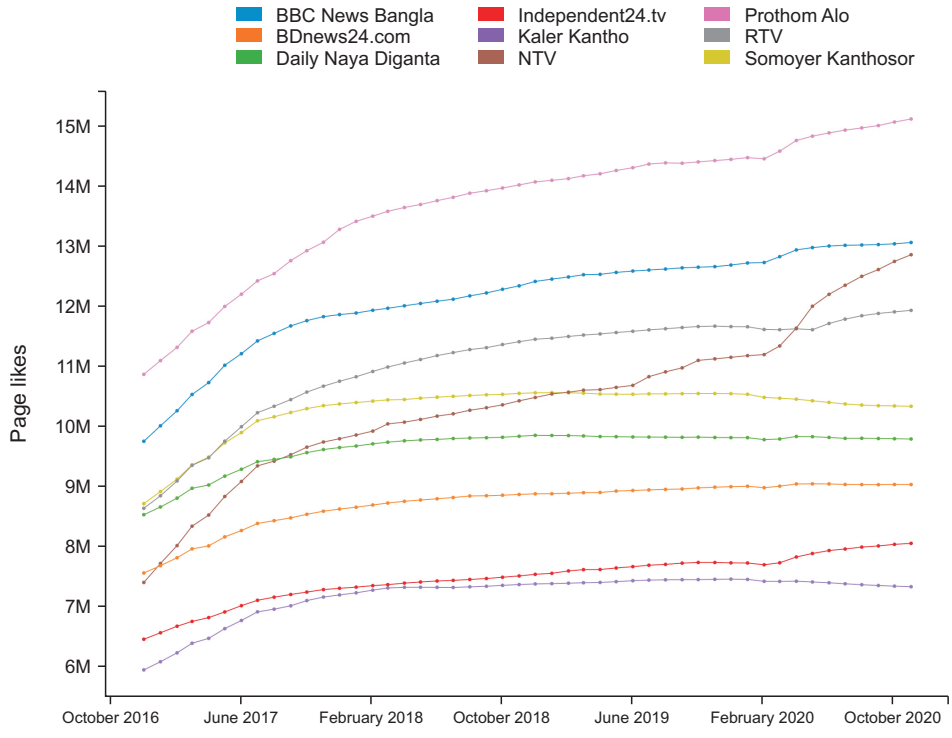


Fig. 1. Likes of individual pages.

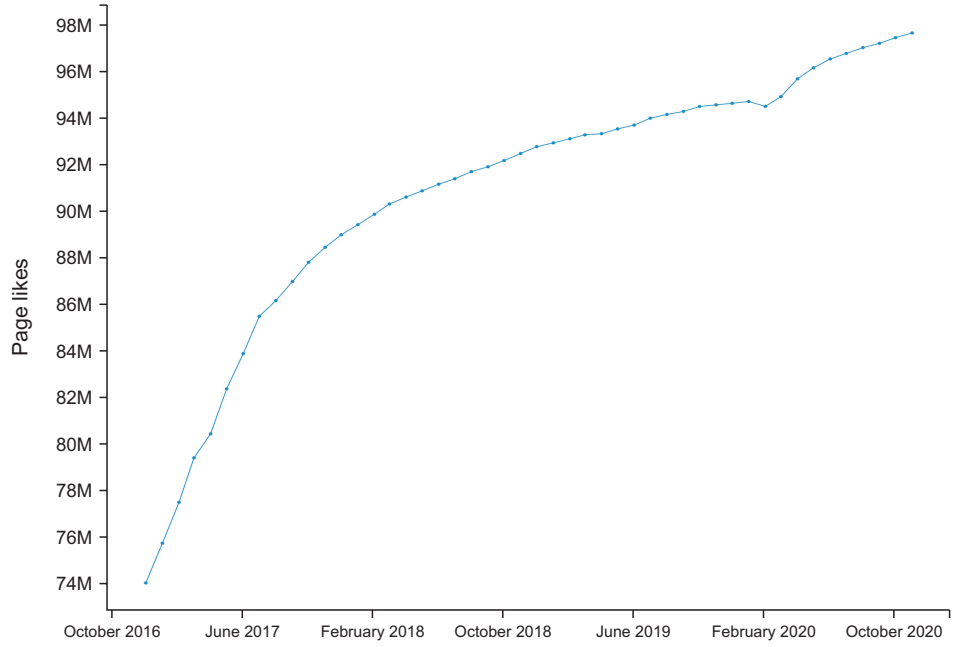


Fig. 2. Combined likes of the pages.

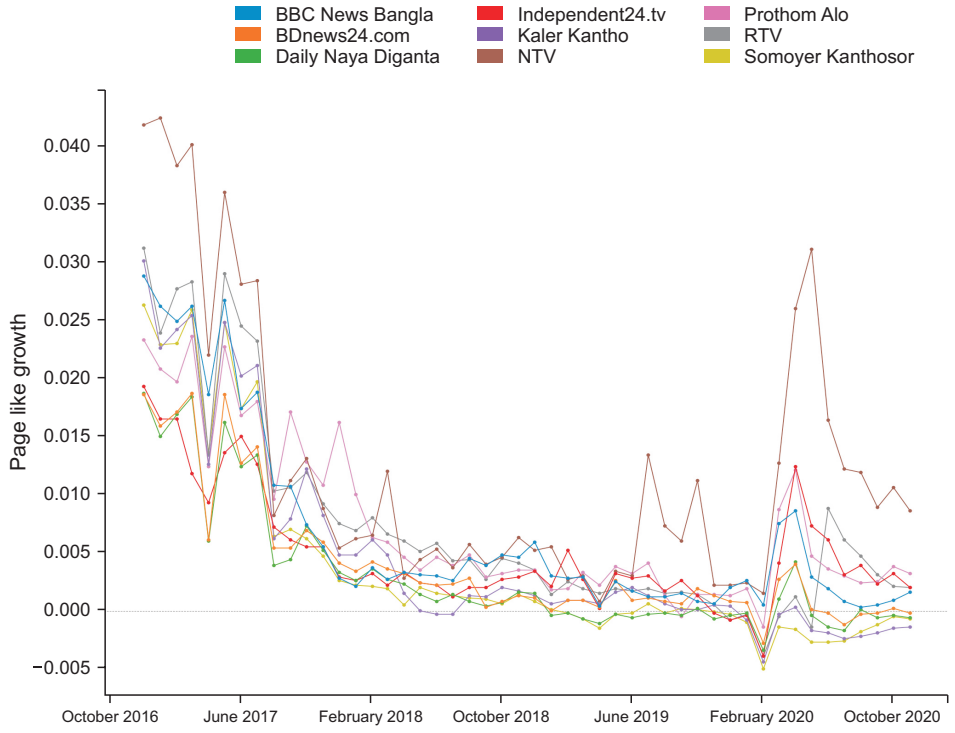


Fig. 3. Likes growth rate of individual pages.

The individual lines of growth rate show regular fluctuations from the very beginning. However, most of the lines were comparatively steadier from August 2017 to January 2020. While other lines were maintaining a rhythm, NTV (3.12%) experienced a remarkable surge in May 2020 before dropping again in the subsequent months (Fig. 3). The line of combined page-like growth rate shows a few heavy fluctuations from March to August 2017, and a steady drop from October 2017 to January 2020, with the largest drop in the next month (Fig. 4). The growth rate increased a little in March and April 2020 and again dropped in the following months. Afterward, the combined growth rate of page likes started to slowly increase again but with regular fluctuations. Although the com-

 https://doi.org/10.1633/JISTaP.2021.9.2.2

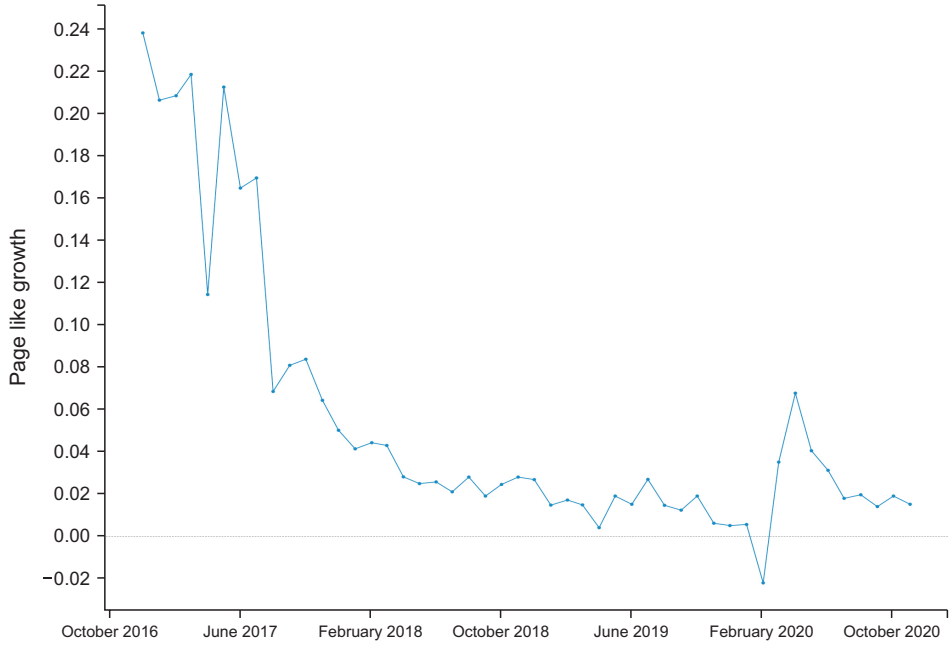


Fig. 4. Combined likes growth rate of the pages.

Table 3. Indicators of page interactions

Page name	Total interactions[a)]	Total posts[a)]	Percentage	Interactions per post	Cumulative interaction (%)	Average interaction (%)
Prothom Alo	897.4	0.11	7.11	8,413	91.84	0.07
NTV	574.77	0.16	10.71	3,579	42.17	0.04
Kaler Kantho	473.11	0.56	37.24	847	-97.26	0.01
Somoyer Kanthosor	346.45	0.09	5.90	3,913	9,832.05	0.04
Daily Naya Diganta	289.96	0.10	6.81	2,840	38.48	0.03
RTV	273.19	0.20	13.53	1,346	592.29	0.02
BDnews24.com	255.19	0.22	14.52	1,172	-26.46	0.01
BBC News Bangla	249.17	0.03	1.78	9,345	115.60	0.08
Independent24.tv	122.56	0.04	2.40	3,400	297.73	0.05

[a)]Values are in the millions.

bined page likes increased, the combined growth rate dropped by 93.56% (0.2236) during this period. A Pearson correlation coefficient analysis of the four years' data shows an almost perfect negative correlation between the combined page likes and page-like growth rate (r=-0.905) with p<0.01, which is statistically significant.

4.2. Interaction Analysis

Table 3 shows the basic descriptions of the interaction indicators. Prothom Alo (n=897.4M) had the highest interactions during the period, followed by NTV (n=574.77M) and Kaler Kantho (n=473.11M), and Independent24.tv (n=122.56M) had the lowest. In terms of the number of posts, Kaler Kantho (n=0.56M; 37.24%) was on the top of the list, followed by BDnews24.com (n=0.22M; 14.52%) and RTV (n=0.20M; 13.53%), and BBC News Bangla (n=0.03M; 1.78%) was on the bottom. On the other hand, BBC News Bangla (n=9,345) had the highest interactions per post, followed by Prothom Alo (n=8,413) and Independenet24.tv (n=3,400), and Kaler Kantho (n=847) had the lowest. The data show that though Kaler Kantho had the highest number of posts, its interactions per post were the lowest of all.

Most of the individual interaction lines show random fluctuations throughout time. For example, Prothom Alo, the highest interactive page, had 30.47M interactions in

December 2016, dropping to 27.93M in February 2017 and surging again after August 2020 (Fig. 5). After several fluctuations, it reached 31.42M interactions in November 2020, which was its highest. Like Prothom Alo, Independent24.tv also experienced an overall increase in interactions: from 4.14M in December 2016 to 6.38M in November 2020. However, unlike these two pages, the interactions of the other seven pages show an overall downward trend with several fluctuations: The total interactions for all pages decreased from 172.28M in December 2016 to 76.95M in November 2020. The pages' combined interaction line shows a fluctuating tendency as well. The line went downward from December 2016 to August 2019 with several ups and downs in regular intervals (Fig.

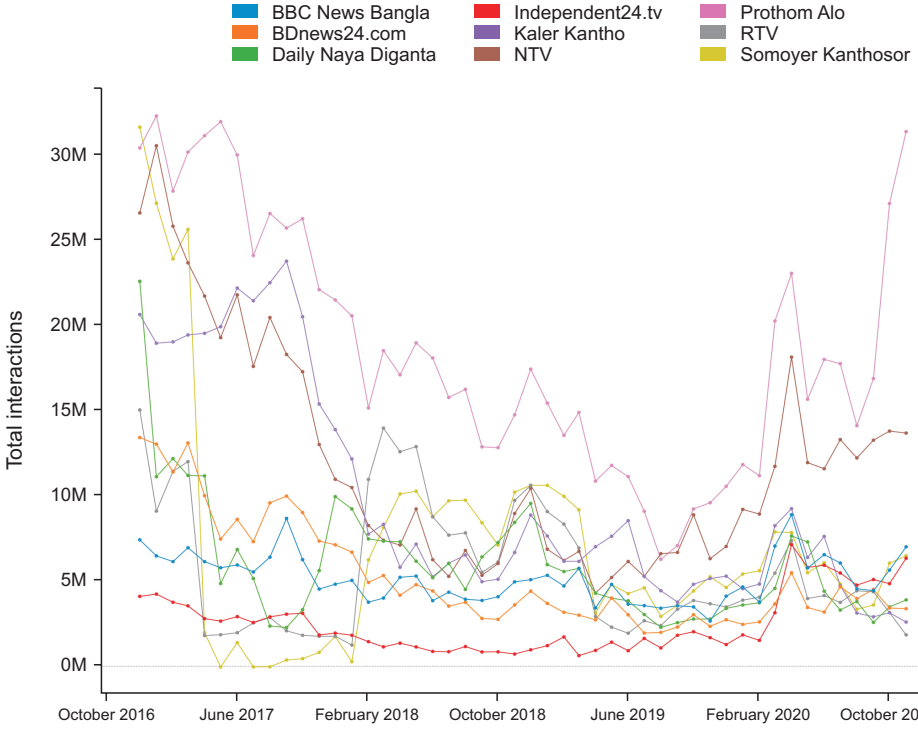


Fig. 5. Interactions of individual pages.

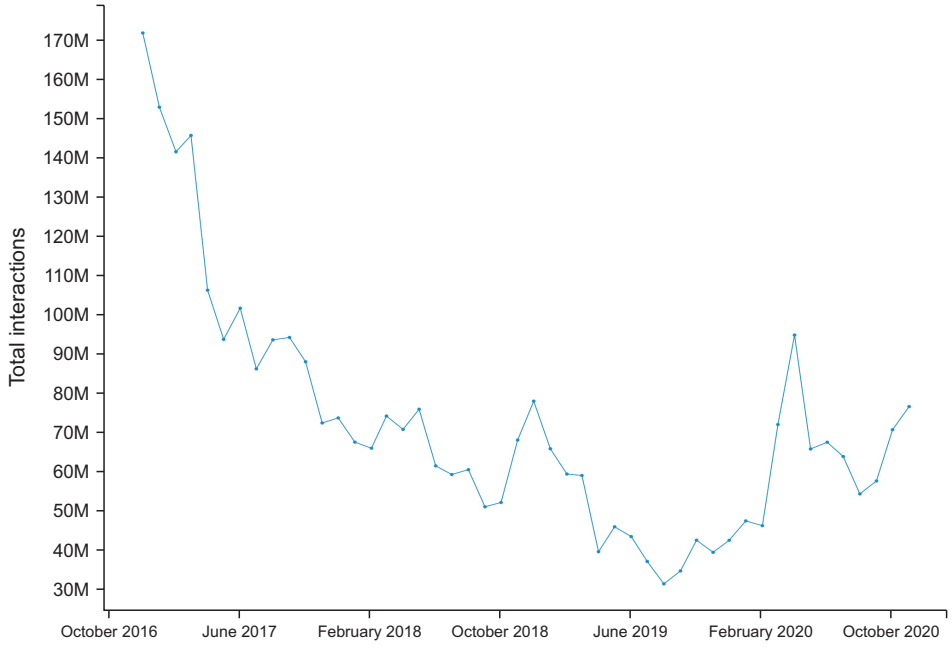


Fig. 6. Combined interactions of the pages.

 https://doi.org/10.1633/JISTaP.2021.9.2.2

6). The combined interactions experienced an irregular upward trend afterward until November 2020. Overall, during this period, the combined interactions dropped by 55.33% (n=95.32M).

Fig. 7 shows the uneven interaction growth rates of individual pages during the period, which somewhat resembles their interaction lines. Interactions of BBC News Bangla remained the highest for most of the time, followed by Prothom Alo, and both experienced several drops in their interaction rates before a gradual rise after September 2020. Interestingly, the overall interaction growth rates of all pages except Independent24.tv were lower in November 2020 than in December 2016: BBC News Bangla from 0.17% to 0.10%, Prothom Alo from 0.13% to 0.07%, RTV from 0.19% to 0%, NTV from 0.11% to 0.03%, BDnews24.com from 0.04% to 0.01%, Kaler Kantho from 0.03% to 0%, and Daily Naya Diganta from 0.09% to 0.01%. However, the interaction rate for Independent24.tv increased from 0.07% in December 2016 to 0.11% in November 2020, with a huge surge (0.18%) in April 2020. After a year of several fluxes, the line of combined interaction growth rate became steadier from January 2018 to March 2020 (Fig. 8). Overall, the combined interaction growth rate dropped by 59.14% (0.0055) from December 2016 to November 2020. A Pearson correlation coefficient test shows almost a perfect positive correlation between the combined interactions and interaction growth rate (r=0.967) with p<0.01, which is statistically significant.

5. DISCUSSION & CONCLUSION

This study aimed at understanding the popularity trends of the top nine Facebook pages of Bangladeshi media. Based on the historical data ranging between December 2016 to November 2020, we analyzed the page likes, like growth, total interactions, and interaction growth to answer the research question.

5.1. Key Findings

This study has two key findings. *First*, the combined likes of the media's Facebook pages increased, but their page-like growth rates dropped. Also, the growth of combined page likes became comparatively steadier compared to two years ago, and the page growth rate had a downward tendency at the end of the period. If both trends continue, after a certain time the combined page likes will be dropping instead of increasing. This result also infers that most of the media pages are failing to attract more Facebook users in recent times. It is possible that the number of Facebook users does not always increase at the same rate, so the numbers of page-likes may not also increase at the same rate. However, it can also be considered that although the page-likes fluctuated throughout the period, the number of Facebook users was likely to

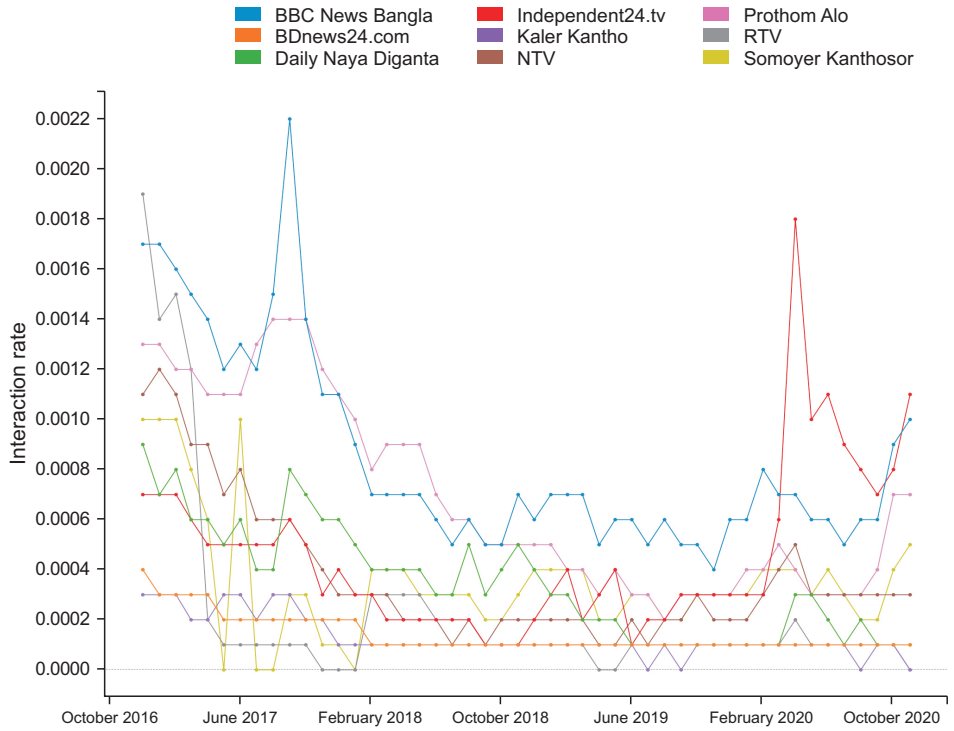


Fig. 7. Interaction growth rates of individual pages.

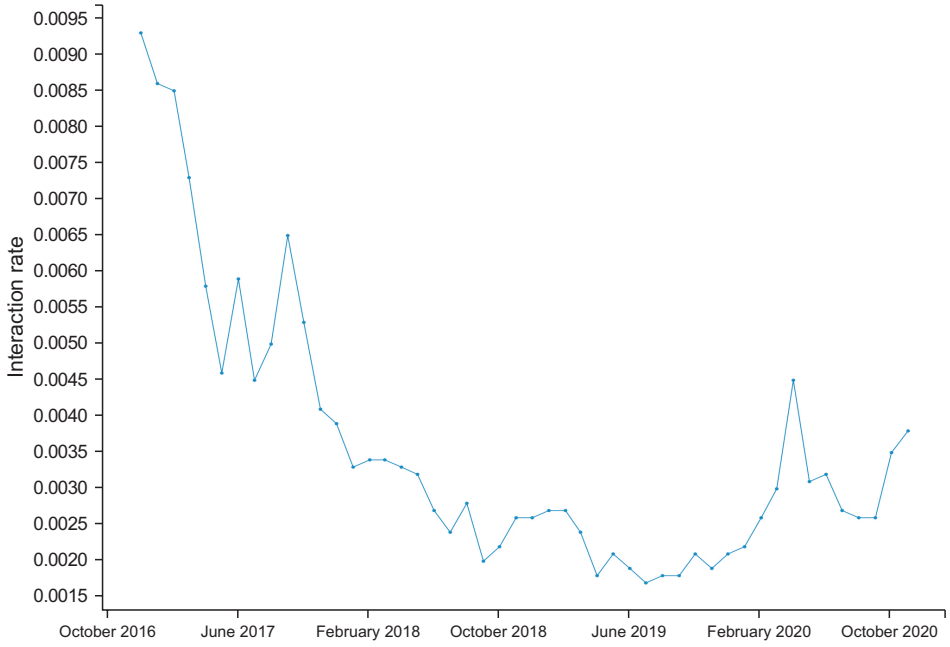


Fig. 8. Combined interaction growth rates of the pages.

increase (Kemp, 2020). *Second,* both the pages' combined interactions and interaction growth rates dropped, although, based on their previous (oscillating) trends, there is a possibility that both numbers could increase again in the future. The results indicate that Facebook users are becoming less interested not only in the media's Facebook pages but also in the contents they share on their pages. However, some pages may be exceptional, generating more interactions, although such instances are insignificant. Overall, the findings suggest that the Bangladeshi media's Facebook pages are perhaps gradually losing their popularity among Facebook users with their reduced user engagement. This may also be interpreted in two other ways: Facebook's aggregate appeal as a news source is decreasing to its users, or Bangladeshi media outlets are losing their appeal to Facebook users. Note that this study analyzed Facebook pages of three different types of media outlets: television channels, newspapers, and online news portals. Interestingly, the popularity trends are almost identical for all types of media, showing no exceptional pattern. It suggests that Facebook users do not differentiate between media outlets while consuming news.

Previous studies demonstrated Facebook's importance as an information source (Bene, 2017; Silver & Matthews, 2017). A few studies also found that Facebook's popularity as a news source is increasing, which means more people are consuming news from Facebook nowadays (Lampe et al., 2012; Shearer & Gottfried, 2017). However, the present findings contradict the previous studies. This may have a few reasons: The data of previous studies were collected mainly from users through a survey or interview, which may have represented a partial picture of the scenario; the previous results were time-specific and temporal; the studies did not consider Facebook's historical data. Considering these limitations of the literature, this study tried to provide more updated and reliable insights into the news consumption behavior of Facebook users, analyzing a large amount of historical and multifaceted public engagement data.

A range of factors can influence Facebook users' news consumption and information-seeking behavior. For example, bandwidth, location of the user, and the user's demographics (e.g., age, income, and education), according to Nguyen (2003), may influence online news consumption. On the other hand, Mishra and Bakshi (2017) identified four factors: credibility, ideology, entertainment, and localization of the news, that influence people's news consumption behavior. In the context of Bangladesh, Hasan and Mohua (2020) found that a user's education level, interest level, and trust in the source mostly influence user news consumption behavior. However, we presume that the discussed factors are inappropriate and/or inapplicable in the present context. Therefore, we presume that the following four factors could be responsible for Facebook's decreased popularity as a news source: information overload, exposure to incidental news, users' selective exposure and different aims of using Facebook, and conflict between the media's agenda and users' interest.

 https://doi.org/10.1633/JISTaP.2021.9.2.2

First, excessive news sharing tendency can be responsible for some pages' decreasing interaction rates. Facebook users' personal and collective preferences determine their news taste, and they may not find more news desirable. Therefore, unimportant information may lead to information overload, which has negative impacts on users' relations to that media (Holton & Chyi, 2012). When users find themselves in an information-overflowed situation, they start to selectively expose themselves to specific Facebook posts, avoiding some overflowed sources (Lee et al., 2017; Pentina & Tarafdar, 2014). For example, Kaler Kantho shared the highest number of posts with a reducing interaction rate, whereas BBC News Bangla's interaction growth rate line remained higher with the lowest number of posts. It suggests that users may consume less news from a page with more news, of which most might be unimportant to them. Moreover, a user may also feel information overload thanks to multiplatform news consumption. The Bangladeshi media industry has experienced three consecutive advancements in print, broadcast, and online media in the 1990s, 2000s, and 2010s, respectively (Haider & Samin, 2014). In September 2020, 126 online news media outlets received clearance through a government-enacted registration process (Dhaka Tribune, 2020). However, in practice, the number of unofficial and low-quality online media outlets is much higher than the officially-certified outlets, and they tend to promote copy-paste journalism instead of intellectual or professional journalism. Most of such outlets usually operate their own Facebook pages to attract more audiences, providing copy-pasted, sensational, unimportant, and clickbait news items (Ercegovac & Sredojević, 2014; Mansurova, 2014). Thus, Facebook users could be affected by redundant news items that may cause information overload as well.

Second, the fluctuations in interactions could be a reason for users' exposure to incidental news, which is a common phenomenon on Facebook due to its algorithms (Boczkowski et al., 2018). Young users interact more with such types of content (Boczkowski et al., 2017). Interestingly, according to Socialbakers, the largest share of Bangladeshi Facebook users is young, aged between 18 to 24. It implies that they could be frequently exposed to incidental news, which defines the fluctuations in pages' interaction and growth rates. Important to note is that users are not always likely to or interested in interacting with incidental news. Rather, their engagement depends on a few factors, such as prior knowledge and interest (Karnowski et al., 2017).

Third, users' preferences and exposure to news are selective (Schmidt et al., 2017) and they often filter out information and avoid information sources that do not match their interests (Lee et al., 2017). Both tendencies explain the increasing and decreasing popularity and interaction rates of different pages. Interestingly, studies also found that the primary reason for many users' Facebook use may not be accessing or seeking news or important information (Bergström & Belfrage, 2018). Rather, they use it to belong and to self-represent: The first goal is an "intrinsic drive to affiliate with others and gains social acceptance," and the second one is a technique of impression management (Nadkarni & Hofmann, 2012, p. 5). This suggests a possibility that the primary concern of Bangladeshi Facebook users is something other than accessing news or news-related content. Though it may define the gradual drop in page likes and interactions, it may not properly explain their irregular fluctuations.

Fourth, the conflicting ideals between the media's agenda and news consumers' interests could be another reason for some pages' decreased popularity (Rather & Ganaie, 2018; Wilson, 1997). We have already explained that users' news preferences are highly motivated by their personal interests and agendas. On the other hand, every media organization has certain agendas that shape its shared contents. Media interests and user preferences could contradict, which may shift users to another news source, reducing the popularity of that particular Facebook page (Schmidt et al., 2017). Bangladeshi media outlets have a few common criticisms: Due to competition in the media industry, some of them often sensationalize or exaggerate a simple news event; some create misperception among the public instead of being an honest watchdog; some are oblivious about their social responsibility and apathetic about public criticisms (Rahman, 2016). For these reasons, Haider and Samin (2014) have termed Bangladeshi media as "mass alienated media." On the other hand, a large number of (online) media outlets produce and spread fake news, deceiving potential news consumers instead of supplying reliable news (Al-Zaman et al., 2020); this hurts news consumers' reputations (Altay et al., 2020). Due to users' lower information and/or digital literacy levels, many of them are also unaware of quality news and information ethics: what types of news they should consume from which sources, what they should expect from the professional media outlets, and which content can be shared publicly or not.

It could be possible that users are only losing interest in the media's Facebook pages, to be specific. This means Facebook's demand as a news source is not at risk, rather

more users seek news from other personalized online sources, such as online opinion leaders, other than the media's Facebook pages (Bergström & Belfrage, 2018; Bodendorf & Kaiser, 2009). If this is the case, media outlets should be worried about it. As stated earlier, Bangladesh's mainstream media outlets are becoming more dependent on social media; they try to capitalize their Facebook pages to attract more audiences and generate more interactions (Fahad, 2014), and Facebook pages help to produce more of the media's followers (Welbers & Opgenhaffen, 2018). In such circumstances, the decreasing popularity of media Facebook pages and their shared contents is alarming. Therefore, for their business to survive, and for their reputation and public reach, they must understand users' motivations and interests, and their news-seeking and consumption behavior.

5.2. Strengths and Limitations

This research has a few limitations. First, the data for this research were collected using CrowdTangle, that relies on automation. For that reason, the data could not be cross-checked manually during and after the collection period to find out whether there is any missing or corrupted data. For example, the dataset shows that the Somoyer Kanthosor page did not post anything in May and July 2017, and it could not be verified. However, there could be some other issues as well that can explain this, such as page inactivity and temporary blocks. Second, this study drew inferences only based on visible users' interaction data, excluding silent users or *lurkers*, who could be active information consumers but produce no digital trace (Benevenuto et al., 2009; Gong et al., 2015). Third, the correlation coefficient analysis between the like counts of the pages and their growth rates showed a negative correlation for the four years' data, which may not explain the real tendency. Therefore, this finding demands further study with an extended time span for a more acceptable and generalizable result.

Besides these limitations, this study has four major strengths. First, academic endeavors to understand different media outlets' popularity and public engagement in social media platforms was limited. In this gap, the present study offered some insights. Second, this study analyzed a larger dataset compared to the previous studies (e.g., Bene, 2017; Lampe et al., 2012; Silver & Matthews, 2017) to explore the news media's online interactions. From this aspect, the present results could be more reliable and inclusive. Third, little was explored about online news consumption in Bangladesh to date based on historical popularity data. Fourth, these findings can provide some insights to media managers regarding the overall climate of Facebook news consumption, which would help them to come up with a more effective plan to increase more public participation in their Facebook pages. More future research should focus on the news consumption patterns of the heterogenous Facebook users of Bangladesh to comprehend their underlying predispositions.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Alam, A. S. (2020). *How newspapers' social media editors in Bangladesh use official social media accounts*. (master's thesis). University of Mississippi, Oxford, MS, USA.

Al-Rawi, A. (2017). News values on social media: News organizations' Facebook use. *Journalism*, 18(7), 871-889. https://doi.org/10.1177/1464884916636142.

Altay, S., Hacquin, A.-S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. *New Media & Society*. https://doi.org/10.1177/1461444820969893.

Al-Zaman, Md. S., Sife, S. A., Sultana, M., Akbar, M., Ahona, K. T. S., & Sarkar, N. (2020). Social media rumors in Bangladesh. *Journal of Information Science Theory and Practice*, 8(3), 77-90. https://doi.org/10.1633/JISTaP.2020.8.3.6.

Bene, M. (2017). Influenced by peers: Facebook as an information source for young people. *Social Media + Society*, 3(2). https://doi.org/10.1177/2056305117716273.

Benevenuto, F., Rodrigues, T., Cha, M., & Almeida, V. (2009, November 4-6). Characterizing user behavior in online social networks. In A. Feldmann, & L. Mathy (Eds.), *Proceedings of the IMC '09: 9th ACM SIGCOMM conference on Internet measurement* (pp. 49-62). Association for Computing Machinery. https://doi.org/10.1145/1644893.1644900.

Bergström, A., & Belfrage, M. J. (2018). News in social media. *Digital Journalism*, 6(5), 583-598. https://doi.org/10.1080/21670811.2018.1423625.

Boczkowski, P. J., Mitchelstein, E., & Matassi, M. (2017, January 3-7). Incidental news: How young people consume news on social media. In T. X. Bui, & R. Sprague Jr. (Eds.), *Proceedings of the 50th Annual Hawaii International Conference on System Sciences* (pp. 1785-1792). University of Hawaii. https://doi.org/10.24251/HICSS.2017.217.

Boczkowski, P. J., Mitchelstein, E., & Matassi, M. (2018). "News comes across when I'm in a moment of leisure": Under-

standing the practices of incidental news consumption on social media. *New Media & Society*, 20(10), 3523-3539. https://doi.org/10.1177/1461444817750396.

Bodendorf, F., & Kaiser, C. (2009, November 2-6). Detecting opinion leaders and trends in online social networks. In I. King, J. Li, G.-r. Xue, & J. Tang (Eds.), *Proceedings of the SWSM '09: 2nd ACM workshop on Social Web Search and Mining* (pp. 65-68). Association for Computing Machinery. https://doi.org/10.1145/1651437.1651448.

Chadwick, A. (2011). The political information cycle in a hybrid news system: The British prime minister and the "Bullygate" affair. *International Journal of Press/Politics*, 16(1), 3-29. https://doi.org/10.1177/1940161210384730.

Dhaka Tribune. (2020). *Dhaka Tribune, 91 other newspapers, get permission to register websites*. https://www.dhakatribune.com/bangladesh/2020/09/04/dt-91-other-newspaper-portals-get-permission-to-register.

Dominick, J. R. (2009). *The dynamics of mass communication: Media in the digital age*. 10th ed. McGraw-Hill.

Ercegovac, I., & Sredojević, M. (2014). Copy/paste journalism and the Internet: A used product in the new media. *Kultura*, 145, 198-213. https://doi.org/10.5937/kultura1445198e.

Fahad, S. A. (2014). The emergence and impact of social media on the mainstream journalism in Bangladesh. *Jahangirnagar University Journal of Journalism and Media Studies*, 1. https://jmsju.com/archives/697.

Fu, S., Li, H., Liu, Y., Pirkkalainen, H., & Salo, M. (2020). Social media overload, exhaustion, and use discontinuance: Examining the effects of information overload, system feature overload, and social overload. *Information Processing & Management*, 57(6), 102307. https://doi.org/10.1016/j.ipm.2020.1023070.

Gong, W., Lim, E., & Zhu, F. (2015, May 26-29). Characterizing silent users in social media communities. In D. Quercia, B. Hogan, M. Cha, C. Mascolo, & C. Sandvig (Eds.), *Proceedings of the 9th International AAAI Conference on Web and Social Media* (pp. 140-149). AAAI Press.

Haider, S., & Samin, S. (2014). *Gonojogajog Totto o Proyog* [Mass communication: Theory and practice]. 2nd ed. Bangladesh Press Institute. Bengali.

Hasan, M., & Mohua, A. A. (2020). Factors influencing online non-main stream news consumption in Bangladesh. *Asian Journal of Management*, 11(2), 147-153. https://doi.org/10.5958/2321-5763.2020.00023.2.

Hermida, A., Fletcher, F., Korell, D., & Logan, D. (2012). Share, like, recommend. *Journalism Studies*, 13(5-6), 815-824. https://doi.org/10.1080/1461670X.2012.664430.

Holton, A. E., & Chyi, H. I. (2012). News and the overloaded consumer: Factors influencing information overload among news consumers. *Cyberpsychology, Behavior, and Social Networking*, 15(11), 619-624. https://doi.org/10.1089/cyber.2011.0610.

Hong, S. (2012). Online news on Twitter: Newspapers' social media adoption and their online readership. *Information Economics and Policy*, 24(1), 69-74. https://doi.org/10.1016/j.infoecopol.2012.01.004.

Karnowski, V., Kümpel, A. S., Leonhard, L., & Leiner, D. J. (2017). From incidental news exposure to news engagement. How perceptions of the news post and news usage patterns influence engagement with news articles encountered on Facebook. *Computers in Human Behavior*, 76, 42-50. https://doi.org/10.1016/j.chb.2017.06.041.

Kemp, S. (2020). Digital 2020: *Bangladesh*. https://datareportal.com/reports/digital-2020-bangaldesh.

Lampe, C., Vitak, J., Gray, R., & Ellison, N. (2012, May 5-10). Perceptions of Facebook's value as an information source. In J. A. Konstan, E. H. Chi, & K. Höök (Eds.), *Proceedings of the CHI '12: SIGCHI Conference on Human Factors in Computing Systems* (pp. 3195-3204). Association for Computing Machinery. https://doi.org/10.1145/2207676.2208739.

Lee, S. K., Lindsey, N. J., & Kim, K. S. (2017). The effects of news consumption via social media and news information overload on perceptions of journalistic norms and practices. *Computers in Human Behavior*, 75, 254-263. https://doi.org/10.1016/j.chb.2017.05.007.

Lysak, S., Cremedas, M., & Wolf, J. (2012). Facebook and Twitter in the newsroom: How and why local television news is getting social with viewers? *Electronic News*, 6(4), 187-207. https://doi.org/10.1177/1931243112466095.

Mansurova, V. D. (2014). Intellectual journalism vs copy-paste journalism. *World of Media: Yearbook of Russian Media and Journalism Studies*, 4, 252-264.

McCombs, M. (2013). *Setting the agenda: The mass media and public opinion*. Polity.

Mishra, P., & Bakshi, M. (2017, June 27-July 1). Determinants of media consumption: Evidences from an emerging market: An abstract. In P. Rossi, & K. Krey (Eds.), *Proceedings of the 2017 Academy of Marketing Science (AMS) World Marketing Congress (WMC)* (pp. 265-266). Springer. https://doi.org/10.1007/978-3-319-68750-6_82.

Nadkarni, A., & Hofmann, S. G. (2012). Why do people use Facebook? *Personality and Individual Differences*, 52(3), 243-249. https://doi.org/10.1016/j.paid.2011.11.007.

Nguyen, A. (2003). The current status and potential development of online news consumption: A structural approach. *First Monday*, 8(9). https://doi.org/10.5210/fm.v8i9.1075.

Paulussen, S., & Harder, R. A. (2014). Social media references

in newspapers. *Journalism Practice*, 8(5), 542-551. https://doi.org/10.1080/17512786.2014.894327.

Pentina, I., & Tarafdar, M. (2014). From "information" to "knowing": Exploring the role of social media in contemporary news consumption. *Computers in Human Behavior*, 35, 211-223. https://doi.org/10.1016/j.chb.2014.02.045.

Pletikosa Cvijikj, I., & Michahelles, F. (2013). Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining*, 3(4), 843-861. https://doi.org/10.1007/s13278-013-0098-8.

Rahman, S. (2016). *Kutorko Bitorke Gonomaddhom* [Mass media in good and bad debates]. Shrabon. Bengali.

Rather, M. K., & Ganaie, S. A. (2018). Information seeking models in the digital age. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology*, 4th ed (pp. 4515-4527). IGI Global. https://doi.org/10.4018/978-1-5225-2255-3.ch392.

Rodriguez, M. G., Gummadi, K., & Schoelkopf, B. (2014, June 1-4). Quantifying information overload in social media and its impact on social contagions. In E. Adar, & P. Resnick (Eds.), *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (pp. 170-179). University of Michigan. https://ojs.aaai.org/index.php/ICWSM/article/view/14549.

Saiful, H., & Park, J. (2019). A study on the integration of UGC in broadcast journalism: An evidence from Bangladesh. *Journal of Digital Convergence*, 17(1), 301-311. https://doi.org/10.14400/JDC.2019.17.1.301.

Schick, A. G., Gordon, L. A., & Haka, S. (1990). Information overload: A temporal approach. *Accounting, Organizations and Society*, 15(3), 199-220. https://doi.org/10.1016/0361-3682(90)90005-F.

Schmidt, A. L., Zollo, F., Vicario, M. D., Bessi, A., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences*, 114(12), 3035-3039. https://doi.org/10.1073/pnas.1617052114.

Shearer, E., & Gottfried, J. (2017). *News use across social media platforms 2017*. https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017.

Silver, A., & Matthews, L. (2017). The use of Facebook for information seeking, decision support, and self-organization following a significant disaster. *Information, Communication & Society*, 20(11), 1680-1697. https://doi.org/10.1080/1369118X.2016.1253762.

Stassen, W. (2010). Your news in 140 characters: Exploring the role of social media in journalism. *Global Media Journal*, 4(1), 116-131. https://doi.org/10.5789/4-1-15.

StatCounter. (2020). *Social media stats Bangladesh*. http://gs.statcounter.com/social-media-stats/all/bangladesh.

Tewksbury, D., & Rittenberg, J. (2012). *News on the Internet: Information and citizenship in the 21st century*. Oxford University Press.

Welbers, K., & Opgenhaffen, M. (2018). Social media gatekeeping: An analysis of the gatekeeping influence of newspapers' public Facebook pages. *New Media & Society*, 20(12), 4728-4747. https://doi.org/10.1177/1461444818784302.

Wilson, T. D. (1997). Information behaviour: An interdisciplinary perspective. *Information Processing & Management*, 33(4), 551-572. https://doi.org/10.1016/S0306-4573(97)00028-9.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(2): 33-46, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.2.3

Drive-thru Library Service in Korea

Seong-Kwan Lim*
School of Librarianship, Graduate School of Education,
Kyonggi University, Suwon, Korea
E-mail: kglimpro@kyonggi.ac.kr

ABSTRACT

The purpose of this paper is to analyze the case of 'drive-thru' services newly introduced and tried in libraries in Korea under the influence of COVID-19, and to develop and propose a service model so that this service can be continuously applied to all libraries in the future. Therefore, the method of study was selected and analyzed by selecting one of the representative libraries that provide related services in Seoul Special City, Incheon Metropolitan City, and Anyang City of Gyeonggi Province. In addition, a focus group interview was conducted with twelve people in charge to find a way to apply the drive-thru service to the library. As a result, the library's drive-thru service is a way to fulfill the library's original purpose of providing information materials while minimizing face-to-face contact with users. It was concluded that it is a suitable method for a library of complex buildings, where there is a lack of parking space. In addition, it was deduced that it may be one of the ways to use the library efficiently for office workers who are unable to use library services during the opening hours. Therefore, if the drive-thru service is implemented according to the developed model, it is expected to increase the library visit rate and data utilization rate.

Keywords: drive-thru, library service, library circulation, library circulation system, drive-thru service system

Received: February 16, 2021 **Revised:** May 3, 2021
Accepted: May 6, 2021 **Published:** June 30, 2021

***Corresponding Author:** Seong-Kwan Lim
https://orcid.org/0000-0002-9735-1717
E-mail: kglimpro@kyonggi.ac.kr

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© Seong-Kwan Lim, 2021

1. INTRODUCTION

The library, which has been established and operated since around 2300 B.C., provides functions related to knowledge, culture, education, and recreation, and emphasizes functions such as knowledge information, culture, education, and recreation in line with the needs of the development and change of the era. Furthermore, at the present time when the Fourth Industrial Revolution era has begun, the development of science and technology represented by artificial intelligence has created a ubiquitous-based smart library to create a much faster and more convenient operating system than before and is working on services.

This change can be said to be the result of efforts not to be dismissed in society, thanks to the necessity and importance of being a necessary place for users, as a library is also an evolving organism. This is because, in accordance with the change of the times, if the library does not develop and implement new services, it will be disregarded by users and eventually shunned. In "The future of employment report: How susceptible are jobs to computerization?", published by Frey and Osborne (2017) at Oxford University in the UK, the prospect of losing librarianships among 702 jobs in the United States is ranked 360, which is a result of this crisis. Since this report focuses on jobs that are likely to be replaced by artificial intelligence, there is no need to broadly interpret that the results themselves are the future of the entire library. This is because the library's role and function of preserving important materials will be maintained regardless of any changes in the future. However, I do not want to see the library turned into a storehouse of books with various artificial intelligence services being implemented instead of a librarian.

In the meantime, much research has been done on library services worldwide. Looking at the results, it can be seen that library services have also made remarkable developments with the development of the times, and it can be predicted that further changes will be made. However, studies on 'drive-thru' services in libraries have yet to be seen. Therefore, in this study, it is meaningful to propose a drive-thru service as a method of library service suitable for the present era.

2. LITERATURE REVIEW

2.1. Public Library Service

According to the International Federation of Library Associations and Institutions (IFLA)/UNESCO Public Library Declaration, public libraries are local information centers that make all kinds of knowledge and information readily available to users. It is a fundamental human right to have access to information and to know it, and now more information is available than ever before. As public services are open to everyone, public libraries not only provide access to a wide variety of information resources, but also play a very important role in collecting, organizing, and utilizing information. Therefore, it can be said that the value of public libraries is assessed according to the services they provide to the community (IFLA, 1994).

Therefore, public libraries provide a variety of services to satisfy the needs of users not only within the library but also within the local community. Libraries should make it easy for everyone to access library services, including those who have difficulty reading printed materials. The services that users can easily access though various forms, media, and the Internet include lending books and other materials for in-house use, information services using printed materials and electronic media, readers' services including reservations, community information services, user education including text-learning program support, programs and events, blogs, mobile phone text services, reference services, and publicity using modern communication tools such as social networking (Koontz & Gubbin, 2010).

2.2. Drive-thru Services

'Drive-thru,' a way to purchase products while in a car, is a service strategy that has spread all over the world since it first started in the United States in the 1930s. The Grand National Bank in St. Louis, a major US financial hub, opened the world's first deposit-only drive-thru window. Later, in 1974, the Red Giant Hamburg in Springfield, Missouri, introduced drive-thru service for the first time in a fast-food restaurant. McDonald's, which has many chain stores around the world, started its first drive-thru service in 1975, and in Korea in 1992 at the McDonald's Busan Haeundae branch.

There is also the concept of "click and collect," which is similar to drive-thru, but is also used interchangeably with "order and collection" or "pickup order." Click and collect allows customers to receive products directly from designated locations outside of their homes. It has become a new and more powerful purchasing method for many people in the current era where online shopping is becoming more and more ubiquitous (even without the influence of the COVID-19 pandemic). According to Dalin-Kaptzan (2020), click and collect, also known as online

 https://doi.org/10.1633/JISTaP.2021.9.2.3

purchase and in-store collection, is a form of omni-channel flow in which shoppers buy online and receive orders from physical locations instead of having items shipped to them. The pickup location is usually a local retail store, but it could also be a post office, a special purpose locker, or another retail store.

When the world was confused by the influence of COVID-19 in 2020, Korea was the first nation in the world to implement a screening clinic that can collect a sample from the nose and mouth, measure body temperature, and disinfect while riding in a car. The drive-thru clinic was evaluated as an innovative idea by major media outlets around the world such as CNN and Agence France-Presse. Since then, the drive-thru screening clinic has become a COVID-19 response model in many countries around the world, and has been extended to funerals, weddings, and library services aimed at avoiding or minimizing face-to-face contact. It was possible to increase the convenience of service by allowing citizens to continue to use the library.

Most of the studies on drive-thru services that have been published have been related to banking, fast food, coffee, and pharmacy services, but there are none about libraries. For this reason, the literature review here focuses on drive-thru services related to banks, fast food, coffee, and pharmacies, and aims to find ways to apply them to library services. We also looked at key resources for libraries in responding to the coronavirus pandemic provided by IFLA, as well as news articles about public libraries in the United States responding to the COVID-19 situation.

Morcos et al. (1992) stated that the employees working in the drive-thru window are the main link between the company and the customer who wants convenient service, and a factor that has a great influence on the reputation with customer return decisions. Therefore, it was necessary to set consistent educational standards for improved job performance accuracy and kindly services.

Song (2002) presents the problems of the existing McDonald's drive-thru system and proposes a multimedia drive-thru system as a solution, and provides an exemplary usage process of a new system, a content flow chart, and conditions for graphical user interface design.

Allon et al. (2011) said that drive-thru stores are a system that allows busy consumers to receive products without getting out of their cars and it is a system that provides consumers with conveniences such as reducing waiting times and travel.

See also Kucukusta et al. (2015), who said that consumers can feel usefulness from the convenience provided with drive-thru, and this usefulness helps them to form a positive attitude towards the products or services.

Abu Farha et al. (2017) surveyed 226 pharmacists in Jordan for their awareness of drive-thru services. As a result, 85.5% of pharmacists said they knew drive-thru service, but only 27% said they would be willing to do it. The reason is that the drive-thru service would make them feel like a fast-food worker, which would negatively affect the job image of a pharmacist. However, pharmacists working in chain community pharmacies were more aware than independent communities were, and Jordan's pharmacists also needed to adopt the service better because the service system was well established worldwide.

According to IFLA's Key Resources for Libraries in responding to the Coronavirus Pandemic, which was last updated on October 13, 2020, libraries around the world face very different circumstances, from widespread maintenance of full services to complete closure. Among them, the methods used by countries with minimal services include fully closing spaces and only offering the possibility to borrow or return books at a counter, or via a book drop. Some countries are experimenting with drive-through pick-up and return of books. Others are only allowing visitors who have pre-booked (IFLA, 2021).

According to the results of a survey with 2,545 United States public libraries conducted by the Public Library Association (2020), 98% of the libraries were closed due to the COVID-19 pandemic. Overall, online renewal policies expanded by 76%, online services such as e-books and streaming media by 74%, and expanded access to digital resources, such as operating a virtual program, 61%. In addition, new services were developed and provided to respond to the COVID-19 situation, such as making curbside pick-up available possible, with 22% of libraries surveyed.

As reported by KSN in Wichita, Kansas during their December 17, 2020 newscast:

> As the pandemic continues to affect the way our community operates, Wichita Public Libraries are finding ways to better serve their readers. In particular, the Westlink branch, the busiest branch in the city, has implemented a new curbside drive-thru system starting last Thursday. The library opened up designated areas for material collection, returns, and printing services, with the staff in charge coming out to support users' requests (Montero, 2021).

As a result of reviewing the literature as described above, drive-thru is a service that helps to form a positive

attitude because it is a system that provides consumers with conveniences to reduce waiting time and movement. However, librarians might consider that the drive-thru service compares them to fast food workers, bestowing a less professional view upon them and negatively affecting further library applications.

3. METHOD

The purpose of this paper is to analyze the case of drive-thru services newly introduced and tried in libraries in Korea under the influence of COVID-19, and to develop and propose a service model so that this service can be applied to all libraries at all times in the future.

Therefore, in order to achieve the purpose of the study, each of the representative libraries that perform related services in Seoul Special City, Incheon Metropolitan City, and Anyang City of Gyeonggi Province was selected for analysis of the operation cases. The reason for the selection of these sites for the analysis of operational cases is that the three regions are the metropolitan areas centered on the capital of Korea, and each library is a place that actively implemented drive-thru services on behalf of the city.

A focus group interview, which was the method chosen for this study, modified the question method suggested by Krueger and Casey (2009). Four types of introductory, main, and closing questions were conducted utilizing an open-ended method of questioning. Analyses of these questions are part of the method suggested by Morgan and Scannell (1998). The focus group interviews were conducted in real time on March 12, 2021 using the ZOOM program (a cloud-based video conferencing service program). The questionnaires were distributed in advance via e-mail, with a total of 12 librarians – one in charge of planning the drive-thru service for each library, and three other librarians involved with the continuing operation of the service.

4. ANALYSIS

4.1. Case Analysis 1: Seoul Special City

As of 2019, there are a total of (180) public libraries in Seoul. Among them, drive-thru service cases were examined focusing on the Namsan Library, belonging to the Seoul Metropolitan Office of Education. Table 1 summarizes some of the drive-thru service promotion plans implemented by Namsan Library.

Looking at drive-thru service cases of Namsan Library, summarized in Table 1, it can be seen that the loan period was extended from two weeks to six weeks a month while maintaining the existing service system. In addition, although existing members can use it, it is also possible to confirm that they tried to minimize face-to-face use while not allowing new members to sign up or reissue due to loss of ID cards. Therefore, the drive-thru service of the Namsan Library has had a positive effect in that the citizens can continue to read and spend free time even while the *force majeure* called COVID-19 can be expected. However, in order to use the service, it is necessary to make a reservation in advance on the Internet homepage, so there is a limitation that it cannot be used for people who do not have a computer or do not know how to use the Internet.

Table 1. Drive-thru service case of Namsan Library

Service appellation	Weekly visit reservation loan
Purpose	The purpose is to help citizens in the COVID-19 situation regain the will to overcome themselves with high quality books.
Period	Monday, March 9, 2020 until reopening after the serious phase of COVID-19.
Time	Monday-Friday 13:00-17:00 (Friday-Sunday reservations will receive Mondays)
Promotion policy	a) Existing members can only make weekly visit reservations. b) You can apply for a weekly visit reservation loan through the Namsan Library website. c) The number and duration of loans are in accordance with the regulations on reading and lending data at Namsan Library (7 books, 4 non-books, 6 weeks of loan period). d) Only books booked the day before can be borrowed, and if the book is not borrowed by the day after the reservation, it will be returned. In addition, if you do not borrow more than 3 times after booking, use is restricted for 30 days. e) In the case of weekly visit reservation loan operation, information materials, reading culture promotion and staff are involved.
How to apply	Search for library website data → Click on the material to be booked → Apply for a weekly reservation loan service (user) → Separately provide user application data (library) → Receive the ID of the person in charge on the following day (user)
Return	Return to library drop box or staff

 https://doi.org/10.1633/JISTaP.2021.9.2.3

4.2. Case Analysis 2: Incheon Metropolitan City

Incheon Metropolitan City has a total of 53 public libraries as of 2019. Among them, drive-thru service cases are planned by the Michuhol Library, which represents 44 libraries belonging to the local government, and are divided into five branches (Cheongna Lake Library, Cheongna International Library, Yeongjong Sky Library, Majeon Library, and Songdo International Organization Library). We focused on the contents carried out together. Table 2 summarizes some contents of the drive-thru service promotion plan planned by the Michuhol Library and implemented with five branches.

Looking at the cases of Michuhol Library and drive-thru service of the five branches, summarized in Table 2, there is a feature that utilizes the smart library located in the Cheongnaru parking lot of Cheongna Lake Park. The smart library is an unmanned library system where anyone can conveniently borrow and return with a device such as a vending machine. Since about 400 books can be stored in the smart library, pre-booked books are put in a designated field, and the latest books or best-selling books are placed in other fields. The strategy that made it possible to borrow or return is outstanding. In addition, the part that made the reservation application available on the library homepage and the mobile app is considered to be the consideration that made it easier for more users to use the drive-thru service. However, the total number of loans including the drive-thru service is five, and the period of the loan is 14 days, which is an unfortunate part considering the current social situation.

4.3. Case Analysis 3: Anyang City of Gyeonggi Province

As of 2019, in Anyang City of Gyeonggi Province, there are a total of ten public libraries belonging to local governments, and all libraries provide drive-thru services. Table 3 summarizes the main contents of the drive-thru service commonly provided by Anyang City Library.

If one looks at the case of the drive-thru service of Anyang City Library, which is summarized in Table 3, there is a feature that users must use e-mail, not a homepage or mobile app, to apply for reservation. Therefore, the process of downloading and filling out the form and sending it back to the e-mail can be cumbersome. In addition, since it is not operated in an integrated manner, it is possible to apply to each library, so one can use the nearest location, but the library may have different materials, so there may be limitations on use. However, there is also a difference that the width of time that can be received is the widest.

As mentioned above, as a result of looking at drive-thru service cases conducted by Namsan Library of Seoul Metropolitan Office of Education, Michuhol Library of Incheon Metropolitan City and its five branches, and Anyang City Library of Gyeonggi province, the service is currently being implemented alternatively due to COVID-19. It was found that the system has a level that can be applied at any time. However, it was also concluded that in order to provide more effective services, it is necessary to find detailed plans for each region or library.

Table 2. Drive-thru service cases of Michuhol Library and 5 branches

Service appellation	Cheongna Book Drive
Period	From Wednesday, April 8, 2020 until re-opening after the serious phase of COVID-19
Time	Monday-Friday 9:00-18:00 (Friday application books will be received on Monday; no application on weekends)
Promotion policy	a) Existing members can use Cheongna Book Drive service. b) To apply for the Cheongna Book Drive service, you can search for materials on the library's website and mobile app, and then apply for a reservation if it is available. c) The target data is limited to the data provided in the general data room and the children's data room. d) The number of reservations is 2 per person and is possible when the total number of loans is within the range of 5 books for 14 days. e) Only books booked on the previous day can be borrowed, and after the reservation is applied, check the identity of the member at the rental location from 13:00 to 17:00 the next day, and receive it. f) Books not received until the next day after reservation are automatically canceled, and rebooking for the same book is possible after one week.
How to apply	Search for materials on the library homepage/mobile app → Check for available loans → Apply for the Cheongna Book Drive service (user) → Submit the user application data to the smart library (library) → Receive the ID the next day (user)
Return	a) Books borrowed through the Cheongna Book Drive service must be returned through Book Drive. b) Smart Library → Return menu → Book barcode recognition → Book input → Return completed

Table 3. Drive-thru service cases of Anyang City Library

Service appellation	Safe book lending service
Period	Tuesday, March 17, 2020 until re-opening after the serious phase of COVID-19
Time	Daily 9:00-18:00 (Excluding holidays)
Promotion policy	a) Existing members can use the secured book lending service. b) The number of applications is 5 per person. c) The collection of application books is available from 14:00 to 20:00 the next day, from 14:00 to 19:00 for children's libraries, and from 14:00 to 17:00 on weekends.
How to apply	a) You can download an application from the homepage and fill out the secured book loan service and send it to the library email where you want to receive the service. b) If you receive a loanable letter from the library after applying for an email, visit the library with an ID card.
Return	Return to library drop box or staff

Table 4. Main participants

Participant	Affiliation library	Library career experience
A	Namsan Library	5 years
B		2 years
C		5 years
D		2 years
E	Michuhol Library	26 years
F		11 years
G		13 years
H		2 years
I	Anyang City Library	18 years
J		17 years
K		17 years
L		10 years

5. FOCUS GROUP INTERVIEW ANALYSIS

FGI (Focus Group Interviewing) is a research method that targets a small number of groups on a specific topic. It is a method of qualitatively grasping each person's perception, meaning, and content by experts participating in the field under the presidency of a modulator. Therefore, it has the advantage of being able to quickly and inexpensively explore the core of the problem and solve it flexibly. FGI was conducted according to the following procedure and results were analyzed.

5.1. Participants

Participants of the FGI are the three library managers introduced in the drive-thru service operation example, the main of which is shown in Table 4 (see participants marked A to L – for experience/classification purposes).

5.2. Write a Question and Send an Email

Prior to the interview, the researcher composed a question appropriate to the purpose of the study and e-mailed the participants. All questions were composed of open-ended questions that could induce participants to experience running a library drive-thru service, taking into account that the time required to conduct the entire interview was one hour. The questionnaire formed by the researcher and sent to the participants in advance was modified to the questioning method suggested by Krueger and Casey (2009), and was selected as four types of starting question, introduction question, main question, and ending question. The contents are shown in Table 5.

The content obtained through the step-by-step focus group interview questions summarized in Table 5 is as follows: First, it tried to confirm the librarian's perception of the drive-thru service. Second, it tried to reveal the relationship between the drive-thru service and the library applied to the special situation of COVID-19. Third, it tried to confirm the possibility of whether a drive-thru can be operated and activated as a regular service of the library even after the pandemic is over.

5.3. Conduct an Interview

The interview was conducted in real time utilizing ZOOM from 18:00 to 19:00 on Friday, March 12, 2021. The researchers briefly explained the purpose of the interview and the study, pledged that the data collected through the interview would not be used for any purpose other than the study, and focused on making possible participants feel free to discuss. The questions were tuned to finish within one hour while taking full advantage of one hour, focusing on the six that were prepared and distrib-

 https://doi.org/10.1633/JISTaP.2021.9.2.3

Table 5. Focus group interview questions

Question process	Item	Apply	Question
Opening questions	Drive-thru service awareness	We started with a simple greeting between researchers and participants and relatively simple questions related to the topic of the interview.	What was your first thought when you heard of drive-thru service?
Introductory questions	Drive-thru service and library relevance	Introduced the topic of the interview and allowed them to express their thoughts freely.	Drive-thru service and library, how can I connect?
Key questions	Drive-thru service operation	The library was able to fully express the experience of operating the drive-thru service.	Why did the library run the drive-thru service? What was the most difficult thing about running a drive-thru service in your library? How about always running the drive-thru service in the library?
Closing questions	Activation of drive-thru service operation	Summarize what you have said about your experiences and see if there are any more to add. Subsequently, the plan to promote the operation and increase the effectiveness of library drive-thru services will be divided.	What do you think is necessary to revitalize the library drive-thru service operation and increase its effectiveness?

Table 6. Analysis result for answering opening question

Participant	Responses/Answers	Category	Subject
A	"I thought of stores like McDonald's and Starbucks first"	Recognized as a service mainly applied to fast food stores	A new service method introduced by the library in accordance with the changing environment
F	"It was reminiscent of services such as the drive-through of the hamburger franchise"		
H	"I thought it was a convenient way to use mainly in cafes and fast-food restaurants"		
I	"I first thought of fast-food stores"		
B	"People can read books anytime, anywhere"	New library service for citizens' reading in accordance with the COVID-19 restrictions	
C	"I think it is a different and good service when closed for a long time due to coronavirus"		
D	"Increased user convenience, good news for those who like books and frequently visit the library"		
E	"I thought I had a job to do in the library"		
G	"Enhancing the convenience of citizens' use of data"		
J	"Borrowing books non-face-to-face outside the library without getting out of the car"		
K	"Borrowing books non-face-to-face"		
L	"There have been many changes due to COVID-19 and it has affected public library services as well"		

uted according to the four-step process. Also, all contents were recorded with the consent of the participants.

5.4. Result Analysis

Krueger and Casey (2009) describe important elements in the analysis of focus group interview data as "systematic, verifiable, sequential, and continuous." The data obtained

through the focus group interview were analyzed in the form of finding important topics and categories by applying open coding among the theory of evidence presented by Morgan and Scannell (1998). In other words, the concept was to name concepts and categorize concepts with similar contents.

Table 6 summarizes the results of analyzing the answer to the opening questions, "What was the first thought that came to your mind when you heard the word drive-thru service?"

Reviewing the answers to the introduction questions summarized in Table 6, librarians are using drive-thru services similar to the method utilized by fast-food stores. This is partly due to the COVID-19 situation and restrictions – libraries have introduced and implemented this new borrowing method for the benefit of the reading public.

Table 7 summarizes the results of analyzing the answers to the introductory question, "How can I connect the drive-thru service and the library?"

In response to the answers with the introductory questions summarized in Table 7, librarians can reduce contact with citizens while maintaining minimal available library services, such as being able to collect books that have been reserved in advance in such a way that connects the library directly with the drive-thru service.

Table 8 summarizes the results of analyzing the answer to the first key question, "Why did the library operate the drive-thru service?"

Looking at the responses to the first key questions listed in Table 8, the reason librarians are operating drive-thru services in libraries are due, in a large part, to the parent agency's COVID-19 guidelines to develop and operate non-face-to-face services.

Table 9 summarizes the results of analyzing the answer to the second key question, "What was the most difficult thing about running the drive-thru service in the library?"

Looking at the answers to the second major question summarized in Table 9, librarians cited an incomplete operating system, such as a lack of manpower and system errors, as the most difficult problems in operating the drive-

Table 7. Analysis result for answering introductory questions

Participant	Responses/Answers	Category	Subject
A	"I think you can make a reservation through the website or mobile app, and then check the ID at the information desk the next day to visit"	Connect to pick up pre-reserved books	Library service practice non-face-to-face
E	"Books reserved through the homepage or mobile app can be borrowed through the smart library without meeting the librarian"		
I	"I think you can make a reservation through the website or mobile app, and then check the ID at the information desk the next day to visit"		
B	"In the non-face-to-face era, I was contemplating what role the library could play"	Pathways to minimize contact with citizens and link library services	
C	"I think it would be good to actively serve users who are uncomfortable to come to the library"		
D	"Third-party services sought by the library when direct service is not possible"		
F	"As a representative non-face-to-face service, it is supported 24 hours a day, so I think it can even play the role of an annex"		
G	"Increase interest and utilization of libraries"		
H	"Pathways to move library services to the outside"		
J	"Provides services that can satisfy citizens' reading needs while minimizing contact when the library is not open due to coronavirus"		
K	"Provide services while minimizing contact between the library and citizens"		
L	"I think it is connected because it has already been implemented"		

 https://doi.org/10.1633/JISTaP.2021.9.2.3

Table 8. Analysis result for answering first key question

Participant	Responses/Answers	Category	Subject
A	"Due to the COVID-19 impact, guidelines were issued from the Lifelong Education Division of the Seoul Metropolitan Office of Education, and the drive-thru service was operated"	Operational guidelines are given by the parent organization through the meeting.	Guidelines for operating non-face-to-face services due to the COVID-19 situations
E	"It was decided through a meeting at Michuhol Library, the representative library of Incheon City Library, so the drive-thru service was operated"		
I	"It was decided through a meeting of Anyang City Library, so the drive-thru service was operated"		
B	"COVID-19"	Due to COVID-19 guidelines and restrictions	
C	"This service will be provided to users due to the library closure due to coronavirus"		
D	"Looking for alternative services due to library closing due to COVID-19"		
H	"For users who cannot visit the library due to COVID-19"		
J	"Minimize contact when the library is not open due to coronavirus and provide loan services to citizens"		
K	"Civil complaints from citizens who cannot use library services because of COVID-19 restrictions"		
L	"This is a method to provide loan services to citizens in situations where it is difficult to open the library due to COVID-19"		
F	"To develop library services without restrictions and to establish a system for common use of library materials"	Development of non-face-to-face library services	
G	"Development of library's active user service and non-face-to-face service"		

thru library service.

Table 10 summarizes the results of analyzing the answer to the third key question, "Why doesn't the library operate the drive-thru service at all times?"

Reviewing the responses to the third major question summarized in Table 10, librarians can provide a good service if a well-structured drive-thru service is always operating in the library. Increased manpower and higher budget are a prerequisite and required, though.

Table 11 summarizes the results of analyzing the answer to the closing question, "What do you think is necessary for the activation and effectiveness of library drive-thru service operation?"

Looking at the answers to the final questions summarized in Table 11, librarians believe that in order to initiate or revitalize drive-thru service operations in the library and increase its effectiveness, it is necessary to further support professional manpower requirements and accompanying budgets, establish a sound system, and promote the resource publicly.

Table 6 to Table 11 summarizes the questioning method suggested by Krueger and Casey (2009), and sets them into four types: opening questions, introductory questions, key questions, and closing questions. The answers were summarized using the analysis method suggested by Morgan and Scannell (1998). As a result, 6 themes, 16 categories, and 72 subjects were derived for the library drive-thru service experience. Among these, 6 themes which can be considered the core results of this study are summarized in Table 12.

In the end, public library librarians thought that the drive-thru service was a newly introduced service method according to the non-face-to-face service operating guidelines of the COVID-19 situation, thus is inefficient and unnecessary at present as it was difficult to use due to its incomplete operating system. However, if the system were well established, such as building a service system, it is expected that it would be a good service.

Table 9. Analysis result for answering second key question

Participant	Responses/Answers	Category	Subject
A	"After preparing the reserved book(s), applicants have to wait for a specified time because they do not know when they will come to pick up the book"	It was difficult due to the preparation process and user involvement	It was difficult because of the incomplete operating system
B	"Feeling of wasting time because of people who do not visit after booking, and the system had to be changed due to constantly changing policies and circumstances"		
C	"It was cumbersome because users had to buy the books they wanted, and there were users who came without making a reservation"		
D	"Users do not visit after making a reservation or request other services"		
E	"After receiving the reservation, it was difficult to repeatedly find the book, disinfect it, and put it in the smart library"		
I	"The service preparation process was also difficult, but the point of waiting outdoors was difficult due to hot or cold weather extremes"		
J	"It took a lot of time to find and prepare books reserved through e-mail"		
K	"It was difficult to find and disinfect the books and make them into separate packages"		
L	"It was difficult because there were more books to borrow than other libraries"		
F	"Equipment error and lack of manpower"	Insufficient manpower and equipment errors	
G	"Difficulty of quick resolution in case of insufficient manpower or equipment error"		
H	"System error"		

6. DISCUSSION

The purpose of this study is to analyze the operation cases of the 'drive-thru' service conducted by public libraries in Korea under the influence of COVID-19, and to find ways to settle this service in the future. Therefore, because the library environment and related policies are different in each country, it is difficult to expand and apply to all countries. Therefore, we hope that discussions can be started from this point to determine whether and how the drive-thru service will operate in the library.

Korea has the lowest number of public libraries among OECD countries. Also, because the total population is larger than the land area, the total area of public libraries, including parking lots, is also narrow. Therefore, there is no public library in place which all citizens can reach within five minutes on foot, and there are many places where there is no parking if users drive a car. Therefore, considering these conditions, the drive-thru service can be an efficient way to increase the library utilization rate of users.

However, in order to operate the drive-thru service in the library at all times, there are the following problems that must be solved first. First, policies should be established. Since the library is a public institution where government or local governments perform official duties, policies should be established in order to proceed in the future. Second, the budget should be secured. Since budgets are a necessary condition to promote everything, libraries need to secure an appropriate budget. Third, it is necessary to recruit manpower. Currently, the number of librarians is insufficient in public libraries in Korea. Therefore, it is difficult to provide efficient services unless dedicated personnel are recruited. Fourth, the reservation system through mobile apps should be expanded. Since modern people use smartphones more than Inter-

 https://doi.org/10.1633/JISTaP.2021.9.2.3

Table 10. Analysis result for answering third key question

Participant	Responses/Answers	Category	Subject
A	"It will be difficult, but if the system is well established, it will be a service for busy users"	It is a necessary service, so it should be organized well	It can be a good service if the system is well established, but at the moment it has been shown to be inefficient and unnecessary
E	"I think this is a necessary service if you consider the library conditions with insufficient parking space, the increase of smart libraries, and the lives of busy modern people"		
G	"It would be okay to designate and operate a specific library, such as a library specialized for each subject"		
I	"There is a need to first find a plan for efficient operation, such as through a smart library or installing a dedicated booth in the library parking lot. If a plan is established first, it will be a good service"		
B	"There are already services available, such as reservation loans and desired books, and it does not seem to suit the purpose of a library that prioritizes visitors"	Feeling inefficient or unnecessary	
C	"Appropriate in unavoidable cases such as closing due to corona, but thought to be unnecessary during opening"		
D	"When the library is open, it is a redundant service and is considered unnecessary because it is a burden in terms of personnel"		
F	"It would be possible if manpower and budget were supported, but I think the regular operation is inefficient"		
H	"I think it is not good in terms of efficiency to run the library normally concurrent with drive through service"		
J	"We oppose parcel delivery services for the socially disadvantaged because it consumes a lot of time and manpower"		
K	"Opposition because it consumes a lot of time and manpower"		
L	"The service is difficult to use for citizens who have difficulty making online reservations, and the number of books should be higher."		

net homepages, access to mobile apps is required rather than Internet homepages to use drive-thru services more efficiently. Fifth, active promotion is necessary. It is necessary to make drive-thru services available to many citizens through various publicity activities such as Internet homepages and social network service. And, finally, it is necessary to prepare appropriate regulatory measures for users who do not use the service after making a reservation.

7. CONCLUSIONS AND SUGGESTIONS

The purpose of this study is to analyze the case of drive-thru services newly introduced and tried in libraries in Korea under the influence of COVID-19, and to develop and propose a service model so that this service can be continuously applied to all libraries in the future. Therefore, in order to achieve the research purpose, we selected one of the representative libraries that are conducting related services in Seoul Special City, Incheon Metropolitan City, and Anyang City in Gyeonggi Province, and looked at the operation cases. Focus group interviews were also conducted through real time video conference calls with twelve service representatives in each library to draw the following conclusions.

First, the library's drive-thru service has a positive effect in that citizens can continue to read and spend free time even in the irresistible situation of COVID-19, and

Table 11. Analysis result for answering closing question

Participant	Responses/Answers	Category	Subject
A	"Professional personnel are needed to handle the service"	Needs professional manpower and budget support	Need to build a service system
D	"Allocate dedicated personnel, allocate budget, require stipulated regulations, prepare measures to boost employee morale"		
F	"Efficient operation is possible if a dedicated team is established"		
G	"Expansion of dedicated manpower and budget"		
I	"Needs support for manpower and budget"		
L	"Recruitment of manpower, securing space for vehicle access, and installation of loan devices should be preceded"		
B	"Codified operating regulations are required, and a system is built for perfect non-face-to-face"	It is necessary to prepare operational regulations and establish a system	
J	"A large space is also required for the computer system and the smooth flow of the vehicle"		
C	"It would be good if you actively use the smart library"	Smart library utilization	
E	"Since it is possible to connect with the mutual lending service system, it is necessary to install many smart libraries in public places used by many people"		
H	"Promotion of the service"	Need to promote	
K	"I hope that the library will be fully opened as before and the situation where citizens can freely use it soon"	We look forward to opening the library entirely	

Table 12. Subjects derived according to the experience of the drive-thru service

Questions	Derived subjects
Opening questions	A new service method introduced by the library in accordance with the changing environment
Introductory questions	Library service practice non-face-to-face
First key questions	Guidelines for operating non-face-to-face services due to the COVID-19 situations
Second key questions	The difficulties incurred by the incomplete operating system
Third key questions	It can be a good service if the system is well established, but at the moment it has been shown to be inefficient and unnecessary
Closing questions	The need of building a service system

also instilled courage to overcome difficulties on their own. However, when service reservation was limited to an Internet homepage, there were also limitations in that it would not be available to those who do not have a computer or do not know how to use the Internet.

Second, it was concluded that it was difficult to apply the library's drive-thru service in a situation where issues such as policy establishment, budget securing, and selection of dedicated personnel were not resolved. Therefore, it is necessary to set the specific application method according to the library after solving the problems first, and it is also necessary to prepare alternatives such as using inter-loan services with the smart library before solving the problem.

Services are convenient for recipients but can be a difficult process for those who provide them. As a result, the service that was being carried out is sometimes stopped, but if the people who have already received the service

Table 13. Drive-thru service operation guidelines in the library

Designation	Drive-thru service
Explanation	A service that allows you to borrow pre-booked materials without getting out of the car
How to make a reservation	1) Check if you are a library member. If you are not a member, you can visit the library and create an ID 2) Book the desired material through the library homepage or mobile app. Reservations are available 24 hours a day 3) If you are aged 65 or over, you can make a reservation by phone from 09:00-17:00
How to pick up	1) You must bring your library ID card and visit the booker yourself 2) If the day after the reservation is ended, it can be picked up after 2 days 3) Materials booked by 15:00 can be picked up within the library operating hours the next day 4) Materials booked after 15:00 can be picked up after 2 days
How to return	1) ID card is not required to return, and someone else can return 2) Return directly to the library 3) Library drop box 4) Return materials to the device when using the smart library
Other	1) The number of loans, loss etc. are in accordance with the library operation regulations 2) Data not received within the designated time will be automatically returned 3) Users who do not receive the data after making a reservation more than 3 times are prohibited from using the service for a month afterwards

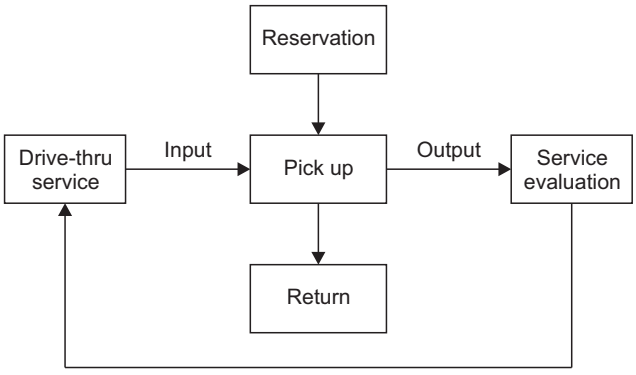


Fig. 1. Drive-thru service model.

experience convenience, the demand for resumption may continue.

In this regard, the drive-thru service in Korea was first applied, which brought difficulties to the library, but it must have been a way to make users experience convenience. Therefore, even if the library is reopened, it is a service that can cause demand for implementation. Therefore, the following are suggestions on how to efficiently apply the drive-thru service in the library.

First, even in public libraries belonging to the same local government, the conditions are different. Therefore, it is necessary to designate and operate one library that is most suitable for the drive-thru service, or actively utilize the smart library. If libraries specify one place in terms of suitability, they will need additional support, such as assigning dedicated manpower or installing a dedicated booth.

Second, reservation of the drive-thru service is possible only through the homepage and mobile app, but it is necessary to allow elderly users to make voice calls.

Third, the operation of the drive-thru service in the library allows users to visit pre-booked materials by visiting the library during the next day. For other loan volumes or due dates, existing operating guidelines should be followed.

Table 13 is the drive-thru service operation guideline established, based on the contents of Table 1, Table 2, and Table 3. It includes the details of the service use methods such as the name, description (definition) of service, reservation method, reception method and return method, and others.

Fig. 1 is the drive-thru service model. Fig. 1 expresses that the results of users' use and evaluation will be calculated when the drive-thru service is prepared and implemented in the library, and the contents should be reflected in the service again. At this time, input factors include dedicated staff, operating budgets, and service space such as a smart library and homepage reservation systems. Calculation factors include the number of reservations made, number of visitors, the quantity of loans, and various satisfaction factors.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Abu Farha, R., Abu Hammour, K., Alefishat, E., Alsaeed, H., & Alma'aiah, S. (2017). Drive-thru pharmacy service: Assessments of awareness, perception and barriers among pharmacists in Jordan. *Saudi Pharmaceutical Journal*, 25(8), 1231-1236. https://doi.org/10.1016/j.jsps.2017.09.008.

Allon, G., Federgruen, A., & Pierson, M. (2011). How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management*, 13(4), 489-507. https://doi.org/10.1287/msom.1110.0343.

Dalin-Kaptzan, Z. (2020). *Retail trends: The growing demand for click and collect*. https://www.bringg.com/blog/industry-trends/click-and-collect-growth/.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting & Social Change*, 114, 254-280. https://doi.org/10.1016/j.techfore.2016.08.019.

IFLA. (1994). *IFLA/UNESCO public library manifesto 1994*. https://www.ifla.org/publications/ifla-unesco-public-library-manifesto-1994.

IFLA. (2021). *COVID-19 and the global library field*. https://www.ifla.org/covid-19-and-libraries.

Koontz, C., & Gubbin, B. (2010). *IFLA public library service guidelines*. https://www.ifla.org/publications/ifla-publications-series-147.

Krueger, R. A., & Casey, M. A. (2009). *Focus groups: A practical guide for applied research*. 4th ed. Sage.

Kucukusta, D., Law, R., Besbes, A., & Legohérel, P. (2015). Re-examining perceived usefulness and ease of use in online booking: The case of Hong Kong online users. *International Journal of Contemporary Hospitality Management*, 27(2), 185-198. https://doi.org/10.1108/IJCHM-09-2013-0413.

Montero, S. (2021). *Westlink Library adds a new "drive-thru" feature*. https://www.ksn.com/news/local/westlink-library-adds-a-new-drive-thru-feature.

Morcos, S. H., Tak, J., & Gregoire, M. B. (1992). Customer perception of drive-thru service. *Hospitality Review*, 10(2), Article 3. https://digitalcommons.fiu.edu/hospitalityreview/vol10/iss2/3.

Morgan, D. L., & Scannell, A. U. (1998). *Planning focus groups*. Sage.

Public Library Association. (2020). *Public libraries respond to COVID-19: Survey of response & activities*. American Library Association.

Song, M. H. (2002). A proposal of multimedia menu system for drive-thru. *Journal of Basic Design & Art*, 3(2), 87-98. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001557078.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(2): 47-65, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.2.4

Topic Analysis of Scholarly Communication Research

Hyun Ji
Library and Information Science, Ewha Womans University, Seoul, Korea
E-mail: Jhyn641@gmail.com

Mikyeong Cha*
Library and Information Science, Ewha Womans University, Seoul, Korea
E-mail: cha@ewha.ac.kr

ABSTRACT

This study aims to identify specific topics, trends, and structural characteristics of scholarly communication research, based on 1,435 articles published from 1970 to 2018 in the Scopus database through Latent Dirichlet Allocation topic modeling, serial analysis, and network analysis. Topic modeling, time series analysis, and network analysis were used to analyze specific topics, trends, and structures, respectively. The results were summarized into three sets as follows. First, the specific topics of scholarly communication research were nineteen in number, including research resource management and research data, and their research proportion is even. Second, as a result of the time series analysis, there are three upward trending topics: *Topic 6: Open Access Publishing, Topic 7: Green Open Access, Topic 19: Informal Communication,* and two downward trending topics: *Topic 11: Researcher Network* and *Topic 12: Electronic Journal*. Third, the network analysis results indicated that high mean profile association topics were related to the institution, and topics with high triangle betweenness centrality, such as *Topic 14: Research Resource Management*, shared the citation context. Also, through cluster analysis using parallel nearest neighbor clustering, six clusters connected with different concepts were identified.

Keywords: scholarly communication, topic modeling, network analysis, topic analysis, text analysis, bibliometric analysis

Received: November 24, 2020 **Revised:** April 24, 2021
Accepted: April 28, 2021 **Published:** June 30, 2021

***Corresponding Author:** Mikyeong Cha
https://orcid.org/0000-0002-7979-6843
E-mail: cha@ewha.ac.kr

This paper is based on data collected during the preparation of the master's thesis of the first author, and completely expanded it.

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© Hyun Ji, Mikyeong Cha, 2021

1. INTRODUCTION

1.1. Background

Scholarly communication has always been considered a fundamental aspect of scientific research. The function could be found in scholarly communication, suggested De Solla Price (1963). He defined it as the connection among scholars, which increases one scholar's awareness of the work by sharing ideas. It can be informal or formal, and more various tools are used for communication within the scholarship. By connecting scholars, scholarly communication has made scholars share and advance their ideas (De Solla Price, 1963; Klain-Gabbay & Shoham, 2018).

The advance of scholarly communication is required to function well in a modern society where technology development is accelerated. It is the reason why research on scholarly communication is essential. At this point, analyzing accumulated studies and reifying the subject area could offer logical information about the scholarly communication to future researchers. The objective of the study is identifying the topics of scholarly communication. More specifically, the study aims to answer the following questions:

- What are the research topics studied in the scholarly communication area?
- What is the trend of scholarly communication research?
- What is the structural character of scholarly communication research?

Three methodologies were used in this study. Latent Dirichlet Allocation (LDA) topic modeling was performed on abstracts of 1,435 articles published from 1970 to 2018 in the Scopus database to find the specific topics. Time series analysis was used to find out the trends of topics. Network analysis was conducted to identify the structural characteristics of scholarly communication research. The results of this study serve as primary data for future scholarly communication research.

1.2. Literature Review

Previous studies, conducted with topic modeling to figure out the research topics, were referenced to construct the proper research processes. The studies were summarized as follows.

Park and Oh (2017) collected 1,027 papers published in two journals related to Korean records management and four journals related to library information science from 1997 to 2016. Then, they conducted LDA topic modeling and Hierarchical Dirichlet Process topic modeling. As a result, they identify ten critical topics, such as 'electronic records,' 'National Archives,' and 'record information service.'

Jin and Song (2016) analyzed the titles and abstracts of the articles published in the top 20 journals of which impact factor was high for the past five years, according to the Information & Library Science subject in Journal Citation Reports (JCR) 2013. Through topic modeling, they identified 50 topics. Networks were created with these topics to classify interdisciplinarity using the average path length of the topic network. After that, they compared the topic networks among academic journals. Through this, it was identified that the text-based index is different from the citation information-based index.

Park and Song (2013) crawled web pages of the journals in library and information science (LIS) in Korea. As a result, 3,834 English abstracts were collected, dating from 1970 to 2012. Then they performed topic modeling. Through this, the topics of LIS were identified. Also, time series analysis was performed on them to identify hot topics and cold topics. Besides this, topic modeling was conducted for each journal to compare and analyze topics for each journal. Research to understand the research trend using topic modeling in other subject fields was also conducted.

Keshner et al. (2019) used topic modeling to analyze how the study of virtual reality (VR) to rehabilitation has been conducted. As a result, they found that the extracted topics of VR rehabilitation are distinguished from general science or engineering fields. In particular, the research areas of VR rehabilitation contained unique topics, such as telerehabilitation.

Amado et al. (2018) conducted a semi-automated text mining study to identify significant big data trends in the marketing field. Using Science Direct, 1,560 published papers from 2010 to 2015 related to big data in marketing were collected. The study found that the topics were categorized into the latest technology and the advantages of big data in marketing.

As a result of integrating these studies, three points were clarified. First, topic modeling was used to find the topics related not only to specific fields but also particular keywords. Second, the data for research could be collected from both journals and databases. Third, network analysis and time series analysis were frequently used to analyze the topics identified from the topic modeling. Based on the three points inferred from previous studies, LDA topic

 https://doi.org/10.1633/JISTaP.2021.9.2.4

modeling was performed on the articles extracted from the database, and time series analysis and network analysis were performed based on those topics.

2. TOPIC ANALYSIS

2.1. Topic Modeling

A topic modeling algorithm is a statistical method that analyzes original texts to discover the themes that run through them. It assumes the literature as a mixture of various topics, and the topic is expressed as a distribution of several words. According to the topic modeling, when composing a document it is repeated in order to select one of the topics and then select a word among the chosen topic based on the proportion of the topics and words (Blei, 2012).

However, in a realistic situation, the distribution of topics in the articles and the distribution of words in the topic were sealed. Therefore, it is necessary to observe the word distribution of the existing literature and then estimate the topic's word distribution. This is the reason that LDA is proposed (Blei, 2012; Song, 2017; Weingart, 2012).

LDA is a statistical model of document collections, and this statistical model reflects the intuition that documents exhibit the topics. Each document shows the topics in different proportions. Each word in each document is drawn from one of the topics, where the selected topic is chosen from the pre-document distribution over topics (Blei et al., 2003; Blei, 2012).

To perform topic modeling, the researcher must set the number of topics. In this study, coherence was used to select the number of topics. Coherence is assessing the consistency of the words in the topic (Newman et al., 2010). In other words, coherence measures the similarity of words in each topic by calculating how often they appear in similar contexts. It is based on the linguistic assumption that words with similar meanings often appear in similar contexts (Syed & Spruit, 2017). When the topic modeling is performed more accurately, the coherence is calculated more highly. Therefore, after repetition to change the number of the topics and see the value of coherence, the number of topics with the highest value of coherence was selected (Choi & Ko, 2019).

2.2. Network Analysis

Networks model various systems in the real world by expressing a person or an object as a node and the relationship between the nodes as a link. The network model could be used to analyze and understand various characteristics scientifically (Lee, 2012).

In LIS, research related to network analysis has been primarily carried out in two categories: the analysis of information users and study about the area of bibliographical science. The starting point of the two fields is the study of the "invisible college." The invisible college means the interaction of scholars apart from each other for productive information exchange. Since then, with the advance of bibliometric analysis techniques, including co-citation analysis, network analysis has extended to the access and links through the publications. Also, bibliographical science has been interested in grasping the connection between documents with network analysis (Lee, 2006c; Zuccala, 2006).

Network analysis performs at various levels. In this research, network-level analysis and node-level analysis were performed. Besides this, centrality analysis was conducted to grasp the relational character of particular topics. Centrality is a concept that indicates how much closer a node is to the center than others. The centrality calculated for the node is not absolute value but relative value in a network. Since its introduction in the 1950s, various methods to measure centrality have been developed. It is necessary to perform centrality analysis by selecting an appropriate indicator among these various centrality indicators, depending on the network's type and structural characteristics (Lee, 2012). In this study, the nearest neighbor centrality (NNC), centrality of mean profile association (Cmp), and triangle betweenness centrality (TBC) proposed by Lee (2006c) were used to prevent data loss during the conversion process.

The pathfinder network (PFNet) was used to describe the topic network in this study. The PFNet, which is one way to express a network, is a network created by removing an edge that violates a triangle inequality. The violation of triangular inequality means a direct edge is longer than an indirect path connected through several short edges. The PFNet is widely used to visualize knowledge networks that describe keywords or concepts in a research field (Lee, 2006a; Lee, 2012; Schvaneveldt, 1990).

3. METHODOLOGY

3.1. Data Extraction

Since scholarly communication was discussed in various fields, data was collected from the database to search for articles from various subject areas. Among the databases, Scopus was selected because it contained the most articles related to scholarly communication. The retrieval term was "Scholarly Communication," according to the LC (Library

of Congress) subject heading. The retrieve was targeted on the title, abstract, keyword, and index search and conducted on the limited period from 1970 to 2018 when citation information was registered in Scopus for accurate time series analysis. As a result, authors, publication year, abstract, and author keywords data of the 1,552 articles were extracted. The abstracts were selected as data for text analysis because they contained more specific information than the author keywords did. Therefore, 96 articles without abstracts were deleted. Besides this, 21 documents not related to scholarly communication were deleted too. Finally, abstracts of 1,435 articles published from 1976 to 2018 were used to analyze the topics. The number of articles about scholarly communication, based on the extracted data, continues to increase as follows (Fig. 1).

The data for topic modeling could be various unstructured text data, and preprocessing is essential to analyze them more accurately (Feldman & Sanger, 2006). For this purpose, abbreviation unification was performed as the first step. For example, in the case of open access, 'OA' and 'open access' were used together. These terms were unified through changing abbreviations into full words. NLTK, an English natural language processing package, was used in other preprocessing steps. Tokenization and lemmatization were performed, and bigram and trigram were applied to reflect noun phrases in which two or more words were used as one word. After stemming, stop words such as articles were deleted. The words with a frequency of less than ten were removed. As a result of the preprocessing, a total of 6,920 words were extracted.

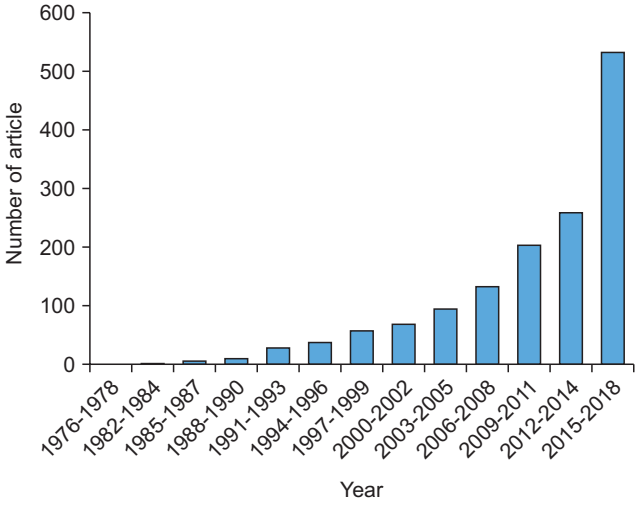


Fig. 1. Number of extracted articles.

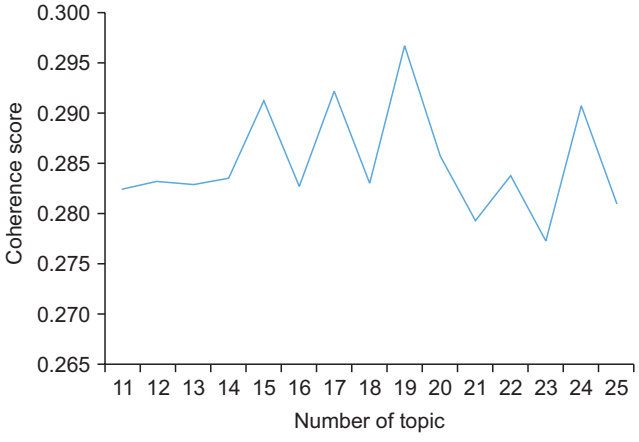


Fig. 2. Coherence by the number of topics.

3.2. Topic Modeling

LDA topic modeling was performed using GENSIM, a Python package, and a corpus and dictionary were constructed on preprocessed data. Then the corpus was written based on term frequency-inverse document frequency (TF-IDF). In the topic modeling, the researchers should decide the number of topics. The coherence depending on the number of topics between 11 and 25 was calculated to select the number of topics. When the number of topics

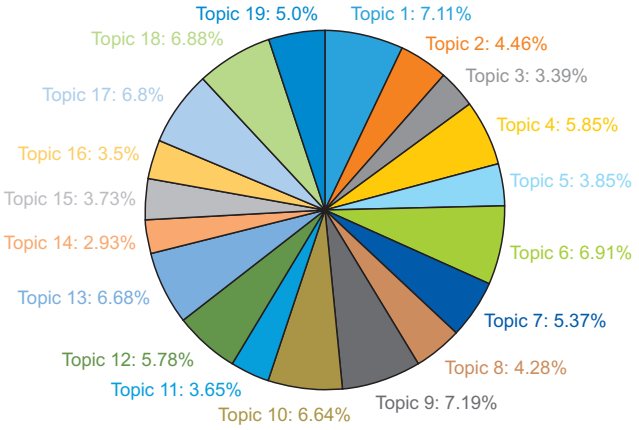


Fig. 3. Topic proportion. The figures were rounded off to second decimal places. Topic 1, Academic Publication; Topic 2, University Library Service (ULS); Topic 3, Technology Element; Topic 4, Copyright; Topic 5, Online Information Search (OIS); Topic 6, Open Access Publishing (OA Publishing); Topic 7, Green Open Access (Green OA); Topic 8, Research Cooperation; Topic 9, Academic Infrastructure (Academic Infra); Topic 10, Digital Communication; Topic 11, Researcher Network; Topic 12, Electronic Journal (E-journal); Topic 13, Preservation; Topic 14, Research Resource Management (RRM); Topic 15, Institute Policy; Topic 16, Institutional Repository (IR); Topic 17, Research Data; Topic 18, Research Evaluation; Topic 19, Informal Communication.

 https://doi.org/10.1633/JISTaP.2021.9.2.4

was nineteen, coherence was highest. As a result, nineteen was chosen as the number of topics (Fig. 2).

Seven representative words of the topics were extracted. The distribution of words in the topic and the ratio of the word's frequency in each topic to the word's total occurrence were considered to choose the words. The largest proportion of the topic was 7.11% (topic 1), and the smallest proportion of the topic was 2.93% (topic 14). The even distribution of the topics means researchers chose the topics evenly (Fig. 3).

3.3. Time Series Analysis

For the time series analysis, the published year of the articles was assigned to the topics based on the extracted document-topic distribution to determine the topic's appearance period. Then the distribution of the topic in each document was summed up for each year. Based on this, the time of the first appearance was identified for each subject.

Among the collected articles, 12 articles were published from 1976 to 1988, and there were no documents published from 1976 to 1984. The whole period was divided into thirteen sections for time series analysis. Also, in the period sections, each topic's appearance was divided by the sum of the all topics appearance and the value was used in because the number of articles increased over the years, affecting time serial analysis.

3.4. Network Analysis

The top 30 words in the topic (Appendix 1) and the term frequency (TF) were extracted, and the inverted topic frequency (ITF) was calculated, applying the concept of inverse document frequency (IDF). The topic × word matrix was calculated, and weight was given to words to identify particular topics using the ITF (Inverted Topic Frequency) value rather than the topic frequency (Jin & Song, 2016). The TfidfVectorizer module was used to make the matrix, the topic × word matrix was transformed to create the topic × topic (19*19) matrix, and the matrix was attached as in Appendix 2. Cosine similarity coefficients were calculated based on the topic × topic matrix, and the matrix was attached it as in Appendix 3. In this study, a PFNet was created using a cosine similarity matrix.

4. RESULT

4.1. The Result of Keyword Analysis

Before topic modeling, keywords analysis was conducted to understand scholarly communication research in an overview. The top 100 stemmed words based on the TF value and the TF-IDF value were visualized with the word cloud for more efficient analysis (Fig. 4).

'Electron*' and 'journal' have high TF and TF-IDF values. It means that both words were used frequently both in one article and over the whole of scholarly communication research. In other words, the research indirectly and directly related to journals and electric resources was conducted frequently. It could be inferred that journal and electronic resources were the most influential academic resources in scholarly communication research.

Words referring to institutions such as 'univers*', 'institut*', 'librari*', and 'associ*' were frequently used. We can infer that a large part of scholarly communication research is related to institutional research because those words referred to the institution deeply related to the scholarly communication. Besides this, 'librari*' and 'librarian' occupy a large proportion of both numbers. In the case of 'librari*', TF-IDF value is less than TF value, whereas in the case of 'librarian,' TF value is less than TF-IDF value. This

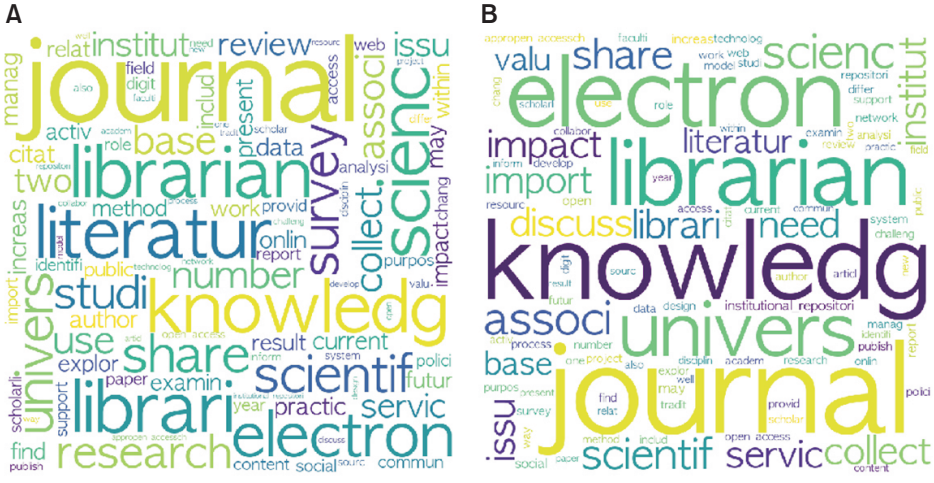


Fig. 4. 100 key words. (A) Term frequency. (B) Term frequency-inverse document frequency.

means that 'librari*' appears in more texts, while 'librarian' appears more frequently in a small number of texts. The research related to the librarian is distinguished from research related to the library. The research related to the librarian is more specifically focused on the librarian.

Many words with high TF value or TF-IDF value refer to academic research, such as 'research,' 'studi*,' 'scienc*,' and 'scientifi*.' Through this, we can confirm that scholarly communication research is being conducted under the definition. In the case of 'knowledg*' and 'journal,' both values were high, and words related to research resources such as 'literatur*' and 'collect*' were also appearing. Through this, we can figure out that discussions on research resources frequently occur in scholarly communication research.

Looking at the keywords, we observed that many kinds of research are about the institution, and among them, many studies are about the library. However, there is a limit to grasping the research topic only by analyzing keywords. It is difficult to identify what the keyword indicates. For example, 'electron*' could be interpreted to mean different topics depending on the word used together. If it is used with a subscription, it could indicate the topic related to the journal crisis; whereas, if it is used with management, it could be interpreted as digital management. Topic modeling was performed to find out more detailed topics in the next chapter to solve this ambiguity.

4.2. Result of Topic Modeling

As a result of LDA topic modeling, we found 19 topics and extracted representative words. Topics were interpreted and named based on the topic-word distribution to identify the significant words in the topics and the document-topic distribution to identify the document with a high proportion of the topic (Table 1).

The nineteen specific topics of scholarly communication research were identified, and researchers evenly selected every topic. The identified topics are subdivided, detailed, and concrete. For example, Open Access (OA) topics were separated into two topics depending on the strategy: *Topic 6: Open Access Publishing (OA Publishing)* and *Topic 7: Green Open Access (Green OA)*. This result would be useful when researchers want to specify their ideas and new researchers want to clarify what scholarly communication research is.

4.3. Result of Time Series Analysis

For the time series analysis of the topic identified through topic modeling, the article's publishing year was assigned to the topic based on the extracted document-topic distribution to determine the topic's appearance year. As the last part of preprocess for the analysis, the whole period was divided by three years to clarify overall trends of each topic and the distribution of each topic in the article was summed up. Overall, because the number of academic communication studies increased, all topics increased (Table 2).

4.3.1. First Appearance of Each Topic

The earliest topic was *Topic 18: Research Evaluation*, which appeared alone in 1976. After then, there was no literature collected until 1982. *Topic 3: Technology Element* first appeared in 1983. In 1984, *Topic 1: Academic Publication* and *Topic 2: University Library Service (ULS)* first appeared together. Subsequently, in 1985, *Topic 12: Electronic Journal (E-Journal)* first appeared, and then, in 1986, *Topic 6: OA Publishing* first appeared. In 1987, *Topic 9: Academic Infrastructure (Academic Infra), Topic 13: Preservation,* and *Topic 17: Research Data* first appeared. In 1988, *Topic 8: Research Cooperation* first appeared, and in 1989, other topics such as *Topic 4: Copyright* and *Topic 5: Online Information Search (OIS)* were started.

In 1963, Science Citation Index (SCI) was published. *Topic 18: Research Evaluation* appeared in the 1970s when the ranking of influence by journals was also published. Besides this, electronic journals started in the late 1970s (Lancaster, 1995), and in the 1980s, the serial pricing crisis was raised as a big problem due to the continuous increase in subscription fees for printed journals (European Commission, 2019). In this context, the discussions on *Topic 3: Technology Element, Topic 1: Academic Publication, Topic 2: ULS,* and *Topic 12: E-Journal* were followed sequentially. Next, *Topic 6: OA Publishing* and *Topic 7: Green OA*, which included discussions on solutions to them, appeared (Table 3).

The topics related to technology appear in the order of *Topic 3: Technology Element, Topic 5: OIS*, and *Topic 10: Digital Communication*. Starting with *Topic 3: Technology Element*, which is related to the impact of science and technology, the focus spread to *Topic 5: OIS*, which focuses on retrieving and searching for online information. Then, *Topic 10: Digital Communication* appeared, which focuses on the whole digital-based communication process. It means that the scholarly communication research topics related to technology became more specified with the advance of technology.

 https://doi.org/10.1633/JISTaP.2021.9.2.4

Table 1. 19 topics of scholarly communication research

1 Academic Publication	2 University Library Service (ULS)	3 Technology Element	4 Copyright
journal	librarian/library	journal	open_access
publish	faculty_member	electron	copyright
policy	service	scientist	electronic
sustain	resource	database	publish
research	profession	impact	embargo
open_access	study	technology	academy
predatory_publish	research/academic	effect	market
5 Online Information Search (OIS)	**6 Open Access Publishing (OA Publishing)**	**7 Green Open Access (Green OA)**	**8 Research Cooperation**
internet	open_access_journal	open_access	co_authorship
resource	open_access	library	orcid
directory	library	institutional_repository	web
frailty	citation	policy	association
web	article	manuscript	publish
manage	public	article	seek
access	publish	share	sparc
9 Academic Infrastructure (Academic Infra)	**10 Digital Communication**	**11 Researcher Network**	**12 Electronic Journal (E-journal)**
academic	research	network	electronic_journal
nation	information_literacy	communication	internet
role	digit	ocid	publish
model	rebut	field	measure
infrastructure	provide	peer_review	impact
patron	tool	subgroup	technology
public	link	reference	system
13 Preservation	**14 Research Resource Management (RRM)**	**15 Institute Policy**	**16 Institutional Repository (IR)**
archive	metadata	research	manage
model	journal	fiscal	institutional_repository
risk	crossref	administration	data
research	grey_literature	institutional_repository	preprint
system	paper	institute	share
librarian/archivist	policy	open_access	policy
digit	collection	mandate	role
17 Research Data	**18 Research Evaluation**	**19 Informal Communication**	
reuse	reward_system	tweet/tweeter	
data	journal	blog	
encoding	factor	change	
masur	evaluate	web	
research	model	collaboration	
collect	impact	site	
system	altmetric	distribution	

4.3.2. Upward Topics and Downward Topics

The amount of scholarly communication research has increased rapidly since 2014, which caused the increase of the topics' distribution. In consideration of this, each topic's distribution was calculated by dividing each topic's distribution by the sum of distribution in the year. By analyzing this, three upward topics and two downward topics were identified among the topics. The three upward top-

Table 2. Trend of topic distribution

Topic No.	Year													Total
	1976 -1978	1982 -1984	1985 -1987	1988 -1990	1991 -1993	1994 -1996	1997 -1999	2000 -2002	2003 -2005	2006 -2008	2009 -2011	2012 -2014	2015 -2018	
1		0.88		1.18	1.70	3.96	1.36	3.51	6.07	8.43	12.32	18.07	32.39	89.87
2		0.87		0.88	2.62	0.64	0.93	2.97	2.30	3.56	11.45	7.07	23.42	56.71
3		0.72		0.86	3.53	4.11	2.06	0.02	1.22	3.86	8.32	8.76	9.66	43.12
4				0.86	0.88	3.47	2.89	2.77	4.67	6.88	14.82	15.57	22.41	75.22
5				0.01	2.28	2.85	4.03	1.66	4.78	4.21	10.34	8.49	13.55	52.20
6			0.15	0.01	0.98	0.69	1.25	2.72	5.87	5.65	16.20	17.56	29.53	80.61
7				0.01	0.02	0.01	1.84	3.02	5.52	5.29	8.41	14.26	28.84	67.22
8				0.86	0.02	2.13	2.08	3.43	3.35	8.17	3.14	8.36	21.99	53.53
9			0.87	0.01	1.91	1.26	8.51	4.96	8.03	7.11	11.58	16.63	30.05	90.92
10				0.83	1.95	2.59	2.05	2.78	4.14	11.15	11.05	12.12	34.42	83.08
11				0.87	1.46	0.87	2.25	2.22	1.61	4.13	5.58	9.13	17.53	45.65
12			0.86	0.82	2.74	2.29	4.14	4.49	3.83	6.23	8.38	7.71	31.73	73.22
13			0.85	0.01	3.78	2.25	2.91	5.78	8.04	5.68	11.96	12.78	30.41	84.45
14				0.01	0.90	0.63	1.41	3.41	1.96	4.04	5.04	8.03	11.21	36.64
15				0.01	1.44	0.84	1.59	2.43	2.43	2.51	8.46	8.86	18.12	46.69
16				0.58	0.02	0.61	1.46	0.04	2.90	3.63	5.65	6.82	22.14	43.85
17			0.85	0.01	0.86	0.46	5.01	5.56	9.36	11.30	7.64	13.38	31.47	85.90
18	0.87		0.86	0.88	0.02	2.49	3.83	1.53	6.14	8.73	9.19	18.68	34.61	87.83
19				0.01	0.90	0.88	1.80	3.75	1.61	4.91	9.48	14.99	24.25	62.58
Total	0.87	2.47	4.44	8.72	28.01	33.04	51.40	57.04	83.82	115.50	179.01	227.26	467.76	1,259.33

The table was rounded off to second decimal places.
Because the number of publication increased, every topic's distribution was increased.
Topic 1, Academic Publication; Topic 2, University Library Service (ULS); Topic 3, Technology Element; Topic 4, Copyright; Topic 5, Online Information Search (OIS); Topic 6, Open Access Publishing (OA Publishing); Topic 7, Green Open Access (Green OA); Topic 8, Research Cooperation; Topic 9, Academic Infrastructure (Academic Infra); Topic 10, Digital Communication; Topic 11, Researcher Network; Topic 12, Electronic Journal (E-journal); Topic 13, Preservation; Topic 14, Research Resource Management (RRM); Topic 15, Institute Policy; Topic 16, Institutional Repository (IR); Topic 17, Research Data; Topic 18, Research Evaluation; Topic 19, Informal Communication.

ics were *Topic 6: OA Publishing*, *Topic 7: Green OA*, and *Topic 19: Informal Communication*. The two downward topics were *Topic 11: Researcher Network* and *Topic 12: E-Journal* (Fig. 5).

OA publishing is a strategy that allows everyone to access articles simultaneously as they are published (European Commission, 2019). *Topic 6: OA Publishing* increased since the Budapest Declaration in 2002 and rapidly increased in the early 2010s when Biomed Central developed the Article Process Cost (APC) model (European Commission, 2019).

Green OA, topic 7, is an OA strategy where researchers deposit their articles in the repository. *Topic 7: Green OA* was studied in a small proportion and the rate started to rise earlier than *Topic 6: OA Publishing*. It means that Green OA is discussed by scholars as a traditional OA strategy. The upward trend of both OA topics shows that more researchers are getting interested in OA and OA is becoming more important in scholarly communication.

In the comparison of *Topic 6: OA Publishing* and *Topic 7: Green OA*, they increased or decreased in the same period before 2012. However, there is a difference between 2012-2014 and 2015-2018. *Topic 7: Green OA* increased, while *Topic 6: OA Publishing* decreased. In that period, a national level OA policy started in various countries, such as Spain (Cha et al., 2017). The policy is deeply related to

 https://doi.org/10.1633/JISTaP.2021.9.2.4

Table 3. First appearance and distribution of the topics

Topic No.	Year							
	1976	1983	1984	1985	1986	1987	1988	1989
1			0.88					1.18
2			0.87				0.87	0.01
3		0.72						0.01
4								0.86
5								0.01
6					0.15			0.01
7								0.01
8							0.85	0.01
9						0.87		0.01
10								0.83
11								0.01
12				0.86				0.82
13						0.85		0.01
14								0.01
15								0.01
16								0.58
17						0.85		0.01
18	0.87				0.86			0.88
19								0.01
Total	0.87	0.72	1.75	0.86	1.01	2.57	1.72	5.28

The table was rounded off to second decimal places.
Topic 1, Academic Publication; Topic 2, University Library Service (ULS); Topic 3, Technology Element; Topic 4, Copyright; Topic 5, Online Information Search (OIS); Topic 6, Open Access Publishing (OA Publishing); Topic 7, Green Open Access (Green OA); Topic 8, Research Cooperation; Topic 9, Academic Infrastructure (Academic Infra); Topic 10, Digital Communication; Topic 11, Researcher Network; Topic 12, Electronic Journal (E-journal); Topic 13, Preservation; Topic 14, Research Resource Management (RRM); Topic 15, Institute Policy; Topic 16, Institutional Repository (IR); Topic 17, Research Data; Topic 18, Research Evaluation; Topic 19, Informal Communication.

Green OA in that the researcher is obligated to open the articles to the public within a specific period. The difference in the fluctuations between *Topic 6: OA Publishing* and *Topic 7: Green OA* is because of the policy trend. Although *Topic 6: OA Publishing* decreased in the corresponding section, the percentage of research on topic 7 was 6.31%, similar to the 6.17% of topic 6.

Informal communication, topic 19, is another one of the types of scholarly communication. It is a personal and social communication method that includes sharing opinions through face-to-face discussion (Garvey, 1979; Mukherjee, 2009). As the development of the Internet affected scholarly communication in general, scholarly communication was actively conducted on the Internet in the 1990s (Barjak, 2006). In the 21st century, informal communication was advanced based on the development of the platform (Van Noorden, 2014). *Topic 19: Informal Communication* also increased significantly in the 1990s and is increasing except for the 2003-2005 section, when discussions of OA began to occur. It means research about informal communication has become more focused, following the advance of technology.

Topic 11: Researcher Networks accounted for 10% of the total research in 1988-1990 but since then has declined. Since 2006, when social media such as Twitter were widespread (Barjak, 2006), research on researchers' networks has increased a little. It occupied only 3.75% in 2015-2018.

Topic 12: E-Journal occupied 9.36% in 1988-1990 and 11.17% in 1991-1993. However, since then it has shown a

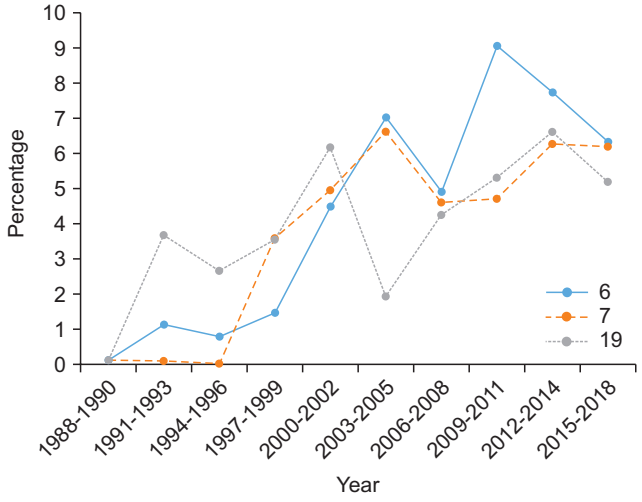


Fig. 5. Upward topics. Topic 6, Open Access Publishing (OA Publishing); Topic 7, Green Open Access (Green OA); Topic 19, Informal Communication.

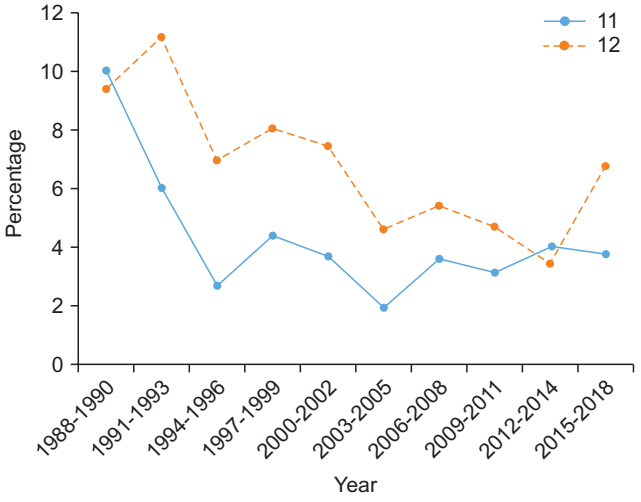


Fig. 6. Downward topics. Topic 11, Researcher Network; Topic 12, Electronic Journal (E-journal).

downward trend. In the early 1990s, large publishing companies such as Elsevier began to develop pricing models for digital publishing. Also, between 1990 and 1993, projects to make electronic journals online accessible were implemented (Lancaster, 1995). Later, as electronic journals became popular, they have replaced print journals in many libraries since 1996 (European Commission, 2019). Because the electronic journal format became common and popularized, the number of studies on the electronic journal decreased. However, research related to electronic journals continued to progress along with subscription issues, and the proportion of *Topic 12: E-Journal* increased during 2015-2018 when they started to develop an OA policy. However, the proportion in 2015-2018, 6.78%, is less than 11.17%, which is the highest proportion between 1991 and 1993 (Fig. 6).

The results of time series analysis infer that scholarly communication research changed following the advance of scholarly communication. According to the analysis of the first appearance of the topic, the topics related to technology became more specified, following various technological advances. More countries try to spread OA and both of two topics, *Topic 6: OA Publishing* and *Topic 7: Green OA*, the most closely related to OA, show an upward trend. It means OA gets more meaningful for both research and scholarly communication in this era. Besides this, many topics, such as *Topic 3: Technology Element*, were neutral. This infers that researchers focus on various aspects of scholarly communication constantly.

4.4. Result of Network Analysis

The top 30 words in each topic and the TF of each word were extracted. Then the Inverted Topic Frequency was calculated applying the concept of Inverted Document Frequency. The topic × word matrix was transformed to create the topic × topic (19*19) matrix. Cosine similarity coefficients were calculated based on the topic × topic matrix. The cosine matrix was used as the input value and a PFNet was created by using the pathfinder algorithm with r=∞ and q=n-1 through WNET ver. 0.4 program (Lee, 2013, 2014). NodeXL was used to visualize the network. The node's size was the distribution of the topic, and the thickness of the link was the cosine similarity between the two topics in the network (Fig. 7).

4.4.1. Centrality of Topics

Three centrality values were identified in the topic network of scholarly communication research. First, the topics with high global centrality are broadly linked to topics across the field of scholarly communication research. Second, mediation centrality is a concept that measures how much one node (topic) performs an intermediary or bridge role in constructing a network with another node (topic). Lastly, high regional centrality topics are influential topics within the cluster (Lee, 2012).

Because the PFNet using cosine similarity is a weighted network, the centrality measures proposed by Lee (2006c) were used. In order to grasp the mediation centrality of the topics, TBC was used. The TBC can be used as both regional centrality and global centrality depending on the measurement range. In this study, since the number of

 https://doi.org/10.1633/JISTaP.2021.9.2.4

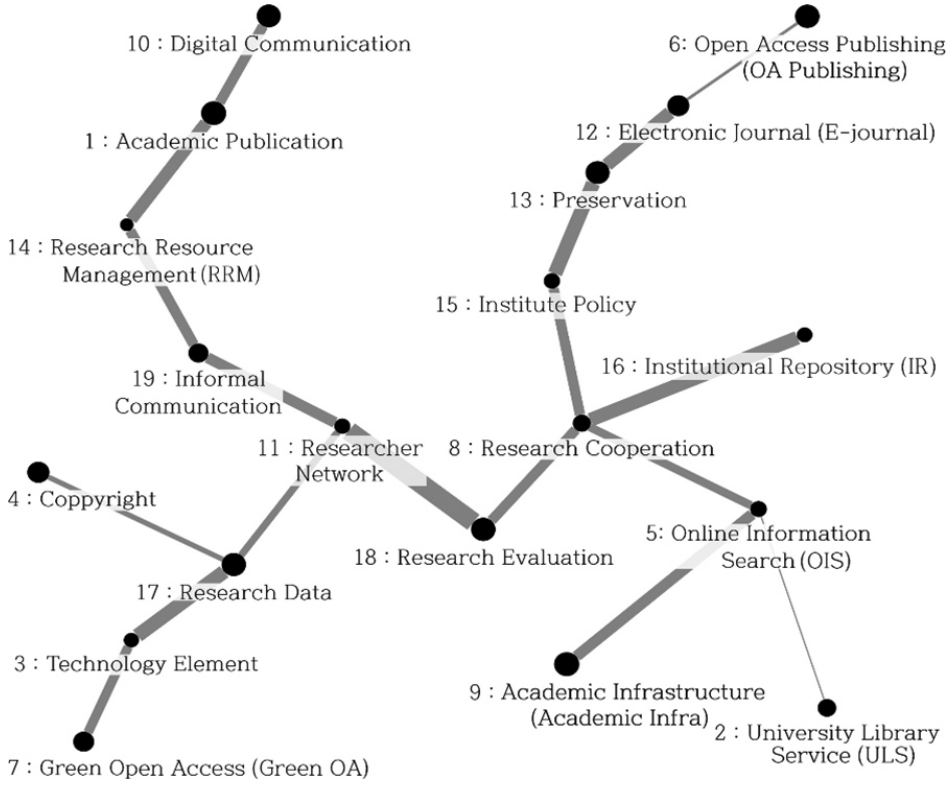


Fig. 7. Scholarly communication topic network.

nodes (topic) is only nineteen, the value was calculated globally.

Cmp was used to analyze the global centrality of the topics. Cmp considers patterns with high and low associations with other nodes. It is a value obtained by averaging the profile association and measuring the similarity between profile vectors by the Pearson correlation coefficient (Table 4).

If the node is located in the middle of the network and the correlation value with other nodes is low, the mean correlation could appear low. In this case, the centrality can be identified using the Cmp. The nodes with high Cmp are located in the center of the network, while the nodes with low Cmp are located in the network's surrounding area. The TBC represents the degree to which one node mediates other nodes, and the value is normalized to 0 to 1 (rTBC). Table 5 shows the top 5 nodes based on each centrality value.

The topics with high Cmp and TBC were *Topic 15: Institute Policy*, and *Topic 19: Informal Communication*. It means that *Topic 15: Institute Policy* and *Topic 19: Informal Communication* were located in the center of the network, and mediated other topics well. Also, they include 'link,' 'internet,' 'digit*,' 'data,' and 'inform*.' It means that the two topics, which are located at the center of the topic network and function as a mediator, share the context of digital information.

The top five topics with high Cmp are *Topic 15: Institute Policy, Topic 16: Institutional Repository (IR), Topic 19: Informal Communication, Topic 18: Research Evaluation*, and *Topic 13: Preservation*. They all included 'academ*,' 'research,' 'data,' and 'articl*.' Through this, it is inferred that the top five topics at the center of the topic network were studies of the research outcomes.

Topics with a high Cmp, including *Topic 14: Research Resource Management (RRM)* with the sixth highest Cmp, are highly related to research institution, except *Topic 19: Informal Communication*. Research evaluation is a method to measure universities' or research institutes' rank, and it is related to the compensation of researchers provided by institutions. Preservation and resource management are essential functions of institutional repositories. It means the topics related to the institute are relatively located on the middle of the scholarly communication research network.

The top five topics with normalized TBC were *Topic 14: RRM, Topic 19: Informal Communication, Topic 15: Institute Policy, Topic 8: Research Cooperation*, and *Topic 12: E-Journal*. They shared 'citat*.' Also, the sixth highest TBC topic was *Topic 18: Research Evaluation*. It could be

Table 4. Mean profile association (Cmp) and triangle betweenness centrality (TBC) of each topic

Topic No.	Topic	Cmp (-1 to 1)	TBC	rTBC (0 to 1)
1	Academic Publication	0.07526	58	0.37908
2	University Library Service (ULS)	-0.09105	50	0.32680
3	Technology Element	-0.08386	48	0.31373
4	Copyright	0.01873	24	0.15686
5	Online Information Search (OIS)	0.06370	58	0.37908
6	Open Access Publishing (OA Publishing)	-0.00257	25	0.16340
7	Green Open Access (Green OA)	-0.05637	16	0.10458
8	Research Cooperation	0.06743	66	0.43137
9	Academic Infrastructure (Academic Infra)	0.01427	43	0.28105
10	Digital Communication	-0.00717	25	0.16340
11	Researcher Network	0.09214	59	0.38562
12	Electronic Journal	0.06725	61	0.39869
13	Preservation	0.09892	55	0.35948
14	Research Resource Management (RRM)	0.09416	80	0.52288
15	Institute Policy	0.13389	69	0.45098
16	Institutional Repository (IR)	0.12997	57	0.37255
17	Research Data	-0.03695	40	0.26144
18	Research Evaluation	0.10675	59	0.38562
19	Informal Communication	0.12846	76	0.49673

rTBC, relative value of TBC.

Table 5. Top five nodes with centrality of mean profile association (Cmp) and highest triangle betweenness centrality (TBC)

Topic No.	Topic	Cmp (-1 to 1)	Topic No.	Topic	TBC	rTBC (0 to 1)
15	Institute Policy	0.13389	14	Research Resource Management (RRM)	80	0.52288
16	Institutional Repository (IR)	0.12997	19	Informal Communication	76	0.49673
19	Informal Communication	0.12846	15	Institute Policy	69	0.45098
18	Research Evaluation	0.10675	8	Research Cooperation	66	0.43137
13	Preservation	0.09892	12	Electronic Journal (E-Journal)	61	0.39869

rTBC, relative value of TBC.

said that topics about citation are instrumental in mediating other topics.

The NNC was used to analyze the regional centrality of the topics. NNC is calculated by analyzing how much the node is considered the nearest neighbor by other nodes. The NNC was used to analyze the characteristics of topics in each cluster.

4.4.2. Cluster of Topics

In analysis of the centrality of a few core nodes, the interpretation is not straightforward because individual nodes may be multifaceted. In this case, cluster analysis can be used to increase the discrimination of the subject analysis, because the common subjects of several nodes in the cluster are extracted, and more specific subjects are identified (Lee, 2006b). In this study, the parallel nearest neighbor clustering technique (PNNC) of Lee (2006b) was applied to extract the cluster from the topic network of scholarly communication research. PNNC analysis was performed through WNET ver 0.4, which was used to apply the Pathfinder algorithm. As a result, six clusters were identified, and the clusters are visualized with different

 https://doi.org/10.1633/JISTaP.2021.9.2.4

colors. The node size in each cluster was decided depending on the NNC value (Fig. 8).

Table 6 summarizes the nearest node of each node belonging cluster and the relative NNC (rNNC, 0 to 1), which is the normalized NNC between 0 and 1.

Cluster 1 consists of *Topic 1: Academic Publication, Topic 10: Digital Communication*, and *Topic 14: RRM. Topic 1: Academic Publication* displays the highest rNNC. The words 'inform*,' 'journal,' 'academ*,' and 'use' are included in all three subjects. Also, three topics shared the word 'provid*' or 'servic*.' This means the three topics share the context of information supply.

Cluster 2 is composed of *Topic 2: ULS, Topic 5: OIS*, and *Topic 9: Academic Infra*. *Topic 5: OIS* displays the highest rNNC. All three subjects share the words 'research' and 'journal,' 'articl*,' 'data,' and 'librari*.' Also, they include 'support' or 'servic*.' This means that topics in cluster 2 linked to each other in the point of research support. *Topic 5: OIS* displays the largest rNNC in cluster 2 and the sixth lowest TBC. Both the TBC and Cmp of *Topic 2: ULS* and *Topic 9: Academic Infra* belong to the lower level. The character of the centrality of topics in cluster 2 shows that the topics of cluster 2 are independent.

Cluster 3 contains *Topic 3: Technology Element, Topic 4: Copyright, Topic 7: Green OA*, and *Topic 17: Research Data*. They share 'academ*,' 'data,' 'journal,' and 'open_access.' More than one topics include 'elecron*' or 'technolog*.' This means that the topics are related in the context of science and technology. The topics of cluster 3 are independent among all academic communication research topics in that they have a lower global centrality. *Topic 3: Technology Element* and *Topic 17: Research Data* are the closest neighbors to each other, so the connection between the two topics is strong. It shows that studies about the research data and studies about the technical elements are related to each other.

Cluster 4 includes *Topic 6: OA Publishing*, *Topic 12: E-Journal, Topic 13: Preservation*, and *Topic 15: Institute Policy*. Topics of cluster 4 share 'data,' 'academ*,' and 'publish.' They include 'manag*' or 'access.' This infers that topics of cluster 4 are linked with the context of research management. *Topic 13: Preservation* is the node with the highest rNNC and the fifth highest Cmp. In that sense, it is located at the center of the whole network and mediates the topics in the cluster. *Topic 12: E-Journal* has the fifth-highest TBC, which significantly functions as mediation in the whole network, while the regional centrality is lower than for *Topic 13: Preservation. Topic 6: OA Publishing* has lower TBC and rNNC. This means that *Topic 6: OA Publishing* is independent. *Topic 15: Institute Policy* displays the highest Cmp and the third highest TBC, and the second lowest rNNC. In that sense, *Topic 15: Institute Policy* locates in the middle of the whole network and mediates the topics at the global level rather than the regional level.

Cluster 5 consists of *Topic 8: Research Cooperation*, and *Topic 16: IR*. They share 'system' and 'model' and link to each other with the system context. *Topic 8: Research Cooperation* has the fourth highest TBC, and *Topic 16: IR* has the second highest Cmp and closest node. In cluster 5, the topic with significant mediation values and the topic located at the center are strongly linked to each other with the system context.

Group 6 includes *Topic 11: Researcher Networks, Topic 18: Research Evaluation*, and *Topic 19: Informal Communication*. All three topics include 'academ*' and 'review,' or

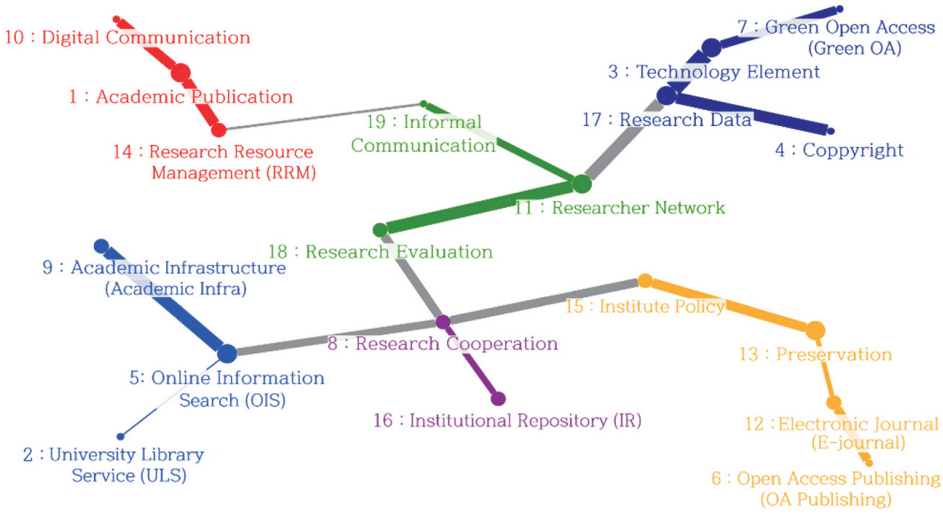


Fig. 8. Cluster of scholarly communication topic network.

Table 6. Nearest neighbor centrality of topics in six clusters

Cluster		Topic	rNNC (0 to 1)		Nearest neighbor
1	1	Academic Publication	0.11111	14	Research Resource Management (RRM)
	10	Digital Communication	0	1	Academic Publication
	14	Research Resource Management (RRM)	0.05556	1	Academic Publication
2	2	University Library Service (ULS)	0	5	Online Information Search (OIS)
	5	Online Information Search (OIS)	0.11111	9	Academic Infrastructure (Academic Infra)
	9	Academic Infrastructure (Academic Infra)	0.05556	5	Online Information Search (OIS)
3	3	Technology Element	0.11111	17	Research Data
	4	Copyright	0	17	Research Data
	7	Green Open Access (Green OA)	0	3	Technology Element
	17	Research Data	0.11111	3	Technology Element
4	6	Open Access Publishing (OA Publishing)	0	12	Electronic Journal (E-Journal)
	12	Electronic Journal (E-Journal)	0.05556	13	Preservation
	13	Preservation	0.11111	15	Institute Policy
	15	Institute Policy	0.05556	13	Preservation
5	8	Research Cooperation	0.05556	16	Institutional Repository (IR)
	16	Institutional Repository (IR)	0.05556	8	Research Cooperation
6	11	Researcher Network	0.11111	18	Research Evaluation
	18	Research Evaluation	0.05556	11	Researcher Network
	19	Informal Communication	0	11	Researcher Network

rNNC, relative nearest neighbor centrality.

'peer_review,' and they include 'scientist' or 'author,' meaning the topics linked in the academic exchange context. *Topic 11: Researcher Networks* has the highest rNNC and lower TBC. Although the role of mediator at the global level is small, it is deeply related to the other topics within the cluster. On the other hand, *Topic 19: Informal Communication* has high TBC and Cmp, but the rNNC of topic 19 is 0. It means that *Topic 19: Informal Communication* has a significant role in mediating the global level, but the connection is weak within the cluster.

As a result of the network analysis, high ranked topics were different depending on the centralities. Through this, the locational characteristics and mediation weight of the topic in the topic network were identified. The words shared by all topics were 'academ*,' 'journal,' and 'data.' These were words related to research resources. It means that all scholarly communication research topics share the context of research resources as shown in the definition.

5. CONCLUSIONS AND IMPLICATIONS

This study analyses scholarly communication research with informatics approaches. It identified scholarly communication research topics based on the articles published from 1970 to 2008 in the Scopus database and figured out research-inflected scholarly communication issues such as OA. Also, the topics shared different contexts. The results are summarized as follows.

First, LDA topic modeling identified nineteen scholarly communication research topics, including research resource management and research data. The proportion of topics were even. The topics were detailed and subdivided. It means that researchers have specified the topics of research over time. Simultaneously, the even proportion of the topics means they have conducted the study broadly and evenly.

Second, time series analysis found the order of topics' first appearance and three upward topics and two downward topics. It infers that more and more various topics are being developed, following the advance of technology,

 https://doi.org/10.1633/JISTaP.2021.9.2.4

and the trend of the topics reflects scholarly communication issues such as OA. Fourteen neutral topics indicate that researchers continue to focus on various aspects of scholarly communication.

Third, network analysis figured out the structural characteristic of scholarly communication research. Topics with high Cmp are related to the institution, and high TBC topics share the citation context. It indicates the topics related to the institution are located in the center of the topic network, and topic related to citation mediates other topics closely. Besides this, all topics are related to the academic resources, and six clusters figured by PNNC shared distinguished contexts. This verified that the research is focused on the academic resources under the definition of scholarly communication and, at the same time, each of them is conducted on a different aspect.

As observed in its various concepts, scholarly communication contains broad issues related to the research society. This could cause ambiguity when researchers understand scholarly communication and start new research about it. In that sense, topics of scholarly communication research, their trend, and structural characteristics identified through this research are helpful in order to understand what scholarly communication is and how researchers studied it. The topics were detailed, and they became more varied over time. Even though there was limited research data because very few articles were published from 1976 to 1988 at the early stage of the research, the characteristics of scholarly communication topics derived can be used as data for researchers to find a new theme in the future.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, 24(1), 1-7. https://doi.org/10.1016/j.iedeen.2017.06.002.

Barjak, F. (2006). The role of the Internet in informal scholarly communication. *Journal of the American Society for Information Science and Technology*, 57(10), 1350-1367. https://doi.org/10.1002/asi.20454.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84. https://doi.org/10.1145/2133806.2133826.

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

Cha, M. K., Song, K. J., & Kim, M. K., (2017). A study on improving laws and regulations for open access of research papers from national research and development projects. *Journal of the Korean Society for Library and Information Science*, 51(1), 147-174. https://doi.org/10.4275/KSLIS.2017.51.1.147.

Choi, S. Y., & Ko, E. J. (2019). Analysis of <Korean Journal of Journalism & Communication Studies> from 1960 to 2018 using metadata with dynamic topic modeling. *Korean Journal of Journalism & Communication Studies*, 63(4), 7-42. https://doi.org/10.20879/kjjcs.2019.63.4.001.

De Solla Price, D. J. (1963). *Little science, big science*. Columbia University Press.

European Commission. (2019). *Future of scholarly publishing and scholarly communication: Report of the Expert Group to the European Commission*. Publications Office of the European Union.

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Garvey, W. D. (1979). *Communication: The essence of science - facilitating information exchange among librarians, scientists, engineers and students*. Pergamon.

Jin, S. A., & Song, M. (2016). Topic Modeling based Interdisciplinarity Measurement in the Informatics Related Journals. *Journal of the Korean Society for information Management*, 16(33), 7-32. https://doi.org/10.3743/KOSIM.2016.33.1.007.

Keshner, E. A., Weiss, P. T., Geifman, D., & Raban, D. (2019). Tracking the evolution of virtual reality applications to rehabilitation as a field of study. *Journal of Neuroengineering and Rehabilitation*, 16(1), 76. https://doi.org/10.1186/s12984-019-0552-6.

Klain-Gabbay, L., & Shoham, S. (2018). *Scholarly communication and the academic library: Perceptions and recent developments*. https://www.intechopen.com/books/a-complex-systems-perspective-of-communication-from-cells-to-societies/scholarly-communication-and-the-academic-library-perceptions-and-recent-developments.

Lancaster, F. W. (1995). The evolution of electronic publishing. *Library Trends*, 43(4), 518-527. https://www.ideals.illinois.edu/bitstream/handle/2142/7981/librarytrendsv43i4c_opt.pdf.

Lee, J. Y. (2006a). A study on the network generation methods

for examining the intellectual structure of knowledge domains. *Journal of the Korean Society for Library and Information Science*, 40(2), 333-355. https://doi.org/10.4275/KSLIS.2006.40.2.333.

Lee, J. Y. (2006b). A novel clustering method for examining and analyzing the intellectual structure of a scholarly field. *Journal of the Korean Society for Information Management*, 23(4), 215-231. https://doi.org/10.3743/KOSIM.2006.23.4.215.

Lee, J. Y. (2006c). Centrality measures for bibliometric network analysis. *Journal of the Korean Society for Library and Information Science*, 40(3), 191-214. https://doi.org/10.4275/KSLIS.2006.40.3.191.

Lee, J. Y. (2013). A comparison study on the weighted network centrality measures of tnet and WNET. *Journal of the Korean Society for Information Management*, 30(4), 241-264. https://doi.org/10.3743/KOSIM.2013.30.4.241.

Lee, J. Y. (2014). *WNET* (Version 0.4) [Computer software].

Lee, S. S. (2012). *Network analysis methods*. Nonhyoung.

Mukherjee, B. (2009). Scholarly communication: A journey from print to web. *Library Philosophy and Practice*, 285.

Newman D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June 2-4). Automatic evaluation of topic coherence. In R. M. Kaplan (Ed.), *Proceedings of the HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100-108). Association for Computational Linguistics.

Park, J. H., & Oh, H. J. (2017). Comparison of topic modeling methods for analyzing research trends of archives management in Korea: Focused on LDA and HDP. *Journal of Korean Library and Information Science Society*, 48(4), 235-258. https://doi.org/10.16981/kliss.48.201712.235.

Park, J. H., & Song, M. (2013). A study on the research trends in library & information science in Korea using topic modeling. *Journal of the Korean Society for information Management*, 30(1), 7-32. https://doi.org/10.3743/KOSIM.2013.30.1.007.

Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: Studies in knowledge organization*. Ablex.

Song, M. (2017). *Text mining*. Chung Ram.

Syed, S., & Spruit, M. (2017, October 19-21). Full-text or abstract? Examining topic coherence scores using Latent Dirichlet allocation. In T. Kamishima (Ed.), *Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 165-174). IEEE. https://doi.org/10.1109/DSAA.2017.61.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature*, 512(7513), 126-129. https://doi.org/10.1038/512126a.

Weingart, S. (2012). *Topic modeling for humanists: A guided tour*. http://www.scottbot.net/HIAL/index.html@p=19113.html.

Zuccala, A. (2006). Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, 57(2), 152-168. https://doi.org/10.1002/asi.20256.

Appendix 1. Topic words (stemmed)

	TOPIC: 1	TOPIC: 2	TOPIC: 3	TOPIC: 4	TOPIC: 5	TOPIC: 6	TOPIC: 7	TOPIC: 8	TOPIC: 9	TOPIC: 10	TOPIC: 11	TOPIC: 12	TOPIC: 13	TOPIC: 14	TOPIC: 15	TOPIC: 16	TOPIC: 17	TOPIC: 18	TOPIC: 19
1	academ	academ	academ	academ	access	academ	analysi	academ	academ	academ	academ	academ	academ	academ	academ	academ	academ	academ	academ
2	activ	articl	articl	author	articl	access	articl	associ	access	access	articl	author	archiv	associ	administr	articl	access	altrmetri	articl
3	analysi	author	author	base	author	articl	citat	author	articl	articl	author	base	articl	author	author	author	articl	chang	author
4	articl	chang	book	book	citat	author	data	citat	base	author	citat	chang	book	citat	articl	chang	author	data	blog
5	citat	collect	databas	chang	cost	citat	develop	co_athorship	collect	blog	communi	citat	chang	collection	citat	collect	citat	digit	chang
6	cost	digit	digit	citat	data	cite	digit	collabor	current	chang	contribut	data	current	complex	cultur	data	collect	electron	collabor
7	data	electron	develop	copyright	directori	copyright	inform	data	data	data	data	develop	data	crossref	data	develop	connect	evalu	content
8	develop	faculti	disciplin	data	effect	data	information_ literaci	digit	digit	develop	field	ejournal	digit	disciplin	differ	futur	data	factor	digit
9	electron	faculty_memb	econom	develop	factor	digit	institutional_ repositori	examin	impact	digit	indic	field	electron	grey_literatur	digit	impact	develop	faculti	distribut
10	exhibit	impact	effect	digit	frailti	electron	journal	experi	import	disciplin	inform	impact	knowledg	impact	faculti	implic	electron	impact	educ
11	impact	institutional_ repositori	electron	electron	grey_literatur	inform	librari	inform	inform	increas	institut	institutional_ repositori	libararian_ archivist	inform	fiscal	includ	encod	inform	exchang
12	inform	issu	factor	embargo	inform	institutional_ repositori	link	infrastructur	infrastructur	inform	journal	internet	librari	journal	impact	inform	inform	journal	inform
13	journal	journal	group	inform	internet	journal	manuscript	journal	journal	information_ literaci	librari	journal	manag	librari	index	intitutional_ repositari	journal	knowledg	journal
14	librarian	librari	impact	institutional_ repositori	issu	librari	model	librari	librari	journal	network	librari	model	metadata	inform	journal	librarian	librari	librari
15	librari	librarian	inform	journal	journal	librarian	new	librarian	model	link	ocid	like	open	network	initi	librari	manag	model	librarian
16	manag	need	journal	knowledg	librari	market	open_access	model	nation	model	onlin	measur	paper	offer	institut	manag	masur	new	literatur
17	new	open_access	librari	librari	look	model	paper	opportun	network	new	peer_review	network	polici	open_ access	institutional_ repositori	manuscript	new	open_access	new
18	open_access	paper	model	market	manag	new	peer_review	orcid	open_access	open_access	polici	new	present	paper	librari	new	open_access	practic	open_access
19	predatory_ publish	profession	network	model	metadata	open_access	provid	peer_review	patron	polici	project	open_access	process	practic	librarian	open_access	public	profession	paper
20	present	public	technolog	nation	new	open_access	public	public	practic	project	public	orcid	public	public	link	polici	research	provid	public
21	public	publish	open_access	open_access	open_access	paper	repositori	publish	process	provid	publish	provid	publish	research	mandat	preprint	resourc	public	publish
22	publish	research	publish	paper	paper	polici	research	research	project	rebutt	reference	publish	research	review	open_access	provid	reuse	publish	resourc
23	refer	scienc	research	public	polici	public	scientif	scienc	public	research	research	research	risk	scientif	paper	research	scholar	research	role
24	research	scientif	role	publish	public	publish	studi	scientif	publish	result	scienc	scholar	scienc	servic	polici	retract	scienc	resourc	support
25	role	servic	scienc	repositori	research	research	subject	scientist	report	survey	scientif	scienc	scientif	social	qualiti	role	servic	review	survey
26	scienc	studi	scientist	research	resourc	share	tool	seek	research	technolog	studi	system	serial	system	research	scienc	system	reward_ system	tweet
27	scientist	technolog	system	role	studi	studi	univers	sparc	role	tool	subgroup	technolog	share	univers	servic	servic	system	scholar	twitter
28	servic	survey	technolog	scienc	system	survey	use	studi	scholar	univers	survey	univers	system	use	studi	share	univers	studi	univers
29	studi	tradit	twitter	studi	technolog	univers	web	system	system	use	tradit	use	univers	web	use	technolog	user	univers	use
30	use	univers	use	use	web	work	work	web	univers	web	web	web	work	work	web	univers	web	use	web

Topic 1, Academic Publication; Topic 2, University Library Service (ULS); Topic 3, Technology Element; Topic 4, Copyright; Topic 5, Online Information Search (OIS); Topic 6, Open Access Publishing (OA Publishing); Topic 7, Green Open Access (Green OA); Topic 8, Research Cooperation; Topic 9, Academic Infrastructure (Academic Infra); Topic 10, Digital Communication; Topic 11, Researcher Network; Topic 12, Electronic Journal (E-journal); Topic 13, Preservation; Topic 14, Research Resource Management (RRM); Topic 15, Institute Policy; Topic 16, Institutional Repository (IR); Topic 17, Research Data; Topic 18, Research Evaluation; Topic 19, Informal Communication.

Appendix 2. TF*ITF matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	0.307077	0.227664	0.286023	0.219918	0.198935	0.300289	0.194994	0.16521	0.192612	0.163056	0.235374	0.298331	0.195985	0.193341	0.202368	0.343718	0.26975	0.165819
2	0.307077	0	0.235808	0.256602	0.23498	0.391725	0.269323	0.190742	0.193832	0.270517	0.316553	0.248572	0.25796	0.172973	0.247729	0.181155	0.392093	0.28102	0.251534
3	0.227664	0.235808	0	0.250729	0.216523	0.26112	0.12549	0.192747	0.250048	0.222345	0.119967	0.201593	0.238591	0.25159	0.209859	0.154322	0.251295	0.224058	0.247494
4	0.286023	0.256602	0.250729	0	0.157611	0.300635	0.351741	0.213404	0.274003	0.227488	0.258853	0.330063	0.366068	0.159864	0.173798	0.159286	0.205822	0.316515	0.212593
5	0.219918	0.23498	0.216523	0.157611	0	0.235152	0.138229	0.12556	0.109479	0.216742	0.150567	0.220089	0.172636	0.104774	0.169229	0.156617	0.240224	0.155488	0.127302
6	0.198935	0.391725	0.26112	0.300635	0.235152	0	0.310809	0.204402	0.264372	0.320475	0.22131	0.296759	0.444901	0.208414	0.206944	0.168496	0.341911	0.291126	0.226036
7	0.300289	0.269323	0.12549	0.351741	0.138229	0.310809	0	0.213279	0.153589	0.218995	0.250884	0.245312	0.256228	0.229461	0.256903	0.179713	0.272818	0.237943	0.167286
8	0.194994	0.190742	0.192747	0.213404	0.12556	0.204402	0.213279	0	0.144659	0.137416	0.174875	0.177434	0.185551	0.238791	0.131421	0.105321	0.262609	0.154715	0.124152
9	0.16521	0.193832	0.250048	0.274003	0.109479	0.264372	0.153589	0.144659	0	0.281733	0.195813	0.29341	0.274745	0.235447	0.144912	0.146069	0.279253	0.257776	0.137943
10	0.192612	0.270517	0.222345	0.227488	0.216742	0.320475	0.218995	0.137416	0.281733	0	0.304509	0.246131	0.211817	0.211715	0.300977	0.234274	0.37526	0.172498	0.300366
11	0.163056	0.316553	0.119967	0.258853	0.150567	0.22131	0.250884	0.174875	0.195813	0.304509	0	0.274708	0.201016	0.156243	0.205034	0.121181	0.207468	0.14303	0.124744
12	0.235374	0.248572	0.201593	0.330063	0.220089	0.296759	0.245312	0.177434	0.29341	0.246131	0.274708	0	0.178068	0.207816	0.213591	0.197406	0.353898	0.283175	0.250702
13	0.298331	0.25796	0.238591	0.366068	0.172636	0.444901	0.256228	0.185551	0.274745	0.211817	0.201016	0.178068	0	0.169123	0.126894	0.194231	0.267662	0.236448	0.151318
14	0.195985	0.172973	0.25159	0.159864	0.104774	0.208414	0.229461	0.238791	0.235447	0.211715	0.156243	0.207816	0.169123	0	0.135724	0.116391	0.247594	0.236496	0.137772
15	0.193341	0.247729	0.209859	0.173798	0.169229	0.206944	0.256903	0.131421	0.144912	0.300977	0.205034	0.213591	0.126894	0.135724	0	0.158718	0.185711	0.211121	0.26271
16	0.202368	0.181155	0.154322	0.159286	0.156617	0.168496	0.179713	0.105321	0.146069	0.234274	0.121181	0.197406	0.194231	0.116391	0.158718	0	0.221694	0.20017	0.130737
17	0.343718	0.392093	0.251295	0.205822	0.240224	0.341911	0.272818	0.262609	0.279253	0.37526	0.207468	0.353898	0.267662	0.247594	0.185711	0.221694	0	0.357572	0.26874
18	0.26975	0.28102	0.224058	0.316515	0.155488	0.291126	0.237943	0.154715	0.257776	0.172498	0.14303	0.283175	0.236448	0.236496	0.211121	0.20017	0.357572	0	0.352924
19	0.165819	0.251534	0.247494	0.212593	0.127302	0.226036	0.167286	0.124152	0.137943	0.300366	0.124744	0.250702	0.151318	0.137772	0.26271	0.130737	0.26874	0.352924	0

TF, term frequency; ITF, inverted topic frequency.

https://doi.org/10.1633/JISTaP.2021.9.2.4

Appendix 3. Cosine matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	0.910221	0.867964	0.882546	0.899958	0.917082	0.846874	0.90596	0.928836	0.935405	0.914424	0.923064	0.89612	0.939694	0.904805	0.882952	0.872441	0.920347	0.928603
2	0.910221	1	0.915258	0.898988	0.91895	0.866065	0.896258	0.910153	0.917094	0.909432	0.85233	0.913992	0.894111	0.914136	0.87004	0.879166	0.880331	0.844847	0.890553
3	0.867964	0.915258	1	0.863282	0.876822	0.911625	0.933835	0.886203	0.895599	0.846489	0.918712	0.897809	0.887583	0.899836	0.885671	0.897625	0.942893	0.893905	0.896627
4	0.882546	0.898988	0.863282	1	0.907934	0.789729	0.820876	0.912713	0.845992	0.904799	0.914557	0.844028	0.896094	0.924014	0.897743	0.907217	0.926265	0.874208	0.882757
5	0.899958	0.91895	0.876822	0.907934	1	0.851065	0.893887	0.932208	0.937113	0.876764	0.903794	0.893958	0.912343	0.869344	0.924758	0.919223	0.866853	0.910582	0.918402
6	0.917082	0.866065	0.911625	0.789729	0.851065	1	0.850756	0.883411	0.902886	0.869943	0.895642	0.922049	0.840017	0.900709	0.883945	0.890509	0.85061	0.915445	0.897812
7	0.846874	0.896258	0.933835	0.820876	0.893887	0.850756	1	0.843386	0.914931	0.757306	0.883849	0.89189	0.878198	0.853796	0.855973	0.871219	0.894134	0.876164	0.929604
8	0.90596	0.910153	0.886203	0.912713	0.932208	0.883411	0.843386	1	0.893099	0.933101	0.882615	0.897511	0.907693	0.901971	0.935125	0.942536	0.891798	0.935597	0.888665
9	0.928836	0.917094	0.895599	0.845992	0.937113	0.902886	0.914931	0.893099	1	0.851725	0.88779	0.884636	0.869747	0.893631	0.913713	0.903234	0.858835	0.879435	0.92676
10	0.935405	0.909432	0.846489	0.904799	0.876764	0.869943	0.757306	0.933101	0.851725	1	0.865025	0.881446	0.900147	0.910501	0.914422	0.855362	0.879464	0.879546	0.84055
11	0.914424	0.85233	0.918712	0.914557	0.903794	0.895642	0.883849	0.882615	0.88779	0.865025	1	0.881717	0.895322	0.923293	0.915146	0.923364	0.929718	0.952723	0.937247
12	0.923064	0.913992	0.897809	0.844028	0.893958	0.922049	0.89189	0.897511	0.884636	0.881446	0.881717	1	0.941224	0.914111	0.911791	0.886526	0.878567	0.915449	0.934455
13	0.89612	0.894111	0.887583	0.896094	0.912343	0.840017	0.878198	0.907693	0.869747	0.900147	0.895322	0.941224	1	0.909746	0.942636	0.883136	0.913805	0.90446	0.926803
14	0.939694	0.914136	0.899836	0.924014	0.869344	0.900709	0.853796	0.901971	0.893631	0.910501	0.923293	0.914111	0.909746	1	0.908126	0.927045	0.919064	0.889671	0.935739
15	0.904805	0.87004	0.885671	0.897743	0.924758	0.883945	0.855973	0.935125	0.913713	0.914422	0.915146	0.911791	0.942636	0.908126	1	0.92711	0.901508	0.912807	0.909773
16	0.882952	0.879166	0.897625	0.907217	0.919223	0.890509	0.871219	0.942536	0.903234	0.855362	0.923364	0.886526	0.883136	0.927045	0.92711	1	0.87232	0.922905	0.926431
17	0.872441	0.880331	0.942893	0.926265	0.866853	0.85061	0.894134	0.891798	0.858835	0.879464	0.929718	0.878567	0.913805	0.919064	0.901508	0.87232	1	0.901354	0.894437
18	0.920347	0.844847	0.893905	0.874208	0.910582	0.915445	0.876164	0.935597	0.879435	0.879546	0.952723	0.915449	0.90446	0.889671	0.912807	0.922905	0.901354	1	0.918114
19	0.928603	0.890553	0.896627	0.882757	0.918402	0.897812	0.929604	0.888665	0.92676	0.84055	0.937247	0.934455	0.926803	0.935739	0.909773	0.926431	0.894437	0.918114	1

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(2): 66-82, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.2.5

Issues and Challenges in the Extraction and Mapping of Linked Open Data Resources with Recommender Systems Datasets

Rosmamalmi Mat Nawi*
Centre for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
The National University of Malaysia, Selangor, Malaysia
E-mail: rosmamalmi@siswa.ukm.edu.my

Shahrul Azman Mohd Noah
Centre for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
The National University of Malaysia, Selangor, Malaysia
E-mail: shahrul@.ukm.edu.my

Lailatul Qadri Zakaria
Centre for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
The National University of Malaysia, Selangor, Malaysia
E-mail: lailatul.qadri@.ukm.edu.my

ABSTRACT

Recommender Systems have gained immense popularity due to their capability of dealing with a massive amount of information in various domains. They are considered information filtering systems that make predictions or recommendations to users based on their interests and preferences. The more recent technology, Linked Open Data (LOD), has been introduced, and a vast amount of Resource Description Framework data have been published in freely accessible datasets. These datasets are connected to form the so-called LOD cloud. The need for semantic data representation has been identified as one of the next challenges in Recommender Systems. In a LOD-enabled recommendation framework where domain awareness plays a key role, the semantic information provided in the LOD can be exploited. However, dealing with a big chunk of the data from the LOD cloud and its integration with any domain datasets remains a challenge due to various issues, such as resource constraints and broken links. This paper presents the challenges of interconnecting and extracting the DBpedia data with the MovieLens 1 Million dataset. This study demonstrates how LOD can be a vital yet rich source of content knowledge that helps recommender systems address the issues of data sparsity and insufficient content analysis. Based on the challenges, we proposed a few alternatives and solutions to some of the challenges.

Keywords: recommender system, linked open data, DBpedia

Received: March 4, 2021 **Revised:** April 25, 2021
Accepted: May 11, 2021 **Published:** June 30, 2021

***Corresponding Author:** Rosmamalmi Mat Nawi
https://orcid.org/0000-0003-1147-3092
E-mail: rosmamalmi@siswa.ukm.edu.my

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© Rosmamalmi Mat Nawi, Shahrul Azman Mohd Noah, Lailatul Qadri Zakaria, 2021

1. INTRODUCTION

Recommendation Systems (RS) technology is currently a successful solution to addressing knowledge overload problem with the ever-increasing amount of online knowledge. Given its widespread use in many web applications, the importance of RS cannot be overstated. Due to the issue of sparseness, cold start, overspecialization, and even the nature of scalability in recommender systems, many researchers have done a lot of research to overcome these limitations by adapting diverse recent technologies. Among the applied technologies are big data (Al-Barznji & Atanassov, 2017; Hammou et al., 2019; Maillo et al., 2017), semantic (Ameen, 2019; Barros et al., 2020; Figueroa et al., 2019), and deep learning (Feng et al., 2019; Sankar et al., 2020; Zhang et al., 2019).

RS quality often degrades when insufficient information or few ratings are available, which is a commonly well-known challenge in RS. RS performance can be improved by enriching the user's profile with relevant features and envisioning the recommendations' improvements using more structured and semantically richer data about the item. This valuable information or data presented in some form of knowledge bases can be effectively utilized in RS to overcome some of its major issues, such as sparsity and cold-start. Linked Open Data (LOD) is one such semantic dataset seen to be potentially exploited by RS to improve the recommendations' quality. Nonetheless, the knowledge of LOD should not be limited only to exploiting relationships among items or to enrich data about items and users. It can also produce implicit knowledge about them and their relationships.

LOD is a relatively new topic with enormous potential in a variety of fields (Yochum et al., 2020), including RS. Linked data-based RS (referring to recommender systems applying LOD) still struggle to develop recommendations for end-users with acceptable accuracy. This is explained by the need for data from both user profiles and item descriptions (Figueroa et al., 2019). There is also very little support for users who consume information published as LOD, most of which focus on searching and visualizing the collected data (Oliveira et al., 2017). They can display related vocabulary and data in facets; however, they do not recommend relevant information to infer users' interests (Franz et al., 2010).

Datasets such as the Million Song, LastFM, and MovieLens can be further enriched by being automatically mapped to the LOD identifiers. Then, additional features can be derived from publicly available knowledge bases like the DBpedia. For example, in the MovieLens dataset, we could get more granular features for movie representations using the LOD cloud. A large number of additional features can provide valuable information for various applications. It depends on how the data are provided for a particular item based on specific purposes. For example, if viewers like a movie with a specific director, they might also like other movies directed by the same director. In the movie domain, information about the movie's actors or directors is available via linked information. If the same actor starred in two movies, the two movies are linked. This will enable us in inferring new facts about movies, which can contribute toward more insightful details in the future. Such structured semantic data can be exploited to improve RS.

Previous research focusing on the mapping and extraction of MovieLens and IMDB can be found in the work of Peralta (2007), but the discussed challenges were mainly on the datasets and not related to linked data. Although works presented in Di Noia et al. (2012), Meymandpour and Davis (2015), and Kushwaha and Vyas (2014) describe approaches to integrate MovieLens and DBpedia, they did not describe and discuss the challenges encountered in extracting knowledge from the linked data (i.e., DBpedia) in terms of data quality and accuracy. Thus, this paper aims to discuss such challenges based on our experience and experiment in exploiting linked data to enhance the performance of recommender systems.

This paper demonstrates our study in exploring data about movies with two datasets: a public dataset of MovieLens 1 Million (ML1M), and DBpedia as the LOD dataset. Furthermore, we also present a general review of the LOD application to RS in existing works involving various elements. We then address the challenges of exploring linked data and the tasks for extracting and mapping for both used datasets. Based on the challenges, alternative solutions are proposed which may resolve some of the challenges.

To this end, we extract and integrate data provided by DBpedia and ML1M by applying the mapping set. The end of this study's contribution is that we provide a clean dataset (ML1M-DBpedia)[1] with additional information for the movie list from the ML1M, enhanced with the '*director*' and '*starring*' data. The primary aim is to tackle sparsity issues in the RS by enriching the dataset with spe-

[1]The ML1M-DBpedia dataset is available for download at https://data.mendeley.com/datasets/ypwk2536gh/1

cific attributes. From these findings, it can allow practitioners in the field to analyze, evaluate, and compare their works with existing approaches.

The paper is organized as follows. Section 2 provides background information of the relevant topics of this paper which are the RS and linked data. Section 3 discusses prior works on the application of LOD in RS. Section 4 then deals with our experimental procedure and setting of mapping ML1M to DBpedia. We discuss the challenges of the linked and extraction process between the two datasets in Section 5 and finally, Section 6 presents conclusion of our work.

2. LINKED OPEN DATA ON RECOMMENDER SYSTEM

The application of Semantic Web (SW) technologies seems to offer promising solutions to the long-standing issues of recommender systems such as data sparsity and cold-start problems (Saat et al., 2018). This is due to the fact that SW contains structured knowledge which can be retrieved and exploited. Defining information using SW concepts will enable data to be retrieved and integrated in more meaningful and effective ways (Hidayat et al., 2012) by incorporating data from a wide range of sources.

Using different knowledge representation languages such as Resource Description Framework (RDF), W3C Web Ontology Language, and Uniform Resource Identifier (URI) protocols has allowed data to be interoperable. As a consequence, the data can be shared and reused across a wide range of platforms, applications, and communities. Yadav et al. (2020) declare that the aim is to link data from various isolated applications, highlight the importance of publicly bringing out data for other applications to use, and connect the data using standard schema.

LOD denotes freely accessible semantic data on the Web. Combining Linked Data and Open Data yields the concept of LOD. LOD's aim is to make data semantically linked and machine-processable (Alshammari & Nasraoui, 2019). The DBpedia is one of LOD's significant contributions (Auer et al., 2007), which has been recognized as one of the valuable sources for building RS. DBpedia data can be easily and readily extracted using the SPARQL query language and the defined properties mentioned in its ontology. This valuable information presented in a linked knowledge base can be effectively utilized in RS.

The existence of interconnected LOD datasets allows federated queries to be performed between different datasets (Cappiello et al., 2016). Thus, users can integrate data obtained from other information sources by leveraging the links between the datasets. Furthermore, with the LOD cloud containing relevant information to extend the feature set and increase granularity (Kuchař, 2015), it could produce new and unique data. So, the newly gained (or inferred) knowledge from LOD is seen as a potential to address the issues of serendipity in recommendations, as discussed by Saat et al. (2018) and Deshmukh et al. (2018).

2.1. DBpedia

DBpedia is a dataset containing Wikipedia data and a project to make it accessible on the Web in a standardized manner. It helps one to query and integrate Wikipedia data with other web sources. Around 4.6 million resources are defined in the English version of DBpedia, with the majority of them (about 4.2 million) classified using the DBpedia ontology. There are about 60 properties for the Film[2] class in DBpedia.

A rich set of semantic data (referring to LOD) is accessible through SPARQL endpoints. DBpedia alone has billions of triples covering a wide range of topics (e.g., people, places, films, and books). Despite the fact that different endpoints may employ different vocabularies and ontologies, they all share a standard platform to access and retrieve semantic data, which is the SPARQL query language. Based on Ferré (2017), composing SPARQL queries is a time-consuming and error-prone task and is primarily inaccessible to most prospective users of semantic data. However, being a W3C standard that is widely utilized, SPARQL's expressivity and scalability for huge RDF stores are two of its advantages. This is likely to have contributed to SPARQL engines that are highly optimized (e.g., Virtuoso, Jena TDB).

There are over 3 billion triples in the DBpedia instance and it is one of the primary interlinking hubs in the LOD (Lehmann et al., 2015). As the data are envisioned for machine-to-machine utilization, however, an acceptable presentation makes them adversely relevant for end-users (Raynaud et al., 2018). Details about users or items are also freely accessible in the LOD cloud in RDF format. This information can be easily retrieved using a SPARQL query endpoint by providing only two types of information: the URI of the required resource and the name of its attribute.

[2]Available at http://mappings.dbpedia.org/server/ontology/classes/Film

 https://doi.org/10.1633/JISTaP.2021.9.2.5

2.1. Recommender Systems

Recommender Systems (RS) delineates a decision-making approach for users on complex information platforms. It can effectively recommend the relevant information to end-users (Zarzour et al., 2018). It is used in our daily lives for various reasons, including the recommendation of products, movies, music, news, and books. These recommendations are based on the user's historical behaviour and other users' behaviour having a similar taste.

The fundamental concept of the RS is based on the premise that if users share the same habits when shopping in online shops, choosing similar artists, reading the same news, or watching the same movies, they are assumed to have similar behaviours in the future. In other words, the RS predicted users' expectations for products and suggested objects so that other users will possibly like them. Currently, recommendation algorithms are widely used in many applications including Amazon, Spotify, Facebook, and TripAdvisor.

The RS can analyze items in two ways: explore items in terms of how users like them or how similar users like the same items, or explore items in terms of how users popularise items, irrespective of the user's similarities. Based on the two fundamental principles of the RS, it can be classified by the main three well-known techniques: Content-based filtering (CB) (Albatayneh et al., 2018; Lops et al., 2019), Collaborative filtering (CF) (Franz et al., 2010; Osman et al., 2020), and Hybrid-based filtering (HB) (Aggarwal, 2016; Jannach et al., 2010). HB technique combines one or more approaches to compensate for the weaknesses of the single methods. Another technique is Knowledge-based (KB), which is inferred between user requirements and items' features described in a certain knowledge base (Dell'Aglio et al., 2010; Ameen, 2019).

A large amount of research has focused on movie recommendation (Jamil et al., 2020; Vilakone et al., 2018), music recommendation (Nguyen et al., 2015; Vall et al., 2019), health recommendation (Nilashi et al., 2020; Stratigi et al., 2020), digital library recommendation (Beel et al., 2016), e-learning (Albatayneh et al., 2018; Pereira et al., 2018), e-commerce recommendation (Chu et al., 2020; Petrova et al., 2019), tourism recommendation (Delic et al., 2018; Yochum et al., 2020), and many other areas.

Recently, Group Recommender Systems (GRS) have become an effective tool for consulting and recommending items according to a group of like-minded users' choices. As the name suggests, a GRS is a system that provides recommendations to a group of users. GRS has been designed, instead of providing one or more items to individuals, to concurrently recommend them to a group with similar interests to satisfy each of them (Nozari & Koohi, 2020). Furthermore, domains such as movies, restaurants, and tourism may typically deal with a group of users. Some works that deal with recommendations to a group of users are Roy et al. (2018), Nawi et al. (2020), and Wang et al. (2020).

2.1.1. Issues in Recommender Systems

Scalability, high computation, diversity (Yadav et al., 2020), and gray sheep (Erion & Maurizio, 2017) are among the concerns when designing an efficient recommendation system. However, above all, two core issues have piqued researchers' interest in the RS area, which are the *cold-start* and *data sparsity* problems. The former occurs during registration of a new user or adding up new resources or items in the system. Apparently, there will be no details in the system regarding the user's interest or preferences for any particular item. In contrast, the latter arises because users usually rate a small portion of the available items. Since there are few ratings, it is uncertain that two users or objects would have similar ratings. As a result, scores are predicted based on a small number of neighbours, which may degrade the recommendation quality since there is insufficient data, or no ratings at all are available (Tomeo et al., 2015). Some works that tackle these issues are by means of item correlation in social networks (Cao et al., 2019; Hong et al., 2015), manipulating user profile demography (Baňas et al., 2015; Xu et al., 2018), enhanced filtering and prediction technique (Mohamed et al., 2019; Ortega et al., 2016), and exploiting external knowledge such as the DBpedia (Meymandpour & Davis, 2015; Di Noia et al., 2016).

3. PRIOR WORK

Several works that address recommendations in numerous domains have been suggested in the literature, even though very few methods take advantage of the LOD initiative to provide efficient recommendations. LOD was often used to mitigate the cold-start and data sparsity issues associated with collaborative recommendation (Vall et al., 2019; Yadav et al., 2020). This section presents the prevalent research work of semantic-based approaches to the RS. As in Table 1, we categorize the previous research on RS applying semantic technology based on both elements of RS and LOD.

LOD-based recommendation research has its emergence in the ground of ontology-based RS pioneered by

Table 1. Application of LOD in Recommender Systems

Study	Domain / Dataset	RS elements	LOD elements
Yadav et al. (2020)	Movie / ML1M + Yahoo! Webscope R4 + DBpedia + FOAF	- New User Cold Start - HB - User profile (social network)	- Item similarity-based ontology - FOAF
Figueroa et al. (2019)	DBpedia + iMDB	- Novelty & accuracy - Heterogeneous information network	- AlLied framework - Simple Knowledge Organization System (SKOS)
Zhao et al. (2019)	Multidomain / Freebase + ML20M + Amazon + LFM-1b	- KB - Heterogeneous information network	- Linkage analysis: popular & recency
Alshammari and Nasraoui (2019)	Multidomain / ML100k + Book crossing + DBpedia	- CF model-based, matrix factorization - Explanation of recommendation	- Semantic knowledge graph - Semantic measure distance: LDSD
Iana et al. (2019)	Conference / SciGraph + WikiCFP	- CB - Recommendation-based (author, keyword, abstract) - Neural network approach	- Computer Science Ontology (CSO)
Nilashi et al. (2018)	Movie / iMDB + ML1M + Yahoo! Webscope R4	- CF model-based (SVD) - Isu scalability and sparseness	- Film ontology (http://www.movieontology.org/) - Semantic similarity between item
Pereira et al. (2018)	Education / Open University	- User profile and interest in education (social network)	- Semantic representation (ontology of FOAF and SIOC) - Semantic search
Srinivasan and Mani (2018)	Movie / ML20M	- topN recommendation - Diversity of recommendation	- Movie based graph (LOD)
Palumbo et al. (2017)	Movie / ML1M	- HB - topN recommendation - Heterogenous information network	- Knowledge graph
Vagliano et al. (2017)	Multidomain / iMDB + LibraryThing + Amazon + DBpedia + Wikidata + Yago + ML1M	- User review - Novelty, diversity recommendation - topN recommendation	- Semantic annotation, AIDA
Musto et al. (2017)	Multidomain / ML1M + DBBook + LastFM	- CF - Graph-based recommendation	- Integrate knowledge from LOD cloud - Feature selection (semantic-aware recommendation)
Piao and Breslin (2016)	Music / DBpedia	- Similarity measure - Vector Space Model	- Semantic measure distance: LDSD
Oliveira et al. (2017)	Movie / DBpedia + LinkedMDB	- Degree of social relation (social network)	- Ontology - Semantic measure distance: LDSD
Di Noia et al. (2016)	Multidomain / ML1M + LibraryThing + LastFM + DBpedia	- CF with topN recommendation - Feature selection, Random Forest model	- Semantic path-based
Meymandpour and Davis (2015)	Film / ML100k + ML1M + DBpedia + LinkedMDB + Freebase + Yago	- HB - Matrix factorization (SVD++)	- Item semantic analysis
Rowe (2014)	Film / MovieTweetings	- SVD++ algorithm	- Semantic aware
Ko et al. (2014)	Mobile TV / MobileIPTV + DBpedia + LinkedMDB + Freebase	- CF - Limited content of the item and user data	- Semantic cluster based on LOD category

 https://doi.org/10.1633/JISTaP.2021.9.2.5

Table 1. Continued

Study	Domain / Dataset	RS elements	LOD elements
Ostuni et al. (2013)	Multidomain / ML1M + LastFM + DBpedia	- HB (CF + CB) - topN recommendation - Random Forests, Gradient Boosted Regression Trees (GBRT)	- Graph semantic - Semantic Vector Space Model (sVSM)
Yang et al. (2013)	Movie / ML100k + DBTropes	- CF - Implicit data - Slope One algorithm	- Implicit item relation LOD-based - Semantic measure distance: LDSD
Mirizzi et al. (2012)	Movie / ML100k + DBpedia + LinkedMDB + Freebase	- CB - Cold start issue - Facebook application	- Semantic similarity between item - sVSM
Bostandjiev et al. (2012)	Music / LastFM + DBpedia + Facebook API + Twitter API	- HB - Explaining aspect (interaction and visualisation)	- Semantic Web API, social Web

LOD, Linked Open Data; RS, Recommendation Systems; ML1M, MovieLens 1 Million; FOAF, Friend of a Friend; SIOC, Semantically-Interlinked Online Communities; HB, hybrid-based filtering; KB, knowledge-based; CF, collaborative filtering; CB, content-based filtering; LDSD, Linked Data Semantic Distance; AIDA, referring to online tool for entity detection and disambiguation.

Middleton et al. (2004). Several research efforts investigate RS based on linked data, and the wealth of data provided by the LOD cloud can be seen in Table 1. In much of the latest literature, authors have exploited DBpedia mainly to extract valuable information. As limited content or information is the core issue in the RS, LOD has significantly played a significant role, and many researchers are taking advantage of using it.

Bostandjiev et al. (2012) developed an application named "*TasteWeights*," which is a kind of RS in which users' preferences for music genres are extracted from Facebook. DBpedia is exploited to discover all of the music played by new artists in the same genre that the active user liked, and then recommends the same to other users. Another DBpedia-based similarity measure has been proposed by Meymandpour and Davis (2015). The authors adopted Partitioned Information Content, a similarity measure inspired by Information Theory. They adapted to LOD's scenario to determine the similarity of two resources defined by LOD properties. The authors have utilized DBpedia extensively to define or modify various similarity measures based on LOD properties (Piao & Breslin, 2016).

More recently, Zhao et al. (2019) present the first public linked KB dataset for RS, named "KB4Rec v1.0," which has linked three commonly used RS datasets with the popular KB Freebase. AlLied, developed by Figueroa et al. (2019), refers to a framework for selecting, evaluating, and creating algorithms to recommend resources from Linked Data (LD) belonging to different application domains. Diversity-Ensured Semantic-aware Item REcommendation (DESIRE), presented by Srinivasan and Mani (2018), deals with a consistent and dependable knowledge source in order to significantly improve quality and provide a diverse topN recommendation list.

The LD can be analyzed through the perspective of its graph structure. The study by Musto et al. (2017), Srinivasan and Mani (2018), Iana et al. (2019), and Ostuni et al. (2013) are among works focusing on graph structure linked-based in RS. LD in graph-based algorithms uses this structure to compute relevance scores for items expressed as nodes in a graph. Semantic exploration with feature selection by Musto et al. (2017) generates semantic-aware recommendations, and the authors studied the impact of the LOD's knowledge on a graph-based recommendation algorithm's overall performance. The results presented in Musto et al. (2017) affirm that knowledge from the LOD cloud can have a significant impact on the recommendation algorithm. In Rowe (2014), the authors developed a semantic-aware extension of the SVD++ model, named "SemanticSVD++." It incorporates semantic item categories into the model. Meanwhile, Ko et al. (2014) apply semantics for clustering television program category and content.

Path-based algorithms compute similarities for producing recommendations using knowledge about semantic paths within a graph structure. Di Noia et al. (2016) introduced "SPRank," which uses LOD to combine machine learning with learning the best path to consider relying on learning to rank for topN recommendation, while Ostuni et al. (2013) demonstrate how to use LOD sources and a

hybrid filtering approach to generate topN recommendations from implicit feedback. Iana et al. (2019) have been recently exploiting SciGraph to provide users with recommendations of conferences to submit their publications to and utilize it for information on past conferences and publications. They also apply WikiCfP for details on upcoming seminars.

Another exciting direction concerning the exploitation of LOD for RS is explored in social networks, as in Mirizzi et al. (2012), Oliveira et al. (2017), and Pereira et al. (2018), by implementing the semantic representation of ontology concept. In the educational domain, Pereira et al. (2018) present an infrastructure able to extract users' profiles and educational context from Facebook, and recommend educational resources. Their study proposes the techniques of information extraction and SW technologies for extraction, enrichment, and definitions of user-profiles, and interests are represented using Friend of a Friend (FOAF). The Semantically-Interlinked Online Communities ontology was used in their study to represent all groups that users join. "Entity2rec," introduced by Palumbo et al. (2017), is an approach to learning user-item relatedness from knowledge graphs for topN item recommendation. They use a knowledge graph via node2vec, encompassing collaborative information from user feedback and item information from LOD.

MORE, introduced in Mirizzi et al. (2012), is a movie recommendation application that works in tandem with Facebook. The remarkable aspect about MORE is that besides analyzing the content, it also considers the user's Facebook profile to overcome the cold-start problem in RS. Oliveira et al. (2016) consume the LOD in RS, adopting the social network and linking to more than one LOD dataset. Their work focuses on identifying related resources from different types, for instance: Books, people, soundtracks, or some other resources considered similar by the algorithm to the user's movie search can be possibly recommended.

Nilashi et al. (2018) tackle scalability and sparseness in RS by proposing a framework implementing film ontology and semantics for item similarity. Meanwhile, Vagliano et al. (2017) combine semantic annotation of user feedback with additional data from the LOD cloud. Yadav et al. (2020) solve the issue of pure new user cold-start in RS by building user profiles based on LOD, and social network-based and collaborative features. The following features are used to calculate user similarity: analyzing the social network, collaborative features from Dense User-Item Matrix, and using the information extracted from the LOD cloud, such as FOAF.

Studies by Meymandpour and Davis (2015), Musto et al. (2014), and Oliveira et al. (2017) use LOD to obtain similarities of items in a CB approach. However, our study focuses on a CF approach with the integration of LOD technologies to overcome sparseness issues. Nilashi et al. (2018) somehow also applied a CF approach in their work. However, they use ontology and dimensionality reduction to tackle sparseness and scalability. Apart from DBpedia, Heitmann and Hayes (2010) considered different sources such as the SmartRadio and Myspace as alternatives to overcome the challenge of the data acquisition problem, as in CF approach. In a greater perspective, we see LOD as an important and influential technological medium applied in RS for various aspects, such as enriching the user and item information.

4. EXPERIMENTAL SETUP

Some of the significant attributes that involve DBpedia in the movie domain are '*dbo:director*,' '*dbo:editor*,' '*dbo:starring*,' '*dbo:producers*' and '*dbo:director*.' We apply two attributes in DBpedia, that the most commonly used in Wikipedia Web pages that portray a movie: '*dbo:starring*' and '*dbo:director*.' These two attributes are relevant in this study because they play an essential role in the selection of movies to watch.

A study by Gmerek (2015) and Carrillat et al. (2018) prove that the starring attribute has a strong correlation with box office movies. According to Rawal and Saavedra (2017), audiences have some set of expectations of a movie produced through its trailers due to the known actors and directors working for the movies. Furthermore, Kim (2013) found that top actors and directors are strongly associated with the success of movies. These arguments support that the director and starring attributes influence a film's commercial success. The same intuition applies when users decide what movies to watch, which is based on the directors and actors. Based on these arguments and findings, it justifies the choice of these two attributes in this study.

Linked data can provide rich semantic representations of the recommended items since they include exciting features (Vagliano et al., 2017). Thus, recommender systems that use linked data should be able to enrich the item or user profiles with LOD sources' information (Sansonetti et al., 2019), particularly from the DBpedia. For example, movies represented in DBpedia contain essential information such as starring actors and directors. As illustrated in Fig. 1, additional information about the actor who starred

 https://doi.org/10.1633/JISTaP.2021.9.2.5

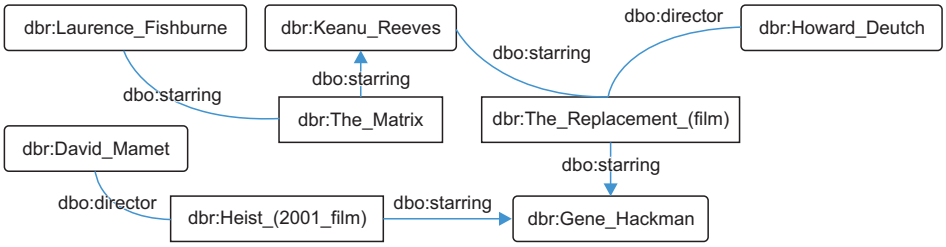


Fig. 1. Link of data relation between attribute through item.

110	*http://dbpedia.org/resource/Braveheart*

Fig. 2. Example of mapping set.

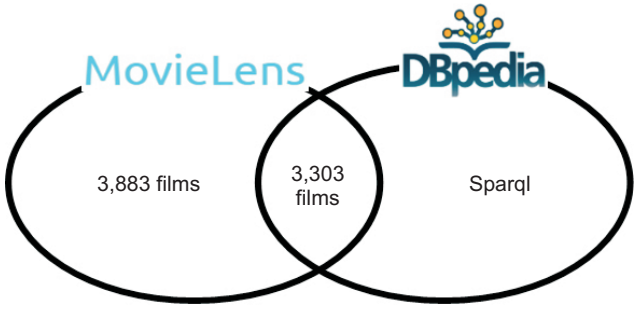


Fig. 3. Mapping MovieLens 1 Million to DBpedia.

in the movie can be explored through the LOD (e.g., the relation '*dbo:starring*' existing between Keanu Reeves and *The Matrix*). This data can then be analyzed and manipulated to enhance recommendation quality based on several aspects, such as semantic analysis of items (Meymandpour & Davis, 2015; Musto et al., 2017), the implicit relationship between items (Ostuni et al., 2013; Yang et al., 2013), search queries (Wenige & Ruhland, 2018), and knowledge graph for item recommendation (Anelli et al., 2020; Palumbo et al., 2017; Piao & Breslin, 2018).

4.1. Mapping MovieLens to DBpedia

To obtain the specific information of resources from the LOD, the URI of such resources are necessary. For that reason, mapping needs to be done in order to identify the required item to the corresponding object in the linked database. It is the only entry point to the LOD. We experimented with two datasets in this study, the ML1M and the DBpedia datasets.

MovieLens is a movie recommendation project created by the University of Minnesota's Department of Computer Science and Engineering. It is a traditional CF system that gathers user movie preferences. Three datasets are available on the MovieLens website. The first one consists of 100,000 ratings (ML100k). The second has around a million ratings (ML1M) for 3,883 movies by 6,040 users, and another one contains 20 million ratings (ML20M).

The mapping set by Meymandpour & Davis (2015)[3] is applied, which provides 3,303 movies' URI. Some other mappings are publicly available, such as those of Kushwaha and Vyas (2014),[4] and Di Noia et al. (2016).[5] Compared to the different mapping sets, Meymandpour and Davis (2015) provide more stable and consistent mappings. The example mapping for the '*Braveheart*' film is shown in Fig. 2, and note that the number 110 is referring to the 'movie_id' of the '*Braveheart*' film in the ML1M dataset.

The URI serves as an identifier for that data related to the associated movie. It can be used to extract additional information. ML1M items have been assigned to the corresponding DBpedia entities, and as mentioned earlier we leverage publicly available mappings. It is derived by running a SPARQL query based on the item's descriptive properties (e.g., the name of the movie). Since not every item in the ML1M data has a corresponding DBpedia entity, out of the 3,883 movies from ML1M, 580 URIs are missing from the mappings. We can say that it happened to reduce 580 films, as illustrated in Fig. 3, thus indicating that we obtained 964,534 ratings from 6,040 users on 3,303 items after this mapping.

Fragment mapping is a function that maps fragments to a set of endpoints, u, and refers to DBpedia in this study. A fragment f correlates to a fragment description, $fd(f)=<u,tp>$, while a triple pattern (tp), $tp;triples(f)$ correspond to the RDF triples of f that fulfill $fd(f)$. The dataset's tp can be accessed using $fd(f).u$. The sample of fragment description can be referred to in Fig. 4. However, the mapping for a film fragment is missing due to some retrieval

[3]DBpedia-MovieLens_Benchmark_Datasets_100K_1M
[4]https://www.researchgate.net/publication/297369577_mapping-movielens-dbpedia
[5]https://github.com/sisinflab/LODrecsys-datasets/tree/master/Movielens1M

of film unavailability attributes in DBpedia. For example, the fragment can retrieve the data from the '*dbo:director*' attribute but no output for '*dbo:starring*.' Some of the explanations are discovered through the issue addressed in Section 5.

4.2. SPARQL Code

LOD features are then extracted from the DBpedia SPARQL endpoint in this study. The following URI pattern is used to access each resource in DBpedia:

http://dbpedia.org/resource/<name_of_resource>

```
triple(f):          {dbr:The_Replacements_(film)
                    dbo:starring  dbr:Gene_Hackman
                    dbr:Heist_(2001_film)
                    dbo:director  dbr:David_Mamet
                    dbr:The_Matrix
                    dbo:starring  dbr:Laurence_Fishburne, ...}
```

Fig. 4. Fragment description.

```
"""SELECT DISTINCT ?label ?abstract ?director ?starring WHERE
{
        <http://dbpedia.org/resource/Braveheart>
        rdfs:label      ?label ;
        dbo:abstract  ?abstract ;
        dbo:director   ?director ;
        dbo:starring   ?starring
        FILTER ( lang(?label) = "en" )
        FILTER ( lang(?abstarct) = "en" )
}"""
```

Fig. 5. SPARQL code for data extracting from DBpedia.

SPARQL code applies in extracting the data as can be referred to in Fig. 5. DBpedia's two properties are considered as mentioned before: '*dbo:director*' and '*dbo:starring*.' Based on this method, we can get the '*Braveheart*' film director and starring actors. Moreover, we can get the related films directed by the same director based on the SPARQL code (Fig. 6), referring to '*dbo:director={mel_gibson}*' matched to the URI *http://dbpedia.org/resource/Mel_Gibson*.

Fig. 7 illustrates this interconnection of the movie from ML1M to the DBpedia based on the director's attribute. Besides this, the related film with the same director could be figured out. A list of movies directed by '*dbo:director={Mel_Gibson}*' other than '*Braveheart*,' includes '*Apocalypto*,' '*Hacksaw Ridge*,' '*The Man Without a Face*' and '*The Passion of the Christ*.'

5. CHALLENGES IN EXTRACTING PROCESS

Although the LOD cloud has indeed been augmented with a large amount of information related to various domains, they still contain inconsistencies, inaccuracy, and missing data problems (Kushwaha & Vyas, 2014). One

```
"""SELECT DISTINCT ?person ?name ?about
WHERE {
  <http://dbpedia.org/resource/Braveheart> dbo:director ?directBy.
  ?person dbo:director ?directBy.
  ?person foaf:name ?name.
  ?person dbo:abstract ?about.
  FILTER (lang(?about) = "en")
} ORDER BY ?name"""
```

Fig. 6. SPARQL code for films with same directors.

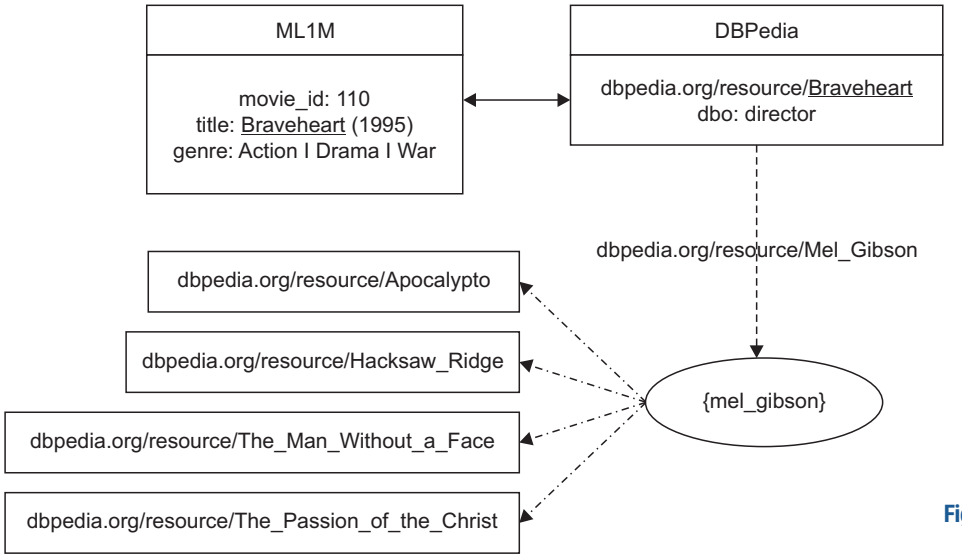


Fig. 7. Data relation for 'Braveheart' film. ML1M, MovieLens 1 Million.

 https://doi.org/10.1633/JISTaP.2021.9.2.5

of the challenging issues in the LOD is data quality. Data quality concerns issues such as inaccuracy, incompleteness, and inconsistency, which entail significant restrictions on the data's optimum utilization. Zaveri et al. (2014) also revealed that data derived from semi-structured or even structured sources frequently contains discrepancies and misrepresentations, as well as incomplete information. Despite that the mapping set is stable and consistent compared to others, we faced some challenges in extracting DBpedia's data process. The discussion is classified into seven issues, as follows.

5.1. Incomplete URI

As shown in Fig. 3, the mapping set provides 3,303 URI for the films out of 3,883, as discussed in Section 4.1. There are some situations where, even though the mapping provides the URI, hardly any result is displayed. Some examples are from the links (films) of '*Champagne* (*http://dbpedia.org/resource/Champagne_(film)*)' and '*Downhill* (*http://dbpedia.org/resource/Downhill_(film)*).' Most probably, it does not come with the appropriate URI link. Based on the SPARQL code of getting films with the same directors (Fig. 6), it resulted that both movies come from the same director of '*Albert Hitchcock*.' We also figure out, based on the same SPARQL code, that the supposed URI that should display the output is (*http://dbpedia.org/resource/Champagne_(1928_film)*) for the '*Champagne*' film and (*http://dbpedia.org/resource/Downhill_(1927_film)*) refers to the '*Downhill*' film.

5.2. Contradiction in the Output Fragment

There exists a contradiction of the output fragment between the attributes and the abstracts of the movies. Some of the output provides one or two actors only. However, referring to the abstract, there are more actors compared to the given output. For example, the film '*Raining Stones* (*http://dbpedia.org/resource/Raining Stones*)' produced starring results with only one actor, '*Bruce Jonas*,' while in abstract details, three actors are displayed: '*Bruce Jonas*,' '*Ricky Tomlinson*,' and '*Tom Hickey*.' Another case in point is the film '*The_Abominable_Snowman (http://dbpedia.org/resource/The_Abominable_Snowman_(film))*.'

5.3. Mappings of Different/Wrong Movies

Incorrect mappings were also found in the mapped dataset. For example, the mapping set for the movie_id of 97, '*Hate (Haine, La) (1995)*' is mapped to a different movie from ML1M. It is mapped to the URI of (*http://dbpedia.org/resource/Whatever_Will_Be,_Will_Be_(1995_film)*), that refers to a 1995 Hong Kong drama film directed by '*Jacob Cheung*,' which starred '*Aaron Kwok*' and '*Kelly Chen*.' In contrast, '*Mathieu Kassovitz*' wrote, co-edited, and directed '*Hate (Haine, La) (1995),*' a French black and white drama film. These two films differ considerably based on the dissimilar given abstracts.

5.4. Actors with Different Names

Another challenge is that some actors have different names in some films, even if it refers to the identical actor. For example, some movies display the data of '*Jada Pinkett*' for the starring information, and some others showed '*Jada Pinkett-Smith*.' Other actors who have similar situations are '*Joanne Whalley/JoanneWhalley-Kilmer*' and '*Tisha Campbell/Tisha Campbell-Martin*.'

5.5. Brotherhood Film Direction

Some movies use the word '*Brothers*' for directors to indicate that the movies are being directed by two or more siblings. '*Hughes Brothers (Albert and Allen Hughes)*,' '*Farrely Brother (Peter* and *Bobby Farrelly)*,' and '*Coen Brothers (Joel and Ethan Coen)*' are some of the example of movies that use the term '*Brothers*' for their directed films. The issue is imminent if it even happened when only one of the brothers directs another film and applies the individual name. For example, '*Dumb and Dumber (1994*)' is directed by '*Peter Farrely*' from the '*Farrely Brothers*.' In contrast, '*Bobby Farrely*' is not involved in the directing. The term '*Brothers*' will be used if both of them direct the same movie. This issue will be affected by the number of the films being directed by a particular director. As a solution, the co-directed films are divided such that the same movies are linked to both individual directors.

5.6. Confusing Output

Confusion exists when the mappings between entities do not make sense. For example, the 'movie_id = 1547' which is mapped to the film '*Shiloh* (*http://dbpedia.org/resource/Shiloh_(film)*)' exhibits a confusing director. The director and abstract information's data are dissimilar: '*Chip Rosenblum*' and '*Dale Rosenblum*.' The director's output shows '*Chip Rosenblum*,' whereas the abstract information reveals '*Dale Rosenblum*.' This scenario requires additional searching information to get the actual data and the result shows '*Dale "Chip" Rosenblum*' as the commercial name. For that issue, we take the '*Dale Rosenblum*' as the director.

Another issue is where the actor list's output does not

properly reflect the film's main cast. It can be clearly associated as some famous actor recognized for portraying a particular character in some films, especially those with sequels. In order to achieve more tangible assurance we also should adhere to the abstract. For instance, the movie '*Mission Impossible II (http://dbpedia.org/resource/Mission:_Impossible_II)*' has four main casts, but does not list the actor '*Tom Cruise*' in the starring output. It is advantageous to attempt retrieving the films' full four main casts, since it will increase the possibility of similarity of a particular item with known actors.

There are also cases where the output displayed for specific attributes does not coincide with the abstract information. An example is '*The Eye of Tammy Faye*' *(http://dbpedia.org/resource/The_Eyes_of_Tammy_Faye).*' According to the abstract detail, there is a film director named '*Fenton Bailey*'. The results for the director attribute, on the other hand, show two directors, '*Randy Barbato*' and '*Fenton Bailey*'. In situations like this, we prioritise the output attribute display.

Other than that, actors with very similar names may also be disconcerting. For cases of nearly identical name ambiguity with different people, data number errors may occur. An example is a name between '*Elaine Stritch*' and '*Elaine Strich*'. Other examples are '*Tom Skerritt*' and '*Tom Skeritt*', and '*Joseph Cotton*' and '*Joseph Cotten*'.

5.7. Special Characters in URI

Some URIs are accompanied by special character symbols representing a text, which could justify why output does not display. Examples can be seen from the URIs of:

(http://dbpedia.org/resource/Who's_Harry_Crumb%3F)
and
(http://dbpedia.org/resource/Cl%C3%A9o_from_5_to_7).

The former refers to the film '*Who's Harry Crumb?*' from the 'movie_id=3387' in ML1M, while the origin film's name for the latter URI is '*Cleo From 5 to 7*'. One of the reasons is that the film's name is not in English or uses non-English symbols and characters. This mapping problem prevents data from being produced. To obtain the actual movie title, we would then have to refer back to the ML1M data set through the mapping set's movie id.

6. DISCUSSION

All of the above-mentioned challenges may impact the data analysis relating to the numbers of directors and actors, and thus will directly affect the process of determining the uniformity and consistency of the data. Therefore, pre-processing involving cleaning and filtering to remove unnecessary data as necessary.

We should also note that the cardinality of attributes may vary. For example, a film may have only one director but ten actors. It is expected that the '*dbo:starring*' attribute values will dominate the encoded vector of a movie as compared to the '*dbo:director.*' However, based on the extracting process some films also provide more than one director as the output, especially for films belonging to the animation genre. As such, we retain the minimum one value when data exists and increase it to a maximum of 4 using the symbol '|' as its separator. The data can be referred to Table 2.

As a result of the aforementioned challenges, a clean additional information of the movie data as illustrated in Table 2 is being provided, namely ML1M-DBpedia. It thus contributes in resolving some of the issues discussed previously. Eventually, the data can then be exploited to overcome the issues of sparsity in RS. With sparse ratings, it is unlikely that two users or objects would have similar ratings. As a result, scores are predicted for a limited number of neighbours. Since sparsity has an effect on identifying item similarities, it has an impact on achieving higher predictive accuracy and recommendation relevance.

For example, the data with ratings given by users in the ML1M can then be analyzed and manipulated to enrich it. For example, the data relating to the relation of users and items for the two films as can be seen in Fig. 8, if further analysed, will result in the recommendation of the film '*Braveheart*' since the user likes the film '*The Man without a Face*', which was directed by the same director. Meanwhile, Table 3 shows the data for the item related to ML1M and DBpedia based on the '*dbo:director*' attri-

Table 2. Additional data for '*dbo:director*' and '*dbo:starring*'

movie_id	dbo:director	dbo:starring
1	John Lasseter	Tom Hanks|Tim Allen
2	Joe Johnston	Robin Williams|Kirsten Dunst|Bradley Pierce|David Alan Grier
3	Howard Deutch	Jack Lemmon|Walter Matthau|Ann-Margret|Sophia Loren

 https://doi.org/10.1633/JISTaP.2021.9.2.5

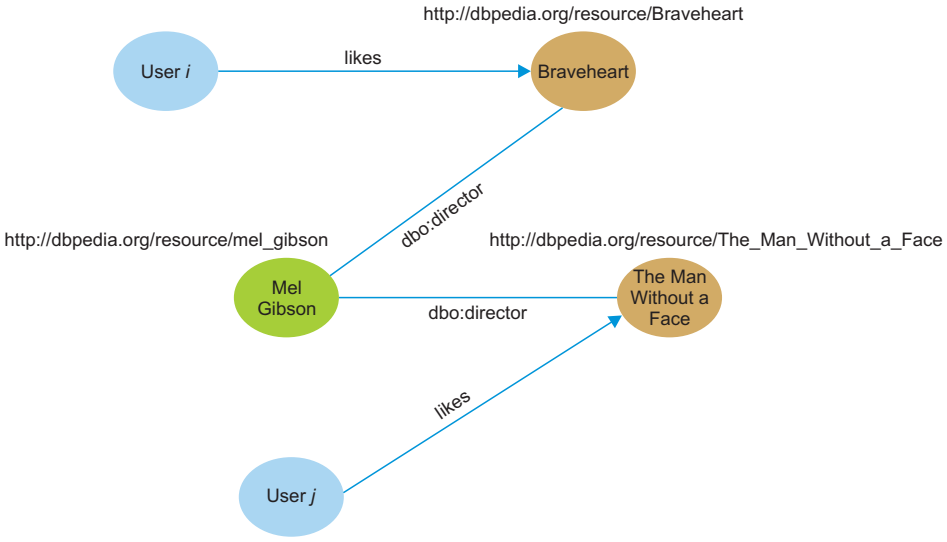


Fig. 8. Data relation of user and item based on '*dbo:director*' attribute.

Table 3. Data of users and ratings based on the relation for the '*dbo:director*' attributes

DBpedia	Film	ML1M	
		User (*U*)	Rating (*R*)
dbo: director {mel_gibson}	Braveheart (movie_id:110) (dbpedia.org/resource/Braveheart)	2	5
		7	5
		8	5
		10	4
		11	3
	The Man Without A Face (movie_id:491) (dbpedia.org/resource/The_Man_Without_a_Face)	26	3
		123	3
		133	4
		148	3
		175	4

Each film display for the five initial ratings only.
ML1M, MovieLens 1 Million.

bute to generate a recommendation. It shows the results of additional information relating to '*dbo:director*' and '*dbo:starring*' extracted from DBpedia and mapped to the *movie_ID* in the ML1M dataset. Given that the recommended items are linked to the LOD dataset, its information may be expanded to ascertain the items considered similar to those used in the past. This rich information can then be exploited to tackle the sparsity issue, by providing new predicted ratings based on the particular attribute. It would collaboratively be helpful for the system to provide more relevant and high-quality recommendations as data sparsity is decreased.

7. CONCLUSION

The usage of LOD data poses new challenges and issues in developing the next-generation RS and, more generally, complex web applications. Several studies on the LOD in RS have been steadily arising over the last few years in different aspects.

This paper has presented collaborative RS that leverage the knowledge encoded in the LOD resources. In particular, since the focus of this study was on the movie domain, DBpedia was exploited to obtain more information about movies, such as actors and directors. Furthermore, we en-

countered and highlighted some of the issues encountered throughout the mapping and extraction of data from the ML1M dataset and the DBpedia resource. Such issues can benefit future research in this area, particularly the cleaning and organising knowledge extracted from similar resources. Thus, resolving such issues may assist data providers in discovering and correcting errors by considering all findings given so that any violations can be fixed before further exploration. Although the solutions to these issues and challenges have been proposed, further work relating to evaluation in terms of the quality and accuracy of the mapped data is necessary and thus becomes the near future work of this study.

Other future work includes integrating to other LOD data sources such as Linked MDB and Freebase. The LOD initiative sets the criteria for cross-domain interoperability and has accumulated vast volumes of knowledge in recent years. It provides various ways by which RS performance can be enhanced by enriching the user's profile with related features.

Furthermore, the refinement of the LOD's role in the recommendation process to a group of users is worth exploring. The cold start problem that arises in individual recommender systems is also being addressed in a group recommender system that collectively recommends items to a group of individuals based on their preferences (Dara et al., 2020). Therefore, it is interesting to investigate the potential of using the wealth of relations embedded within the LOD resources so as to produce a more accurate and diversified recommendation to a group of users.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the sponsorship received to carry out this study from Tun Hussein Onn University of Malaysia and from the Malaysia Ministry of Higher Education.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Aggarwal, C. C. (2016). *Recommender systems: The textbook*. Springer.

Al-Barznji, K., & Atanassov, A. (2017, October 4-6). Collaborative filtering techniques for generating recommendations on big data. In V. Sgurev, K. Boyanov, M. Hadjiski, K. Stoilova, K. Boshnakov, E. Nikolov, & V. Andreev (Eds.), *Proceedings of the International Conference Automatics and Informatics' 2017* (pp. 225-228). John Atanasoff Society of Automatics and Informatics.

Albatayneh, N. A., Ghauth, K. I., & Chua, F.-F. (2018). Utilizing learners' negative ratings in semantic content-based recommender system for e-learning forum. *Journal of Educational Technology & Society*, 21(1), 112-125. https://www.jstor.org/stable/26273873.

Alshammari, M., & Nasraoui, O. (2019, September 17-19). Augmented semantic explanations for collaborative filtering recommendations. In A. Fred, & J. Filipe (Eds.), *Proceedings of the IC3K 2019: 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (pp. 83-88). SciTePress.

Ameen, A. (2019). Knowledge based recommendation system in semantic web - a survey. *International Journal of Computer Applications*, 182(43), 20-25. https://doi.org/10.5120/ijca2019918538.

Anelli, V. W., Deldjoo, Y., Di Noia, T., Di Sciascio, E., & Merra, F. A. (2020, May 31-June 4). SAShA: Semantic-aware shilling attacks on recommender systems exploiting knowledge graphs. In A. Harth, S. Kirrane, A.-C. N. Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, & M. Cochez (Eds.), *Proceedings of the 17th International Conference, ESWC 2020* (vol. 12123, pp. 307-323). Springer. https://doi.org/10.1007/978-3-030-49461-2_18

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, & P. Cudré-Mauroux (Eds.), *Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007* (vol 4825, pp. 722-735). Springer.

Baňas, D., Havrilová, C., & Paralič, J. (2015, September 3-5). Combination of user profile information and collaborative filtering in recommendations. In A. Szakal (Ed.), *Proceedings of the 2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES)* (pp. 359-363). IEEE.

Barros, M., Moitinho, A., & Couto, F. M. (2020, April 14-17). Hybrid semantic recommender system for chemical compounds. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, & F. Martins (Eds.), *Proceedings of the 42nd European Conference on IR Research, ECIR 2020* (vol. 12036, pp. 94-101). Springer.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *Interna-*

tional Journal on Digital Libraries, 17(4), 305-338. https://doi.org/10.1007/s00799-015-0156-0.

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012, September 9-13). TasteWeights: A visual interactive hybrid recommender system. In P. Cunningham, N. Hurley, I. Guy, & S. S. Anand (Eds.), *Proceedings of the RecSys '12: 6th ACM Conference on Recommender Systems* (pp. 35-42). Association for Computing Machinery.

Cao, D., He, X., Miao, L., Xiao, G., Chen, H., & Xu, J. (2019). Social-enhanced attentive group recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 33(3), 1195-1209. https://doi.org/10.1109/TKDE.2019.2936475.

Cappiello, C., Di Noia, T., Marcu, B. A., & Matera, M. (2016, June 6-9). A quality model for linked data exploration. In A. Bozzon, P. Cudre-Maroux, & C. Pautasso (Eds.), *Proceedings of the 16th International Conference on Web Engineering*, *ICWE 2016* (vol. 9671, pp. 397-404). Springer.

Carrillat, F. A., Legoux, R., & Hadida, A. L. (2018). Debates and assumptions about motion picture performance: A meta-analysis. *Journal of the Academy of Marketing Science*, 46(2), 273-299. https://doi.org/10.1007/s11747-017-0561-6.

Chu, P.-M., Mao, Y.-S., Lee, S.-J., & Hou, C.-L. (2020). Leveraging user comments for recommendation in e-commerce. *Applied Sciences*, 10(7), 2540. https://doi.org/10.3390/app10072540.

Dara, S., Chowdary, C. R., & Kumar, C. (2020). A survey on group recommender systems. *Journal of Intelligent Information Systems*, 54(2), 271-295. https://doi.org/10.1007/s10844-018-0542-3.

Delic, A., Neidhardt, J., Nguyen, T. N., & Ricci, F. (2018). An observational user study for group recommender systems in the tourism domain. *Information Technology & Tourism*, 19(1-4), 87-116. https://doi.org/10.1007/s40558-018-0106-y.

Dell'Aglio, D., Celino, I., & Cerizza, D. (2010, November 8). Anatomy of a semantic web-enabled knowledge-based recommender system. In A. Bernstein, P. Grace, M. Klusch, & M. Paolucci (Eds.), *Proceedings of the 4th International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web (SMR2 2010)* (vol. 667, pp. 115-130). CEUR Workshop Proceedings.

Deshmukh, A. A., Nair, P., & Rao, S. (2018, November 17-20). A scalable clustering algorithm for serendipity in recommender systems. In H. Tong, Z. Li, F. Zhu, & J. Yu (Eds.), *Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 1279-1288). IEEE.

Di Noia, T., Mirizzi, R., Ostuni, V. C., Romito, D., & Zanker, M. (2012, September 5-7). Linked open data to support content-based recommender systems. In H. Sack, & L. T. Pellegrini (Eds.), *Proceedings of the I-SEMANTICS 2012: 8th International Conference on on Semantic Systems* (pp. 1-8). Association for Computing Machinery.

Di Noia, T., Ostuni, V. C., Tomeo, P., & Di Sciascio, E. (2016). SPrank: Semantic path-based ranking for top-N recommendations using linked open data. *ACM Transactions on Intelligent Systems and Technology*, 8(1), 9. https://doi.org/10.1145/2899005.

Erion, Ç., & Maurizio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, 21(6), 1487-1524. https://doi.org/10.3233/IDA-163209.

Feng, X., Zhang, H., Ren, Y., Shang, P., Zhu, Y., Liang, Y., Guan, R., & Xu, D. (2019). The deep learning–based recommender system "Pubmender" for choosing a biomedical publication venue: Development and validation study. *Journal of Medical Internet Research,* 21(5), e12957. https://doi.org/10.2196/12957.

Ferré, S. (2017). SPARKLIS: An expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web*, 8(3), 405-418. https://doi.org/10.3233/SW-150208.

Figueroa, C., Vagliano, I., Rocha, O. R., Torchiano, M., Zucker, C. F., Corrales, J. C., & Morisio, M. (2019). Executing, comparing, and reusing linked-data-based recommendation algorithms with the allied framework. In M. Lytras, N. Aljohani, E. Damiani, & K. Chui (Eds.), *Semantic Web science and real-world applications* (pp. 18-47). IGI Global.

Franz, T., Koch, J., Dividino, R., & Staab, S. (2010, March 22-24). LENA-TR: Browsing linked open data across knowledge-aspects. In D. Brickley, V. K. Chaudhri, H. Halpin, & D. McGuinness (Eds.), *Proceedings of the AAAI 2010 Spring symposium* (Technical Report SS-10-07, pp. 46-51). AAAI Press.

Gmerek, N. (2015). The determinants of Polish movies' box office performance in Poland. *Journal of Marketing and Consumer Behaviour in Emerging Markets*, 1(1), 15-35. https://doi.org/10.7172/2449-6634.jmcbem.2015.1.2.

Hammou, B. A., Lahcen, A. A., & Mouline, S. (2019). A distributed group recommendation system based on extreme gradient boosting and big data technologies. *Applied Intelligence*, 49(12), 4128-4149. https://doi.org/10.1007/s10489-019-01482-9.

Heitmann, B., & Hayes, C. (2010, March 22-24). Using linked data to build open, collaborative recommender systems. In D. Brickley, V. K. Chaudhri, H. Halpin, & D. McGuinness (Eds.), *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium 2010: Linked Data Meets Artificial Intelligence* (pp. 76-81). AAAI

Publications.

Hidayat, R., Yahya, Y., Mohd Noah, S. A., Ahmad, M. Z., & Hamdan, A. R. (2012). Semantic web portal in university research community framework. *International Journal on Advanced Science, Engineering and Information Technology*, 2(6), 449-453. https://doi.org/10.18517/ijaseit.2.6.242.

Hong, M., Jung, J. J., & Lee, M. (2015, November 26-27). Social affinity-based group recommender system. In P. C. Vinh, & V. Alagar (Eds.), *Proceedings of the 4th International Conference, ICCASA 2015* (vol. 165, pp. 111-121). Springer.

Iana, A., Jung, S., Naeser, P., Birukou, A., Hertling, S., & Paulheim, H. (2019, September 9-12). Building a conference recommender system based on SciGraph and WikiCFP. In M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, & Y. Sure-Vetter (Eds.), *Proceedings of the 15th International Conference on Semantic Systems, SEMANTiCS 2019* (vol. 11702, pp. 117-123). Springer.

Jamil, N., Mohd Noah, S. A., & Mohd, M. (2020). Collaborative item recommendations based on friendship strength in social network. *International Journal of Machine Learning and Computing*, 10(3), 437-443. https://doi.org/10.18178/ijmlc.2020.10.3.954.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: An introduction*. Cambridge University Press.

Kim, M. H. (2013). Determinants of revenues in the motion picture industry. *Applied Economics Letters*, 20(11), 1071-1075. https://doi.org/10.1080/13504851.2013.783678.

Ko, H., Kim, E., Ko, I.-Y. & Chang, D. (2014, January 15-17). Semantically-based recommendation by using semantic clusters of users' viewing history. In S.-M. Choi, & J.-C. Chen (Eds.), *Proceedings of the 2014 International Conference on Big Data and Smart Computing (BIGCOMP)* (pp. 83-87). IEEE.

Kuchař, J. (2015, August 2-5). Augmenting a feature set of movies using linked open data. In N. Bassiliades, P. Fodor, A. Giurca, G. Gottlob, T. Kliegr, G. J. Nalepa, M. Palmirani, A. Paschke, M. Proctor, D. Roman, F. Sadri, & N. Stojanovic (Eds.), *Proceedings of the RuleML 2015 Challenge, the Special Track on Rule-based Recommender Systems for the Web of Data* (vol. 1417, pp. 1-9). CEUR Workshop Proceedings.

Kushwaha, N., & Vyas, O. P. (2014, October 9-11). SemMovieRec: Extraction of semantic features of DBpedia for recommender system. In P. Bhattacharya, P. J. Narayanan, & S. Padmanabhuni (Eds.), *Proceedings of the COMPUTE '14: 7th ACM India Computing Conference* (pp. 1-9). Association for Computing Machinery.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2015). DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167-195. https://doi.org/10.3233/SW-140134.

Lops, P., Jannach, D., Musto, C., Bogers, T., & Koolen, M. (2019). Trends in content-based recommendation. *User Modeling and User-Adapted Interaction*, 29(2), 239-249. https://doi.org/10.1007/s11257-019-09231-w.

Maillo, J., Ramírez, S., Triguero, I., & Herrera, F. (2017). kNN-IS: An iterative spark-based design of the k-nearest neighbors classifier for big data. *Knowledge-Based Systems*, 117, 3-15. https://doi.org/10.1016/j.knosys.2016.06.012.

Meymandpour, R., & Davis, J.G. (2015, January 27-30). Enhancing recommender systems using linked open data-based semantic analysis of items. In J. G. Davis, & A. Bozzon (Eds.), *Proceedings of the 3rd Australasian Web Conference (AWC 2015)* (vol. 166, pp. 11-17). Australian Computer Society.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1), 54-88. https://doi.org/10.1145/963770.963773.

Mirizzi, R., Di Noia, T., Ragone, A., Ostuni, V. C., & Di Sciascio, E. (2012). Movie recommendation with DBpedia. *IIR*, 101-112. https://dblp.org/db/conf/iir/iir2012.html#MirizziNROS12.

Mohamed, M. H., Khafagy, M. H., Elbeh, H., & Abdalla, A. M. (2019). Sparsity and cold start recommendation system challenges solved by hybrid feedback. *International Journal of Engineering Research and Technology*, 12(12), 2735-2742. https://www.ripublication.com/irph/ijert19/ijertv12n12_87.pdf.

Musto, C., Basile, P., Lops, P., de Gemmis, M., & Semeraro, G. (2014, October 6). Linked open data-enabled strategies for top-N recommendations. In T. Bogers, M. Koolen, & I. Cantador (Eds.), *Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems co-located with the 8th ACM Conference on Recommender Systems, CBRecSys@RecSys 2014* (vol. 1245, pp. 49-55). CEUR Workshop Proceedings.

Musto, C., Basile, P., Lops, P., de Gemmis, M., & Semeraro, G. (2017). Introducing linked open data in graph-based recommender systems. *Information Processing & Management*, 53(2), 405-435. https://doi.org/10.1016/j.ipm.2016.12.003.

Nawi, R. M., Mohd Noah, S. A., & Zakaria, L. Q. (2020). Evaluation of group modelling strategy in model-based collaborative filtering recommendation. *International Journal of Machine Learning and Computing*, 10(2), 330-338. https://

doi.org/10.18178/ijmlc.2020.10.2.939.

Nguyen, P. T., Tomeo, P., Di Noia, T., & Di Sciascio, E. (2015, October 11-15). Content-based recommendations via DBpedia and freebase: A case study in the music domain. In M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, K. Thirunarayan, & S. Staab (Eds.), *Proceedings of the 14th International Semantic Web Conference* (vol. 9366, pp. 605-621). Springer.

Nilashi, M., Asadi, S., Abumalloh, R. A., Samad, S., & Ibrahim, O. (2020). Intelligent recommender systems in the COVID-19 outbreak: The case of wearable healthcare devices. *Journal of Soft Computing and Decision Support Systems*, 7(4), 8-12. https://www.jscdss.com/index.php/files/article/view/233/pdf_286.

Nilashi, M., Ibrahim, O., & Bagherifard, K. (2018). A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. *Expert Systems with Applications*, 92, 507-520. https://doi.org/10.1016/j.eswa.2017.09.058.

Nozari, R. B., & Koohi, H. (2020). A novel group recommender system based on members' influence and leader impact. *Knowledge-Based Systems*, 205, 106296. https://doi.org/10.1016/j.knosys.2020.106296.

Oliveira, J., Delgado, C., & Assaife, A. C. (2017). A recommendation approach for consuming linked open data. *Expert Systems with Applications*, 72, 407-420. https://doi.org/10.1016/j.eswa.2016.10.037.

Ortega, F., Hernando, A., Bobadilla, J., & Kang, J. H. (2016). Recommending items to group of users using matrix factorization based collaborative filtering. *Information Sciences*, 345, 313-324. https://doi.org/10.1016/j.ins.2016.01.083.

Osman, N. A., Noah, S. A. M., & Darwich, M. (2020). Contextual sentiment based recommender system to provide recommendation in the electronic products domain. *International Journal of Machine Learning and Computing*, 9(4), 425-431. https://doi.org/10.18178/ijmlc.2019.9.4.821.

Ostuni, V. C., Di Noia, T., Di Sciascio, E., & Mirizzi, R. (2013, October 12-16). Top-N recommendations from implicit feedback leveraging linked open data. In Q. Yang, I. King, Q. Li, & P. Pu (Eds.), *Proceedings of the RecSys '13: 7th ACM conference on Recommender systems* (pp. 85-92). Association for Computing Machinery.

Palumbo, E., Rizzo, G., & Troncy, R. (2017, August 27-31). entity2rec: Learning user-item relatedness from knowledge graphs for top-N item recommendation. In P. Cremonesi, F. Ricci, S. Berkovsky, & A. Tuzhilin (Eds.), *Proceedings of the RecSys '17: 11th ACM Conference on Recommender Systems* (pp. 32-36). Association for Computing Machinery.

Peralta, V. (2007). Extraction and integration of MovieLens and IMDb data, Technical Report. *Accès Personnalisé à des Masses de Données*. https://www.researchgate.net/publication/228429288_Extraction_and_Integration_of_MovieLens_and_IMDb_Data.

Pereira, C. K., Campos, F., Ströele, V., David, J. M. N., & Braga, R. (2018). BROAD-RSI – educational recommender system using social networks interactions and linked data. *Journal of Internet Services and Applications*, 9, 7. https://doi.org/10.1186/s13174-018-0076-5.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R. L. (2019, July 10-12). Semantic data mining and linked data for a recommender system in the AEC industry. In J. O'Donnell, A. Chassiakos, D. Rovas, & D. Hall (Eds.), *Proceedings of the 2019 European Conference on Computing in Construction* (pp. 172-181). European Council on Computing in Construction.

Piao, G., & Breslin, J. G. (2016, April 4-8). Measuring semantic distance for linked open data-enabled recommender systems. In S. Ossowski (Ed.), *Proceedings of the SAC '16: 31st Annual ACM Symposium on Applied Computing* (pp. 315-320). Association for Computing Machinery.

Piao, G., & Breslin, J. G. (2018, June 3-7). Transfer learning for item recommendations and knowledge graph completion in item related domains via a co-factorization model. In A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, & M. Alam (Eds.), *Proceedings of the 15th International Conference*, *ESWC 2018* (vol. 10843, pp. 496-511). Springer.

Rawal, M., & Saavedra, J. L. (2017). Mediating effect of word-of-mouth in movie theatre industry. *Journal of Media and Communication Studies*, 9(3), 17-23. https://doi.org/10.5897/JMCS2016.0529.

Raynaud, T., Subercaze, J., & Laforest, F. (2018, October 22-26). Fouilla: Navigating DBpedia by topic. In A. Cuzzocrea, J. Allan, N. Paton, D. Srivastava, R. Agrawal, A. Broder, M. Zaki, S. Candan, A. Labrinidis, A. Schuster, & H. Wang (Eds.), *Proceedings of the CIKM '18: 27th ACM International Conference on Information and Knowledge Management* (pp. 1907-1910). Association for Computing Machinery.

Rowe, M. (2014, August 11-14). SemanticSVD++: Incorporating semantic taste evolution for predicting ratings. In D. Ślęzak, B. Dunin-Kęplicz, M. Lewis, & T. Terano (Eds.), *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (pp. 213-220). IEEE.

Roy, A., Banerjee, S., Bhatt, C., Badr, Y., & Mallik, S. (2018). Hybrid group recommendation using modified termite colony

algorithm: A context towards big data. *Journal of Information & Knowledge Management*, 17(2), 1850019. https://doi.org/10.1142/S0219649218500193.

Saat, N. I. Y., Noah, S. A. M., & Mohd, M. (2018). Towards serendipity for content–based recommender systems. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4-2), 1762-1769. https://doi.org/10.18517/ijaseit.8.4-2.6807.

Sankar, A., Zhang, X., Krishnan, A., & Han, J. (2020, February 3-7). Inf-VAE: A variational autoencoder framework to integrate homophily and influence in diffusion prediction. In J. Caverlee, X. B. Hu, M. Lalmas, & W. Wang (Eds.), *Proceedings of the WSDM '20: 13th International Conference on Web Search and Data Mining* (pp. 510-518). Association for Computing Machinery.

Sansonetti, G., Gasparetti, F., Micarelli, A., Cena, F., & Gena, C. (2019). Enhancing cultural recommendations through social and linked open data. *User Modeling and User-Adapted Interaction*, 29(1), 121-159. https://doi.org/10.1007/s11257-019-09225-8.

Srinivasan, U. S., & Mani, C. (2018). Diversity-ensured semantic movie recommendation by applying linked open data. *International Journal of Intelligent Engineering & Systems*, 11(2), 275-286. https://doi.org/10.22266/ijies2018.0430.30.

Stratigi, M., Kondylakis, H., & Stefanidis, K. (2020). Multidimensional group recommendations in the health domain. *Algorithms*, 13(3), 54. https://doi.org/10.3390/a13030054.

Tomeo, P., Di Noia, T., de Gemmis, M., Lops, P., Semeraro, G., & Di Sciascio, E. (2015, September 16-20). Exploiting regression trees as user models for intent-aware multi-attribute diversity. In T. Bogers, & M. Koolen (Eds.), *Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015)* (vol. 1448, pp. 1-8). CEUR Workshop Proceedings.

Vagliano, I., Monti, D., Scherp, A., & Morisio, M. (2017, December 4-6). Content recommendation through semantic annotation of user reviews and linked data - an extended technical report. In Association for Computing Machinery (Ed.), *Proceedings of the 9th International Conference on Knowledge Capture* (pp. 1-4). Association for Computing Machinery.

Vall, A., Dorfer, M., Eghbal-zadeh, H., Schedl, M., Burjorjee, K., & Widmer, G. (2019). Feature-combination hybrid recommender systems for automated music playlist continuation. *User Modeling and User-Adapted Interaction*, 29(2), 527-572. https://doi.org/10.1007/s11257-018-9215-8.

Vilakone, P., Park, D.-S., Xinchang, K., & Hao, F. (2018). An efficient movie recommendation algorithm based on improved *k*-clique. *Human-centric Computing and Information Sciences*, 8(1), 38. https://doi.org/10.1186/s13673-018-0161-6.

Wang, H., Chen, D., & Zhang, J. (2020). Group recommendation based on hybrid trust metric. *Automatika,* 61(4), 694-703. https://doi.org/10.1080/00051144.2020.1715590.

Wenige, L., & Ruhland, J. (2018). Retrieval by recommendation: Using LOD technologies to improve digital library search. *International Journal on Digital Libraries*, 19(2-3), 253-269. https://doi.org/10.1007/s00799-017-0224-8.

Xu, J., Yan, Z., Cao, G., & Zhao, J. (2018, August 17-19). Family shopping recommendation system using behavior sequence data and user profile. In ICIMCS (Ed.), *Proceedings of the ICIMCS '18: 10th International Conference on Internet Multimedia Computing and Service* (article no.: 41). Association for Computing Machinery.

Yadav, U., Duhan, N., & Bhatia, K. K. (2020). Dealing with pure new user cold-start problem in recommendation system based on linked open data and social network features. *Mobile Information Systems*, 2020, 8912065. https://doi.org/10.1155/2020/8912065.

Yang, R., Hu, W., & Qu, Y. (2013). Using semantic technology to improve recommender systems based on slope one. In J. Li, G. Qi, D. Zhao, W. Nejdl, & H.-T. Zheng (Eds.), *Semantic web and web science* (pp. 11-23). Springer.

Yochum, P., Chang, L., Gu, T., & Zhu, M. (2020). Linked open data in location-based recommendation system on tourism domain: A survey. *IEEE Access*, 8, 16409-16439. https://doi.org/10.1109/ACCESS.2020.2967120.

Zarzour, H., Maazouzi, F., Soltani, M., & Chemam, C. (2018, May 8-10). An improved collaborative filtering recommendation algorithm for big data. In A. Amine, M. Mouhoub, O. A. Mohamed, & B. Djebbar (Eds.), *Proceedings of the 6th IFIP TC 5 International Conference, CIIA 2018* (vol. 522, pp. 660-668). Springer.

Zaveri, A., Maurino, A., & Equille, L.-B. (2014). Web data quality: Current state and new challenges. *International Journal on Semantic Web and Information Systems*, 10(2), 1-6. http://doi.org/10.4018/ijswis.2014040101.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1), 1-38. https://doi.org/10.1145/3285029.

Zhao, W. X., He, G., Yang, K., Dou, H., Huang, J., Ouyang, S., & Wen, J.-R. (2019). KB4Rec: A data set for linking knowledge bases with recommender systems. *Data Intelligence*, 1(2), 121-136. https://doi.org/10.1162/dint_a_00008.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 9(2): 83-98, 2021

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2021.9.2.6

Influencing Factors of Research Collaboration Intention in Virtual Academic Communities in China

Chunlai Yan
Rattanakosin International College of Creative Entrepreneurship, Rajamangala University of Technology Rattanakosin, Phutthamonthon, Thailand
E-mail: wingvestige@163.com

Hongxia Li*
Rattanakosin International College of Creative Entrepreneurship, Rajamangala University of Technology Rattanakosin, Phutthamonthon, Thailand
School of Management Science and Engineering, Chongqing Technology and Business University, Chongqing, China
E-mail: lihongxia@ctbu.edu.cn

ABSTRACT

Research collaboration is an important strategy to improve research output, and virtual academic communities (VACs) have become an important platform to collaborate on. This paper reveals the influencing factors of researchers' collaboration intention in VACs from two attributes: individual, and inter-members. On the basis of the Social Cognitive Theory, Social Exchange Theory, social network theory, and Five-Factor Model, this paper constructed a model demonstrating the influencing factors of VACs researchers' collaboration intention. A self-administered questionnaire was employed on members of four VACs in China to collect data; subsequently, 558 usable responses were analyzed using structural equation modeling. The result showed that openness, conscientiousness, reciprocity, trust, and the social network characteristic had a significant influence on the collaboration intention of researchers in VACs, while self-efficacy, agreeableness, extroversion, neuroticism, and experience had no significant effects on the collaboration intention of researchers in VACs. This model plays a positive role in promoting the research collaboration intention of Chinese VACs researchers and in guiding the construction of VAC platforms.

Keywords: virtual academic communities, research collaboration, influencing factors, Five-Factor Model, China, social network characteristic

Received: May 11, 2021 **Revised:** May 25, 2021
Accepted: May 27, 2021 **Published:** June 30, 2021

***Corresponding Author:** Hongxia Li
https://orcid.org/0000-0002-6386-8152
E-mail: lihongxia@ctbu.edu.cn

All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

© Chunlai Yan, Hongxia Li, 2021

1. INTRODUCTION

1.1. Research Background

With the development of new knowledge communication carriers, the mode of academic exchange has also changed, and more scientific research scholars share their knowledge and collaborate in scientific research through online academic exchange platforms. The advancement of Web 2.0 technology makes it possible for researchers from different nations to work together with the support of online collaborative platforms, such as virtual academic communities (VACs) (Berdun & Armentano, 2019). VACs are online learning communities for researchers to exchange academic-related views and resources (Lin, 2007). With the coming era of big science and the globalization of science and technology, research collaboration has increasingly become the mainstream style way of scientific research. The collaborative behavior of research is becoming more and more critical in the process of scientific research and is becoming an essential way of academic innovation. VACs such as ResearchGate, Academic.edu, Mendeley, and muchong.com have gradually become sites for research collaboration.

1.2. Need for the Study

At present, VACs have been regarded by researchers as an essential place for the exchange of academic knowledge and data sharing. Nevertheless, more scholars have studied the exchange and sharing of information or knowledge in VACs. There is limited state-of-the-art research on the collaborative behavior of researchers in VACs, and the related research is still in its start. The collaboration between researchers in VACs has been carried out, but no research scholars have yet studied what factors affect collaborative intention. Although VACs have become a platform for researchers to communicate and collaborate, more researchers tend to collaborate offline. To fill the above research gaps and promote the collaboration behavior of researchers in VACs, this paper addresses the following two research requestions:

(1) What factors influence the collaboration intention of researchers in VACs?
(2) What strategies can promote the collaboration intention of researchers in VACs and involve more researchers in collaborative research projects online?

This study is helpful to enhance the enthusiasm and efficiency of researchers' participation in collaboration in VACs and to promote the healthy development of research collaboration activities in VACs. In addition, this research also constructs the theoretical basis of VACs and research collaboration and establishes a relatively perfect theoretical system of collaborative behavior in VACs.

2. LITERATURE REVIEW

2.1. Virtual Academic Communities

Rheingold (1993) concludes that a "virtual community" relates to a group of people who interact with each other over the social network to share data and knowledge, to develop a group of personal connecting society in the cyberspace world. Harrison and Stephen (1995) proposed for the first time that academic communities can be understood as discourse communities that share data and communication protocols. Shade (1995) believes that the Internet has become an essential tool to maintain and cultivate academic culture, and the concept of the "invisible college" has mainly been transformed into a virtual community. Lin (2007) define VACs as learning communities, which they believe are virtual learning communities for researchers to exchange ideas, experiences, documents, comments, and reviews related to the academic culture. Therefore, VACs not only have the principle of the virtual community, but also make academic knowledge exchanges, which require them to meet three key conditions, firstly the computer as the medium, secondly academic knowledge as the exchange content, and finally, the researchers as the participants.

2.2. Research Collaboration

At present, there are different definitions of research collaboration in academia. Katz and Martin (1997) define research collaboration as the common desire of researchers to work together to create new knowledge. Ziman (1994) proposes that research collaboration is the product of scientific development to a certain "stable period," and the collaborative effect after entering the "stable period" plays an increasingly important role in improving the output of scientific knowledge. Heffner (1981) believes that research collaboration is a form of researcher interaction. Researchers in the process of scientific activities through ideological and intellectual communication form the ideal research collaboration model. Bergstrom (2008) found that some scholars opposed the idea of teacher-student collaboration as a kind of research collaboration, and believed that real research collaboration should occur between researchers of the same rank or similar status. This

 https://doi.org/10.1633/JISTaP.2021.9.2.6

research work defines research collaboration based on the above analysis: Research collaboration refers to a scientific activity in which two or more researchers or organizations achieve the goal of enhancing research output through collaboration to complete the same research task or proposed project.

2.3. State-of-the-Art of Research Collaboration in VACs

The activities of researchers in VACs are mainly scientific research, which will gradually produce research collaborative behavior within the community. Research about research collaboration in VACs is still lacking. Most of the previous research efforts focus on knowledge sharing and communication in VACs. By reading the literature related to virtual academic community topics, the proposed research study found that virtual academic community research collaboration is gradually developed by knowledge exchange and sharing among researchers. Collaboration behavior occurs mainly because of the research demands of researchers in VACs (Qu, 2010). In the process of reaching cooperation, communication within the community is an essential linkage. The state-of-the-art on knowledge exchange of VACs researchers mainly focuses on the following three aspects: (1) the influencing factors of knowledge exchange in the virtual academic community; (2) the evaluation of knowledge exchange in the virtual academic community; and (3) the strategy improvement of knowledge exchange in the virtual academic community. There is limited literature related to the collaborative topic of virtual academic community research. Chia and Pritchard (2014) use the case study method to analyze transnational cooperation in the virtual academic community; it is concluded that the community is of considerable significance to promote transnational collaboration. Maxwell et al.'s (2017) research found that medium and long-term cooperation in VACs is better than short-term collaboration. With an in-depth understanding of each other's relationship, they can form a great academic understanding.

By reviewing the literature, we find that there are few literature studies on research collaboration in VACs, however, and those that exist have limited scope. There is no empirical study on influencing factors of research collaboration in VACs, which bridges the research gaps for our study.

3. THEORETICAL FRAMEWORK AND RESEARCH HYPOTHESES

This study explores the factors that affect the intention of research collaboration in VACs from the dimension of individual factors and inter-member factors and takes Social Cognitive Theory (SCT), Social Exchange Theory (SET), Big-five personality model, and Social Network Theory (SNT) as the theoretical basis. Based on the above theories, this study puts forward the following hypotheses from six dimensions: self-efficacy, personality traits, reciprocity, experience, trust, and the social network characteristic.

3.1. Self-Efficacy and Research Collaboration Intention

Self-efficacy refers to individuals' anticipation of the impact of their behavior, which is the individual's perception of their own ability and self-confidence (Lu et al., 2006). The higher the self-efficacy is, the more feasible it is for individuals to choose and decide to carry out an action. Many researchers pointed out that knowledge sharing self-efficacy can affect users' sharing behavior in a virtual community (Almahamid et al., 2010; Cabrera & Cabrera, 2002; Hazzan & Seger, 2010).

This study anticipates that researchers who believe they can contribute and acquire knowledge in VACs also prefer to collaborate with others to finish research projects. Therefore, the following hypothesis is proposed.

H1: Self-efficacy has a positive influence on research collaboration intention in VACs.

3.2. Personality Trait and Research Collaboration Intention

The personality trait is a stability index used to describe comprehensive individual traits, and it is also the fundamental factor (Costa & McCrae, 1992a) that affects individual psychological attitudes and behavior activity. Nonnecke and Preece (2001) proposed that shyness and privacy are the main factors that affect researchers' reluctance to share in virtual communities. Among the research efforts on researcher personality trait correlation, the Five-factor model of personality structure is one of the most widely used personality trait theories (Costa & McCrae, 1992b). The FFM of personality traits divides personality traits into five dimensions, such as agreeableness, openness, extroversion, conscientiousness, and neuroticism (Gosling et al., 2003).

People with high agreeableness are friendly and helpful, and they are willing to give up their interests (Wang et al., 2012) for others. Moore and McElroy (2012) believe that researchers with high agreeableness regret sharing incorrect information. Ryan and Xenos (2011) find that researchers with high agreeableness are more willing to share information in virtual communities. Therefore,

researchers with higher agreeableness tend to prefer to collaborate with virtual community members rather than compete, for altruistic reasons. Based on the above analysis, the following hypothesis is made in this proposed research work.

H2(a): Agreeableness positively influences research collaboration intention in VACs.

People with high openness are imaginative and creative and are more sensitive to wonderful things. Cabrera et al. (2006) believe that openness is a significant antecedent variable of knowledge sharing, and 'High-Open' researchers are more willing to share knowledge and meet the needs of the team for knowledge acquisition. Therefore, high openness people have more intention of knowledge sharing in the virtual academic community. Based on the above analysis, the following hypothesis is made in this research work.

H2(b): Openness positively influences research collaboration intention in VACs.

Extroversion reflects the strength and dynamic characteristics of the individual nervous system. Highly extroverted people are passionate and adventurous; in a group, they are talkative and enjoy attracting attention (Barrick et al., 1998). These characteristics indicate that highly extroverted people prefer to communicate and collaborate with others in virtual communities. Accordingly, the hypothesis is made as below.

H2(c): Extroversion positively influences research collaboration intention in VACs.

Conscientiousness reflects the ability to self-discipline and the sense of responsibility to achieve. Highly conscientious people have a strong sense of responsibility, for which they may be more willing to help others. Yoo and Gretzel (2011) believe that people with high conscientiousness are more willing to share knowledge. Therefore, conscientiousness has a significant positive effect on knowledge sharing. Based on the above analysis, the following hypothesis is made in this proposed research work.

H2(d): Conscientiousness positively influences research collaboration intention in VACs.

Neuroticism refers to the situation in which the emotional state of a person fluctuates with the change of external conditions. People with high neuroticism are more likely to feel negative emotions such as anger, anxiety, and depression. Their response to external stimuli is stronger than the average person's, and their ability to regulate emotions is relatively weak. They are often unwilling to collaborate with others. Wang et al. (2012) find that highly neurotic people have fewer status updates in virtual communities. The following hypothesis is made.

H2(e): Neuroticism negatively influences research collaboration intention in VACs.

3.3. Experience and Research Collaboration Intention

The father of experience marketing, B. H. Schmitt (1999), believes that experience is the perception generated by brain stimulation during activity. Nambisan and Watt (2011) put forward the idea of "online community experience," which is defined as the all-around feeling that people get in the process of network community interaction. This study defines experience as the subjective feelings gained by researchers in the process of using VACs. Scholars have identified cognitive and emotional aspects that affect people's experiences in the field of psychological research (Hirschman & Holbrook, 1982; Overby & Lee, 2006; Voss et al., 2003). Hars and Ou (2002) believes that creating reciprocal social norms within a virtual community allows people to experience intimacy. Moreover, the activities of members of VACs mainly include obtaining valuable academic information and looking for academic partners. Therefore, the experience of VACs includes not only utilitarian experience and emotional experience but also a social experience. When the researcher acquires the knowledge needed from others in the academic community, and the process of communicating and contacting the researcher produces a pleasant experience, it will promote the intention of the researcher to collaborate in research. Based on the above analysis, the following hypothesis is made in this research work.

H3: Experience positively influences research collaboration intention in VACs.

3.4. Reciprocal Benefit and Research Collaboration Intention

The theory of social exchange shows that the mechanism to ensure the success of exchange is reciprocity, and any interaction in society can be explained by reciprocity (Coyle-Shapiro & Conway, 2005). Many researchers believe that reciprocity promotes collaborative research in the virtual academic community, because it improves relationships among researchers and increases the expectations of helping others to ensure ongoing collaboration (Bock et al., 2005; Davenport & Prusak, 1998; Kankanhalli et al., 2005; Lin, 2007; Lin et al., 2009). Research collabo-

 https://doi.org/10.1633/JISTaP.2021.9.2.6

ration involves the contribution and absorption of knowledge. On the surface, research collaboration is a kind of cooperation based on knowledge exchange, but it is a kind of reciprocity based on material or spiritual return. Research collaboration depends on reciprocal mechanisms between collaborators. If one party's actions benefit the other's, it can move the whole group towards reciprocity. On the contrary, if one party's efforts cannot be rewarded, or if both parties agree that the gains are not worth the losses, the collaborative relationship will not be sustainable. Therefore, the following hypothesis is proposed.

H4: Reciprocal benefit positively influences research collaboration intention in VACs.

3.5. Trust and Research Collaboration Intention

According to the theory of social exchange, trust is the basis of human communication and collaboration (Jarvenpaa et al., 2004; Zaheer et al., 1998). Trust is essentially a kind of interpersonal relationship (Powell et al., 2004). Emotional trust can effectively adjust the relationship between enterprise employees, which is considered one of the tacit assets of the enterprise (Reina & Reina, 2006; Vangen & Huxham, 2003). Researchers in VACs do not know others and need to increase their trust in each other to promote collaboration and communication. If there is no trust, deception may occur. If trust arises among researchers in VACs, they will enhance their knowledge sharing intention and collaborate with other researchers in the community. The stronger this emotional trust, the more urgent the desire for knowledge sharing (Black et al., 2003). Therefore, the following hypothesis is proposed.

H5: Trust positively influences research collaboration intention in VACs.

3.6. Social Networks Characteristics and Research Collaboration Intention

The research collaboration behavior in VACs depends on a multi-dimensional relationship of social networks. Because human beings are creatures with social attributes who pay attention to the construction of interpersonal relationships, individuals tend to show their tacit knowledge in their respective small-scale networks to realize the exchange and sharing of knowledge. These kinds of offline activities also hold the attitude and tendency of sharing after moving to the virtual academic community. Kotlarsky and Oshri (2005) finds that relationship orientation can promote knowledge transfer between enterprises. Okoli and Oh (2007) believe that the higher density of the network nodes, the higher the performance of their knowledge contribution; Lu and Yang (2011) shows that online interactive connections in virtual communities have a significant impact on knowledge sharing behavior among community members. Coleman (1988) believes that social connections between people can promote group coherence and facilitate group knowledge sharing. Many studies show that node density and connection strength in social networks promote the transfer and sharing of knowledge (Reagans & McEvily, 2003; Uzzi & Lancaster, 2003). In summary, this research work proposes the hypothesis:

H6: Social network characteristics are positively related to research collaboration intention.

3.7. Research Model

From the above discussion, the final research model of the study is shown in Fig. 1. The variables in the research model include independent variables (self-efficacy, agreeableness, openness, extroversion, conscientiousness, neuroticism, experience, reciprocity, trust, social networks characteristic) and dependent variables (research collaboration intention). In terms of individual attributes, researchers' self-efficacy, agreeableness, openness, extroversion, conscientiousness, and experience positively affect research collaboration intention, but neuroticism negatively affects research collaboration intention. From the aspect of the inter-member attribute, the reciprocity, trust, and social network characteristics between researchers positively affect research collaboration intention.

4. METHODOLOGY

The methodology of this study is qualitative, and we adopt a survey-based approach to collect data. The quantitative method in this study refers to statistical and mathematical methods to analyze data to confirm the findings of the qualitative analysis.

4.1. Measurement of Variables

All the variables in the research model are latent, and their measures are mainly carried out by designing the corresponding scale. The items of the scale mainly refer to the mature scale in authoritative journals and are closely combined with the design of research requirements. The scale design of self-efficacy is by Chen and Hung (2010); The scale design of personality trait is from Hsu et al. (2007a); The scale design of experience is from Leung (2003) and Flanagin and Metzger (2001); The scale design of reciprocity is from Kankanhalli et al. (2005); the scale design of trust is from Fang and Chiu (2010) and Hsu et

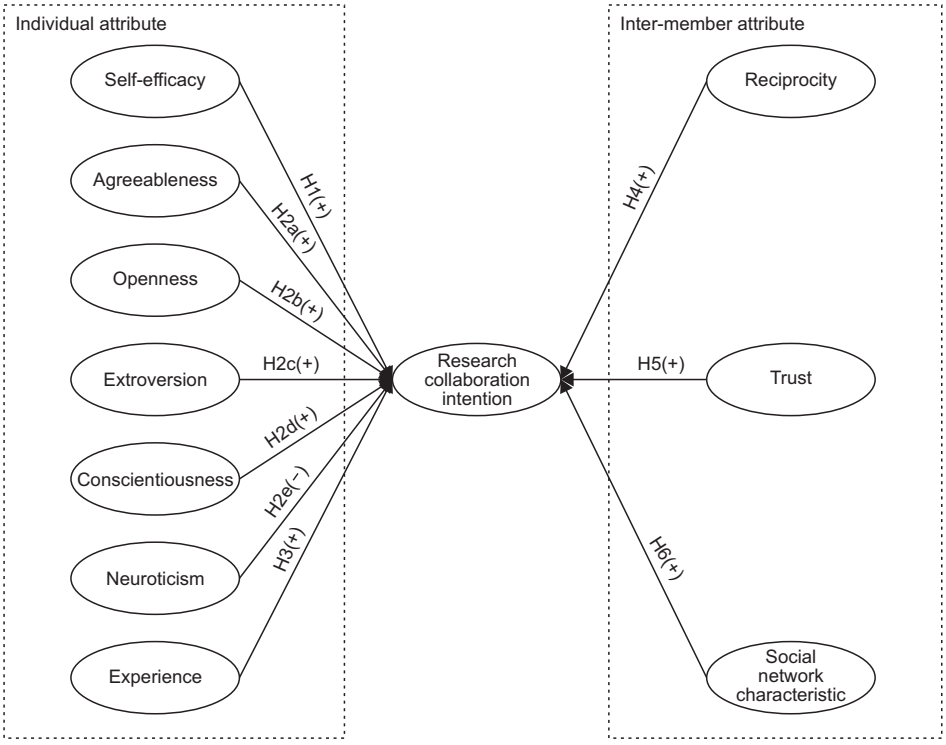


Fig. 1. Research model.

al. (2007b); the scale design of social networks is from Chen et al. (2009). All measurement items are presented in Appendix.

4.2. Population and Sample

The target populations of this study are online registered researchers who usually use VACs. Sample selection is limited to VACs with a large number of researchers in China, such as muchong.com, jg.com.cn, sciencenet.cn, and DXY.CN. The sampling technology combines two approaches, such as simple random sampling and snowball sampling. The questionnaires were made on the Wenjuanxing Website, and sent to the relevant VACs. We collected 558 valid questionnaires in this survey. Table 1 shows the descriptive statistics of respondents.

As shown in Table 1, 50.7% of the respondent were male and 49.3% were female, which indicated that the sex distribution was uniform. The age range of 18-25 years and 26-35 years of age accounted for the most, and 61.3% of them had a master's degree or above, which was consistent with the fact that most of the respondents were researchers and students. The largest number of respondents used VACs once per week, accounting for 50 percent, indicating that half of the respondents were active in VACs; about 65 percent of the respondents had more than half a year of experience using VACs.

5. DATA ANALYSIS AND HYPOTHESIS TEST

5.1. Reliability and Validity Test

The Cronbach's alpha coefficient was used to test the reliability of the scale. Cronbach's alpha value is between 0 and 1. If α value is lower than 0.6, this indicates that the reliability of the scale is not good; when it reaches 0.7-0.8, it means that the scale has considerable reliability. When it reaches 0.8-0.9, it means that the reliability of the scale is very good. In the aspect of validity, the scale of this paper comes from the mature scale, and a small part has been modified to combine the research practice. Exploratory factor analysis (EFA) and confirmatory factor analysis were used to test the structural validity of the scales. In EFA, we delete items with factor loading less than 0.5 and cross-factor loading greater than 0.4. The reliability and validity analysis of the scale is shown in Table 2.

As shown in Table 2, the values of Cronbach's of each variable are more than 0.7, which indicates that the scale of this study has good reliability. The values of Corrected Item-Total Correlation are between 0.497 and 0.631, which means the correlation between items is strong. The values of factors loading of each item are all more than 0.5, indicating that factors have a good explanation for the corresponding latent variables. Therefore, the scale of this study has good structural validity.

 https://doi.org/10.1633/JISTaP.2021.9.2.6

Table 1. Descriptive statistics

Measure	Items	Frequency	Percent
Sex	Male	283	50.7
	Female	275	49.3
Age	Under 18 years	0	0.0
	Between 18 and 25 years	240	43.0
	Between 26 and 35 years	218	39.1
	Between 36 and 45 years	67	12.0
	Over 45 years	33	5.9
Education background	College and below	20	3.6
	Bachelor degree	196	35.1
	Master degree	235	42.1
	Doctor degree and above	107	19.2
Disciplinary background	Science and engineering	163	29.2
	Literature and philosophy	87	15.6
	Economics and management	266	47.7
	Medical	35	6.3
	Law	7	1.3
Position	Students	228	40.9
	Researchers and teachers	160	28.7
	Civil servants	139	24.9
	Private owners	31	5.6
Frequency of use	Every day	199	35.7
	Weekly	279	50.0
	Monthly	80	14.3
VACs member experience (in years)	Within one month	39	7.0
	Between 1 and 6 months	157	28.1
	Between 6 months and 1 year	136	24.4
	Between 1 and 3 years	109	19.5
	More than 3 years	117	21.0

VACs, virtual academic communities.

Before factor analysis, the Kaise-Meryer-Olkin test and Bartlett's test of sphericity were used to verify the partial correlation and simple correlation coefficients among the variables. When the correlation is high, the data is suitable for factor analysis. The commonly used Kaiser-Meyer-Olkin (KMO) metrics are as follows: 0.9 or more is strongly suitable for factor analysis; 0.8-0.9 is relatively suitable; 0.7-0.8 is suitable; 0.6-0.7 is acceptable; 0.5-0.6 means relatively not suitable; less than 0.5 is strongly not suitable. The KMO and Bartlett's test results are shown in Table 3.

As shown in Table 3, the KMO value of the scale is 0.946, and Bartlett's Test of Sphericity is 16141.902, $p<0.01$, which means the structure of each item of the scale is reasonable and suitable for factor analysis.

5.2. Model and Hypothesis Testing

After the reliability and validity test, the software AMOS 24.0 (IBM Corp., Armonk, NY, USA) was used to study the fitting degree of the model. The structural

Table 2. Reliability and validity analysis of scale

Variables	Items	Corrected item-total correlation	Cronbach's alpha if item deleted	Cronbach's alpha of variables	Factor loading
Self-efficiency	SE1	0.597	0.935	0.832	0.815
	SE2	0.589	0.935		0.858
	SE3	0.624	0.935		0.717
Agreeableness	AG1	-0.193	0.944	0.913	0.869
	AG2	-0.191	0.944		0.866
	AG3	-0.167	0.944		0.912
Openness	OP1	0.497	0.936	0.751	0.764
	OP2	0.485	0.936		0.777
	OP3	0.498	0.936		0.611
Extroversion	EX1	0.535	0.935	0.808	0.687
	EX2	0.629	0.935		0.826
	EX3	0.576	0.935		0.781
Conscientiousness	CO1	0.584	0.935	0.757	0.750
	CO2	0.509	0.936		0.696
	CO3	0.518	0.936		0.681
Neuroticism	NEU1	0.528	0.935	0.907	0.850
	NEU2	0.557	0.935		0.890
	NEU3	0.545	0.935		0.885
Experience	UE1	0.567	0.935	0.886	0.787
	UE2	0.505	0.936		0.742
	UE3	0.553	0.935		0.674
Reciprocal	RE1	0.621	0.935	0.868	0.737
	RE2	0.551	0.935		0.765
	RE3	0.626	0.935		0.798
	RE4	0.631	0.935		0.793
Trust	TR3	0.582	0.935	0.881	0.711
	TR4	0.614	0.935		0.772
	TR5	0.571	0.935		0.765
	TR6	0.618	0.935		0.768
Social networks characteristic	CS1	0.596	0.935	0.899	0.843
	CS2	0.640	0.935		0.846
	ND1	0.557	0.935		0.844
	ND2	0.579	0.935		0.819
Research collaboration intention	RCI1	0.596	0.935	0.861	0.777
	RCI2	0.582	0.935		0.820
	RCI3	0.599	0.935		0.828

 https://doi.org/10.1633/JISTaP.2021.9.2.6

Table 3. KMO and Bartlett's test

KMO measure of sampling adequacy		0.946
Bartlett's test of sphericity	Approx. Chi-Square	16141.902
	df	903
	Sig.	0.000

KMO, Kaiser-Meyer-Olkin; Approx., approximate chi-square test; df, degree of freedom; Sig., significance.

Table 4. Model fit index

Fit index	Standard	Value
χ^2/df	<3	2.120
GFI	>0.9	0.908
NFI	>0.9	0.918
IFI	>0.9	0.955
TLI (NNFI)	>0.9	0.945
CFI	>0.9	0.954
RMSEA	<0.05	0.044

df, degree of freedom; GFI, Goodness-of-fit index; NFI, Normed fit index; IFI, Incremental fit index; TLI, Tucker-Lewis index; NNFI, Non-normed fit index; CFI, Comparative fit index; RMSEA, Root Mean Square Error of Approximation.

equation model (SEM) was used to study the relationship between variables. The goodness of fit index is an indicator used in the structural equations to test the fit degree of the model and the data. Commonly used indicators are the χ^2/df, Root Mean Square Error of Approximation (RMSEA), Goodness-of-fit index (GFI), Non-normed fit index (NNFI), Comparative fit index (CFI), Normed fit index (NFI), and Incremental fit index (IFI). It is generally believed that if χ^2/df is below 3, RMSEA is below 0.08, and GFI, NNFI, CFI, NFI, and IFI are above 0.9, the fitted model is a "good" model. The model fit index of this study is shown in Table 4.

From Table 4, we find that absolute fit index χ^2/df is 2.120 and smaller than 3; RMSEA is 0.044 and smaller than 0.05; relative fit index NFI is 0.918 and greater than 0.9; IFI is 0.955 and greater than 0.9; NNFI is 0.945, and greater than 0.9; CFI is 0.954, and greater than 0.9. That means all model parameters meet reasonable standards and the proposed research model fit data well. Based on the results of Table 4, we tested the previously proposed hypothesis by using the path coefficient. The path diagram is shown in Fig. 2. The path coefficient of the SEM is shown in Table 5.

Fig. 2 is a model path diagram drawn after importing the questionnaire data into the AMOS, which can visually express the linear relationship between the variables. Within the elliptical box are the latent variables, containing all the independent variables of this paper. Within the rectangular box are the observation variables, which consist of the scale items of each variable. Within small circles are measurement errors (such as e1 to e37). A one-way arrow connecting two variables assumes a causal relationship between them, and the arrow points from the independent variable to the factor variable. The curved two-arrow represents a correlation between the two variables. The value on the connector is the path coefficient to measure the degree of influence or correlation between variables. The path coefficient values between the independent variables and dependent variable in Fig. 2 are included in Table 5.

In Table 5, path coefficient is the non-standardized regression coefficient of the research model. S.E. is used to estimate the standard error of parameters, and C.R. is used to test the critical ratio of statistics. If the C.R. is greater than 1.96 and $p<0.05$, it indicates that the path coefficient has reached a significant level, and the hypothesis is supported. If the contrary applies, the hypothesis is not supported.

From Fig. 2 and Table 5, we found that for individual attributes the path coefficient of self-efficiency, agreeableness, extroversion, neuroticism, and experience impacting on research collaboration intention are -0.142, 0.002, -0.282, 0.023, and -0.097, and their p-value is greater than 0.05 ($p>0.05$), therefore H1, H2a, H2c, H2e, and H3 are not supported; openness and trust have a significant positive function on research collaboration intention (C.R.>1.96, $p<0.05$), therefore H2b and H2d are both proven. For inter-members attributes, we know that reciprocal, trust, and social network characteristics have a significant positive function on research collaboration intention (C.R.>1.96, $p<0.05$), so H4, H5, and H6 are all supported.

For the supported hypothesis, reciprocal (H4) is an important factor of research collaboration intention in VACs. Reciprocal has the most significant influence on research collaboration intention, and this is consistent with the results of Lin (2007). Hence, members were more likely to have a higher tendency to collaborate with others in VACs, if they obtain reciprocal benefits from other researchers by contributing knowledge. Social network characteristics (H6) have a significant positive correlation to research collaboration intention, and this is consistent

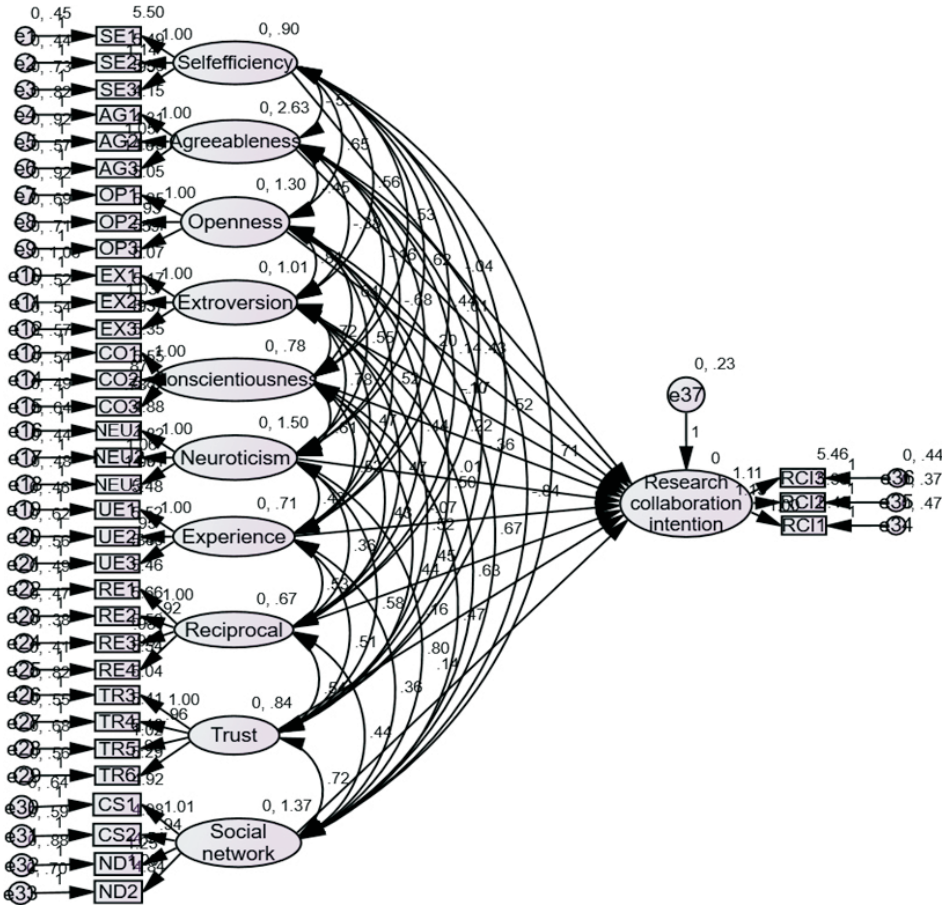


Fig. 2. Path diagram of the model.

Table 5. Path coefficient of structural equation model

Hypothesis	Path hypothesis	Path coefficient	S.E.	*z* (C.R.)	*p*-value	Test results
H1	RCI ← Self-efficiency	-0.142	0.112	-1.267	0.205	Not Supported
H2a	RCI ← Agreeableness	0.002	0.024	0.072	0.942	Not Supported
H2b	RCI ← Openness	0.146	0.063	2.334	0.020	Supported
H2c	RCI ← Extroversion	-0.282	0.156	-1.805	0.071	Not Supported
H2d	RCI ← Conscientiousness	0.393	0.179	2.195	0.028	Supported
H2e	RCI ← Neuroticism	0.023	0.043	0.540	0.589	Not Supported
H3	RCI ← Experience	-0.097	0.112	-0.862	0.389	Not Supported
H4	RCI ← Reciprocal	0.453	0.098	4.656	<0.001	Supported
H5	RCI ← Trust	0.164	0.060	2.789	0.005	Supported
H6	RCI ← Social network characteristics	0.144	0.046	3.139	0.002	Supported

S.E., standard error; C.R., critical ratio.

with the results of Lu and Yang (2011) and Okoli and Oh (2007). Therefore, social network characteristic is an important factor in research collaboration in VACs. Trust (H5) is a significant positive correlation with research collaboration intention, and this is consistent with the results of Vangen and Huxham (2003), Reina and Reina (2006), and Black et al. (2002). Trust building and maintaining trust mechanisms are the topic for most research-

ers, which suggests that trust is a core factor in the three terms of collaboration, coordination, and cooperation. Openness (H2b) and conscientiousness (H2d) of the FFM have a significant influence on the research collaboration intention in VACs, which are consistent with the results of Cabrera et al. (2006) and Yoo and Gretzel (2011).

For the hypotheses that are not supported, self-efficacy (H1) has no significant effect on research collaboration intention. Analysis of the reasons may be related to the respondents from different VACs. Because VACs have many uncertain factors, such as trust, capital, completion time, and so on, agreeableness (H2a) and extroversion (H2c) have no significant effect on research collaboration intention, which is consistent with the result of Hughes et al. (2012). The individuals with higher agreeableness and extroversion have better interpersonal relationships offline, therefore they do not need to find partners in the VACs. On neuroticism (H2e), this study found that it was not significantly related to research collaboration intention in VACs, which is consistent with the results of Ross et al. (2009). Highly neurotic people often feel anxious and lonely in their research work (Ebeling-Witte et al., 2007), and will use VACs more frequently to seek psychological comfort. But because of their inner sensitivity and suspicion, it is difficult for them to believe in others, and then they do not want to cooperate with others in VACs. The researcher's experience (H3) in VACs has no significant impact on research collaboration intention. Most studies have found that good experience can increase the intention of researchers to use VACs continuously (Qu & Zhang, 2015; Qu et al., 2016), but it has no significant correlation with the research collaboration intention.

6. DISCUSSION AND CONCLUSION

This study provides a new theoretical model for the research of VACs. On the basis of the theories of SCT, SET, SNT, and FFM, this paper constructed a model demonstrating the influencing factors of VACs researchers' research collaboration intention. The result shows that openness, conscientiousness, reciprocity, trust, and social network characteristic have a significant influence on the collaboration intention of researchers in VACs. In order to strengthen the research collaboration behavior and accelerate the development of VACs, we should improve the regulatory system, punish the researchers who have broken their promises, and create a safe cooperation environment. We also should create a good collaboration culture, atmosphere, and incentive mechanism to encourage interaction between members, so that they can establish an effective network connection and mutual understanding. From the attribute of the individual, increasing openness, enhancing conscientiousness, and self-reputation will gain more partners in VACs.

According to the above research results, this paper proposes several suggestions to promote the research collaboration intention and community development. First, managers of VACs should establish management norms and trust mechanisms, because the community management norms and trust mechanisms are the necessary guarantees for the trust between researchers. The perfect trust management norms and mechanism will promote the researchers' intention to collaborate. Second, community managers should create a cultural atmosphere and enhance researchers' loyalty, belonging, and identity to the community, thus prompting them to find partners in the VACs to complete research tasks. Finally, the reward and punishment system in VACs should be improved. Managers should join research institutions and academic conferences to release research tasks and strengthen collaborative researchers' motivation to collaborate again by improving the reputation and distributing community gold coins.

We select the sample from four famous VACs in China. Future research will select more types of VACs to further verify the theoretical model. The sample mainly comes from college students and teachers, while the other occupations' groups are relatively few. We will expand the target population scope of the questionnaire survey in future study. The research model of this paper mainly considers the individual attributes and inter-member attribute of researchers in VACs, and future study will add the attributes of community to perfect the theoretical model.

ACKNOWLEDGMENTS

We would like to thank all respondents for participating in our questionnaire. This study would have been impossible without support by the National Natural Science Foundation of China and Chongqing Education Commission of China (project: 70901080, 19SKGH091).

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Almahamid, S., McAdams, A. C., & Kalaldeh, T. (2010). The relationship among organizational knowledge sharing practices, employees' learning commitments, employees' adaptability, and employees' job satisfaction: An empirical investigation. *Interdisciplinary Journal of Information, Knowledge, and Management*, 5, 327-356. https://doi.org/10.28945/1225.

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, 83(3), 377-391. https://doi.org/10.1037/0021-9010.83.3.377.

Berdun, F. D., & Armentano, M. G. (2019). Modeling users' collaborative behavior with a serious game. *IEEE Transactions on Games*, 11(2), 121-128. https://doi.org/10.1109/TG.2018.2794419.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor™ metrics. *Journal of Neuroscience*, 28(45), 11433-11434. https://doi.org/10.1523/jneurosci.0003-08.2008.

Black, L. J., Cresswell, A. M., Luna, L. F., Pardo, T. A., Martinez, I. J., Thompson, F., Andersen, D. F., Canestraro, D. S., Richardson, G. P., & Cook, M. (2003, January 6-9). A dynamic theory of collaboration: A structural approach to facilitating intergovernmental use of information technology. In R. H. Sprague, Jr. (Eds.), *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (pp. 12). IEEE.

Bock, G.-W., Zmud, R. W., Kim, Y.-G., & Lee, J.-N. (2005). Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. *MIS Quarterly*, 29(1), 87-111. https://doi.org/10.2307/25148669.

Cabrera, A., & Cabrera, E. F. (2002). Knowledge-sharing dilemmas. *Organization Studies*, 23(5), 687-710. https://doi.org/10.1177/0170840602235001.

Cabrera, Á., Collins, W. C., & Salgado, J. F. (2006). Determinants of individual engagement in knowledge sharing. *International Journal of Human Resource Management*, 17(2), 245-264. https://doi.org/10.1080/09585190500404614.

Chen, C.-J., & Hung, S.-W. (2010). To give or to receive? Factors influencing members' knowledge sharing and community promotion in professional virtual communities. *Information & Management*, 47(4), 226-236. https://doi.org/10.1016/j.im.2010.03.001.

Chen, I. Y. L., Chen, N.-S., & Kinshuk (2009). Examining the factors influencing participants' knowledge sharing behavior in virtual learning communities. *Educational Technology & Society*, 12(1), 134-148. https://eric.ed.gov/?id=EJ833422.

Chia, H. P., & Pritchard, A. (2014). Using a virtual learning community (VLC) to facilitate a cross-national science research collaboration between secondary school students. *Computers & Education*, 79, 1-15. https://doi.org/10.1016/j.compedu.2014.07.005.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94, S95-S120. https://doi.org/10.1086/228943.

Costa, P. T. Jr., & McCrae, R. R. (1992a). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4(1), 5-13. https://doi.org/10.1037/1040-3590.4.1.5.

Costa, P. T. Jr., McCrae, R. R. (1992b). Four ways five factors are basic. *Personality and Individual Differences*, 13(6), 653-665. https://doi.org/10.1016/0191-8869(92)90236-i.

Coyle-Shapiro, J. A-M., & Conway, N. (2005). Exchange relationships: Examining psychological contracts and perceived organizational support. *Journal of Applied Psychology*, 90(4), 774-781. https://doi.org/10.1037/0021-9010.90.4.774.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business School Press.

Ebeling-Witte, S., Frank, M. L., & Lester, D. (2007). Shyness, internet use, and personality. *CyberPsychology & Behavior*, 10(5), 713-716. https://doi.org/10.1089/cpb.2007.9964.

Fang, Y.-H., & Chiu, C.-M. (2010). In justice we trust: Exploring knowledge-sharing continuance intentions in virtual communities of practice. *Computers in Human Behavior*, 26(2), 235-246. https://doi.org/10.1016/j.chb.2009.09.005.

Flanagin, A. J., & Metzger, M. J. (2001). Internet use in the contemporary media environment. *Human Communication Research*, 27(1), 153-181. https://doi.org/10.1111/j.1468-2958.2001.tb00779.x.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504-528. https://doi.org/10.1016/s0092-6566(03)00046-1.

Harris, T. M., & Stephen, T. D. (1995). The electronic journal as the heart of an online scholarly community. *Library Trends*, 43(4), 592-608.

Hars, A., & Ou, S. (2002). Working for free? Motivations of participating in open-source projects. *International Journal of Electronic Commerce*, 6(3), 25-39. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.2007&rep=rep1&type=pdf.

Hazzan, O., & Seger, T. (2010). Recruiting software practitioners: The importance of self-efficacy. *CrossTalk: Journal of Defense Software Engineering*, 23(3), 8-11. https://citeseerx.

ist.psu.edu/viewdoc/download?doi=10.1.1.432.8448&rep=rep1&type=pdf.

Heffner, A. G. (1981). Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, 3(1), 5-12. https://doi.org/10.1007/bf02021860.

Hirschman, E. C., & Holbrook, M. B. (1982). Hedonic consumption: Emerging concepts, methods and propositions. *Journal of Marketing*, 46(3), 92-101. https://doi.org/10.1177/002224298204600314.

Hsu, B.-F., Wu, W.-L., & Yeh, R.-S. (2007a, August 5-9). Personality composition, affective tie and knowledge sharing: A team level analysis. In D. F. Kocaoglu, T. R. Anderson, T. U. Daim, A. Jetter, & C. M. Weber (Eds.), *Proceedings of the PICMET '07 - 2007 Portland International Conference on Management of Engineering & Technology* (pp. 2583-2592). Portland State University.

Hsu, M.-H., Ju, T. L., Yen, C.-H., & Chang, C.-M. (2007b). Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. *International Journal of Human-Computer Studies*, 65(2), 153-169. https://doi.org/10.1016/j.ijhcs.2006.09.003.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2), 561-569. https://doi.org/10.1016/j.chb.2011.11.001.

Jarvenpaa, S. L., Shaw, T. R., Staples, D. S. (2004). Toward contextualized theories of trust: The role of trust in global virtual teams. *Information Systems Research*, 15(3), 215-310. https://doi.org/10.1287/isre.1040.0028.

Kankanhalli, A., Tan, B. C. Y., & Wei, K. K. (2005). Contributing knowledge to electronic knowledge repositories: An empirical investigation. *MIS Quarterly*, 29(1), 113-143. https://doi.org/10.2307/25148670.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1-18. https://doi.org/10.1016/s0048-7333(96)00917-1.

Kotlarsky, J., & Oshri, I. (2005). Social ties, knowledge sharing and successful collaboration in globally distributed system development projects. *European Journal of Information Systems*, 14(1), 37-48. https://doi.org/10.1057/palgrave.ejis.3000520.

Leung, L. (2003). Impacts of Net-generation attributes, seductive properties of the Internet, and gratifications-obtained on Internet use. *Telematics & Informatics*, 20(2), 107-129. https://doi.org/10.1016/s0736-5853(02)00019-9.

Lin, H.-F. (2007). Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions. *Journal of Information Science*, 33(2), 135-149. https://doi.org/10.1177/0165551506068174.

Lin, H.-F., Lee, H.-S., & Wang, D.W. (2009). Evaluation of factors influencing knowledge sharing based on a fuzzy AHP approach. *Journal of Information Science*, 35(1), 25-44. https://doi.org/10.1177/0165551508091310.

Lu, L., Leung, K., & Koch, P. T. (2006). Managerial knowledge sharing: The role of individual, interpersonal, and organizational factors. *Management and Organization Review*, 2(1), 15-41. https://doi.org/10.1111/j.1740-8784.2006.00029.x.

Lu, Y., & Yang, D. (2011). Information exchange in virtual communities under extreme disaster conditions. *Decision Support Systems*, 50(2), 529-538. https://doi.org/10.1016/j.dss.2010.11.011.

Maxwell, A. E., Crespi, C. M., Arce, A. A., & Bastani, R. (2017). Exploring the effects of longstanding academic-community partnerships on study outcomes: A case study. *Preventive Medicine Reports*, 8, 101-107. https://doi.org/10.1016/j.pmedr.2017.09.003.

Moore, K., & McElroy, J. C. (2012). The influence of personality on Facebook usage, wall postings, and regret. *Computers in Human Behavior*, 28(1), 267-274. https://doi.org/10.1016/j.chb.2011.09.009.

Nambisan, P., & Watt, J. H. (2011). Managing customer experiences in online product communities. *Journal of Business Research*, 64(8), 889-895. https://doi.org/10.1016/j.jbusres.2010.09.006.

Nonnecke, B., & Preece, J. (2001). Why lurkers lurk. Paper presented at the AMCIS 2001, Boston, UK.

Okoli, C., & Oh, W. (2007). Investigating recognition-based performance in an open content community: A social capital perspective. *Information & Management*, 44(3), 240-252. https://doi.org/10.1016/j.im.2006.12.007.

Overby, J. W., & Lee, E.-J. (2006). The effects of utilitarian and hedonic online shopping value on consumer preference and intentions. *Journal of Business Research*, 59(10-11), 1160-1166. https://doi.org/10.1016/j.jbusres.2006.03.008.

Powell, A., Piccoli, G., & Ives, B. (2004). Virtual teams: A review of current literature and directions for future research. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, 35(1), 6-36. https://doi.org/10.1145/968464.968467.

Qu, B. (2010). Research and think of cooperation behavior within the network academic forum taking the emuch forum as an example. *Science and Technology Management Research*, 30(10), 215-218. http://dx.chinadoi.cn/10.3969/j.issn.1000-7695.2010.10.070.

Qu, F., & Zhang, H. Y. (2015). A research on the influence of user experience on the continuance intention of non-trade virtual community. *Journal of Intelligence*, 34(9), 185-191. http://www.qbzz.net/oa/DArticle.aspx?type=view&id=150920.

Qu, F., Zhang, H. Y., & Chen, T. (2016). Research on the relationship between the user experience of virtual community and the intention of information continuous use in colleges. *Research on Library Science*, 3, 64-71. https://doi.org/10.15941/j.cnki.issn1001-0424.2016.03.013.

Reagans, R., & McEvily, B. (2003). Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly*, 48(2), 240-267. https://doi.org/10.2307/3556658.

Reina, D. S., & Reina, M. L. (2006). *Trust & betrayal in the workplace: Building effective relationships in your organization*. 2nd ed. Berrett-Koehler.

Rheingold, H. (1993). *The virtual community: Finding connection in a computerized world*. Addison-Wesley Longman.

Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., & Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2), 578-586. https://doi.org/10.1016/j.chb.2008.12.024.

Ryan, T., & Xenos, S. (2011). Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*, 27(5), 1658-1664. https://doi.org/10.1016/j.chb.2011.02.004.

Schmitt, B. H. (1999). *Experiential marketing: How to get customers to sense, feel, think, act, and relate to your company and brands*. Free Press.

Shade, L. R. (1995). Wired in the ivory tower: Access and copyright issues surrounding the Internet and higher education in North America. *Education for Information*, 13(3), 211-228. https://doi.org/10.3233/efi-1995-13304.

Uzzi, B., & Lancaster, R. (2003). Relational embeddedness and learning: The case of bank loan managers and their clients. *Management Science*, 49(4), 383-399. https://doi.org/10.1287/mnsc.49.4.383.14427.

Vangen, S., Huxham, C. (2003). *Building trust in inter-organizational collaboration*. Paper presented at the EURAM 2003, Milan, Italy.

Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring the hedonic and utilitarian dimensions of consumer attitude. *Journal of Marketing Research*, 40(3), 310-320. https://doi.org/10.1509/jmkr.40.3.310.19238.

Wang, J.-L., Jackson, L. A., Zhang, D.-J., & Su, Z.-Q. (2012). The relationships among the Big Five Personality factors, self-esteem, narcissism, and sensation-seeking to Chinese University students' uses of social networking sites (SNSs). *Computers in Human Behavior*, 28(6), 2313-2319. https://doi.org/10.1016/j.chb.2012.07.001.

Yoo, K.-H., & Gretzel, U. (2011). Influence of personality on travel-related consumer-generated media creation. *Computers in Human Behavior*, 27(2), 609-621. https://doi.org/10.1016/j.chb.2010.05.002.

Zaheer, A., McEvily, B., & Perrone, V. (1998). Does trust matter? Exploring the effects of interorganizational and interpersonal trust on performance. *Organization Science*, 9(2), 141-159. https://doi.org/10.1287/orsc.9.2.141.

Ziman, J. Z. (1994). *Prometheus bound: Science in a dynamic steady state*. Cambridge University Press.

APPENDIX. Data collection tool and measurement items

The 7-point Likert scale was used in the survey of questionnaires (i.e., ranging from 1=strongly disagree to 7=strongly agree). The questionnaire had 43 items, and all derived from the previous maturity scale. We used a pilot study to correct ambiguous places. The questionnaires were made on the Wenjuanxing Website, and sent to the relevant VACs.

Self-efficiency

SE1: I have confidence in my ability to provide knowledge that other members in this virtual community consider valuable.
SE2: I have the expertise, experiences, and insights needed to provide knowledge valuable for other members of this virtual community.
SE3: I have confidence in responding or adding comments to messages or articles posted by other members of this virtual community.

Agreeableness

AG1: Some people think I'm selfish and egotistical.
AG2: I tend to be cynical and skeptical of others' intentions.
AG3: Some people think of me as cold and calculating.

Openness

OP1: Poetry has little or no effect on me.
OP2: Sometimes when I am reading poetry or looking at a work of art, I feel a chill or wave of excitement.
OP3: I have little interest in speculating on the nature of the universe or the human condition.

Extroversion

EX1: I often feel as if I'm bursting with energy.
EX2: I am a cheerful, high-spirited person.
EX3: I really enjoy talking to people.

Conscientiousness

CO1: I am a productive person who always gets the job done.
CO2: I strive for excellence in everything I do.
CO3: I try to perform all the tasks assigned to me conscientiously.

Neuroticism

NEU1: I am seldom sad or depressed.
NEU2: I rarely feel lonely or blue.
NEU3: I rarely feel fearful or anxious.

Experience

UE1: I can get the information or knowledge I need in this community.
UE2: When I'm in trouble, I come to the community for help.
UE3: The community is well organized and rich in content.
UE4: I can relax in the community.
UE5: I feel less tense in the community.
UE6: Joining the virtual academic communities makes me meet new research partners.
UE7: I can get support and encouragement from other members in the community.
UE8: I can communicate with people who have similar ideas in the community.

Reciprocal

RB1: When I share my knowledge through VACs, I believe that I will get an answer for answering.
RB2: When I share my knowledge through VACs, I expect somebody to respond when I'm in need.

RB3: When I contribute knowledge to VACs, I expect to get back knowledge when I need it.
RB4: When I share my knowledge through VACs, I believe that my queries for knowledge will be answered in the future.

Trust

TR1: Management at the VACs shows concern for members' needs.
TR2: Management at the VACs does not take advantage of members when the opportunity arises.
TR3: Management at the VACs is very capable of performing tasks in managing the community.
TR4: Researchers in the VACs do not take advantage of other members when the opportunity arises.
TR5: The other members at VACs would not knowingly do anything to disrupt conversation/interactions.
TR6: The other members at VACs have adequate knowledge about the subject we discuss.

Social networks characteristic

CS1: I maintain close social relationships with some members in the virtual learning community.
CS2: I intensively exchange ideas with some members in the virtual learning community.
ND1: I know some members in the virtual learning community on a personal level.
ND2: I have frequent communication with some members in the virtual learning community.

Research collaboration intention

RCI1: If I can, I would like to continue collaborating with others at VACs in the future.
RCI2: I will likely continue collaborating with others at VACs in the future.
RCI3: I expect to continue collaborating with others at VACs in the future.

 https://doi.org/10.1633/JISTaP.2021.9.2.6

Call for Paper

Journal of Information Science Theory and Practice (JISTaP)

We would like to invite you to submit or recommend papers to Journal of Information Science Theory and Practice (JISTaP, eISSN: 2287-4577, pISSN: 2287-9099), a fast track peer-reviewed and no-fee open access academic journal published by Korea Institute of Science and Technology Information (KISTI), which is a government-funded research institute providing STI services to support high-tech R&D for researchers in Korea. JISTaP marks a transition from Journal of Information Management to an English-language international journal in the area of library and information science.

JISTaP aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics.

We welcome materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. Topics covered by the journal include: information processing and management; information policy; library management; knowledge organization; metadata and classification; information seeking; information retrieval; information systems; scientific and technical information service; human-computer interaction; social media design; analytics; scholarly communication and bibliometrics. Above all, we encourage submissions of catalytic nature that explore the question of how theory can be applied to solve real world problems in the broad discipline of information science.

Co-Editors in Chief : Gary Marchionini & Dong-Geun Oh

Please click the "Online Submission" link in the JISTaP website (http://www.jistap.org), which will take you to a login/account creation page. Please consult the "Author's Guide" page to prepare your manuscript according to the JISTaP manuscript guidelines.

Any question? Ji-Young Kim (managing editor) : jistap@kisti.re.kr

JISTaP
Journal of Information Science Theory and Practice

Information for Authors

The Journal of Information Science Theory and Practice (JISTaP), which is published quarterly by the Korea Institute of Science and Technology (KISTI), welcomes materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. JISTaP is an open access journal run under the Open Access Policy. See the section on Open Access for detailed information on the Open Access Policy.

A. Originality and Copyright

All submissions must be original, unpublished, and not under consideration for publication elsewhere. Once an article is accepted for publication, all papers are accessible to all users at no cost. If used for other researches, its source should be indicated in an appropriate manner and the content can only be used for uncommercial purpose under Creative Commons license.

B. Peer Review

All submitted manuscripts undergo a single-blind peer review process in which the identities of the reviewers are withheld from the authors.

C. Manuscript Submission

Authors should submit their manuscripts online via Article Contribution Management System (ACOMS). Online submission facilitates processing and reviewing of submitted articles, thereby substantially shortening the paper lifecycle from submission to publication. After checking the manuscript's compliance to the Manuscript Guidelines, please follow the "Online Submission" hyperlink in the top navigation menu to begin the online manuscript submission process.

D. Open Access

With the KISTI's Open Access Policy, authors can choose open access and retain their copyright or opt for the normal publication process with a copyright transfer. If authors choose open access, their manuscripts become freely available to public under Creative Commons license. Open access articles are automatically archived in the KISTI's open access repository (KPubS, www.kpubs.org). If authors do not choose open access, access to their articles will be restricted to journal users.

E. Manuscript Guidelines

Manuscripts that do not adhere to the guidelines outlined below will be returned for correction. Please read the guidelines carefully and make sure the manuscript follows the guidelines as specified. We strongly recommend that authors download and use the manuscript template in preparing their submissions.

1. Page Layout :

All articles should be submitted in single column text on standard Letter Size paper (21.59 × 27.94 cm) with normal margins.

2. Length :

Manuscripts should normally be between 4,500 and 9,000 words (10 to 20 pages).

3. File Type :

Articles should be submitted in Microsoft Word format. To facilitate the manuscript preparation process and speed up the publication process, please use the manuscript template.

4. Text Style :

- Use a standard font (e.g., Times New Roman) no smaller than size 10.
- Use single line spacing for paragraphs.
- Use footnotes to provide additional information peripheral to the text. Footnotes to tables should be marked by superscript lowercase letters or asterisks.

5. Title Page :

The title page should start with a concise but descriptive title and the full names of authors along with their affiliations and contact information (i.e., postal and email addresses). An abstract of 150 to 250 words should appear below the title and authors, followed by keywords (4 to 6).

Author1
Affilliation, Postal Address. E-mail
Author2
Affilliation, Postal Address. E-mail

ABSTRACT
A brief summary (150-250 words) of the paper goes here.

Keywords: 4 to 6 Keywords, separated by commas.

6. Numbered Type :

1. INTRODUCTION

All articles should be submitted in single column text on standard letter size paper (21.59 × 27.94 cm) with normal margins[1. Text should be in 11-point standard font (e.g., Times New Roman) with single line spacing. [1 Normal margin dimensions are 3 cm from the top and 2.54 cm from the bottom and sides.

2. SECTIONS

The top-level section heading should be in 14-point bold all uppercase letters.

2.1. Subsection Heading 1

The first-level subsection heading should be in 12-point bold with the first letter of each word capitalized.

2.1.1. Subsection Heading 2

The second-level subsection heading should be in 11-point italic with the first letter of each word capitalized.

7. Figures and Tables :

All figures and tables should be placed at the end of the manuscript after the reference list. To note the placement of figures and tables in text, "Insert Table (or Figure) # here" should be inserted in appropriate places. Please use high resolution graphics whenever possible and make sure figures and tables can be easily resized and moved.

Figure

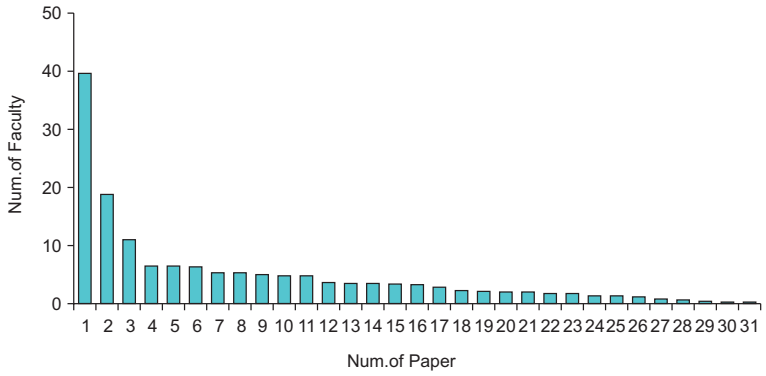


Fig. 1. Distribution of authors over publication count.

Table

Table 1. The title of table goes here

Study	Time period study	Data
Smith Wesson (1996)	1970 - 1995	684 papers in 4 SSCI journals
Reeves [a (2002)	1997 - 2001	597 papers in 3 SSCI journals
Jones Wilson [b (2011)	2000 - 2009	2,166 papers in 4 SSCI journals

[a Table footnote a goes here
[b Table footnote b goes here

8. Acknowledgements :

Acknowledgements should appear in a separate section before the reference list.

9. Conflicts of Interest :

Authors of manuscripts must disclose any potential conflicts of interest at the time of submission. Statements on conflict of interest have no influence on the editorial decision to publish.

10. Citations :

Citations in text should follow the author-date method (authors' surname followed by publication year).

- Several studies found... (Barakat et al., 1995; Garfield, 1955; Meho & Yang, 2007).
- In a recent study (Smith & Jones, 2011)...
- Smith and Jones (2011) investigated...

11. Reference List :

Reference list, formatted in accordance with the American Psychological Association (APA) style, should be alphabetized by the first authors last name.

Journal article

- Author, A., Author, B. & Author, C. (Year). Article title. *Journal Title*, volume(issue), start page-end page.
- Smith, K., Jones, L. J., & Brown, M. (2012). Effect of Asian citation databases on the impact factor. *Journal of Information Science Practice and Theory*, 1(2), 21-34.

Book

- Author, A., & Author, B. (Year). *Book title*. Publisher Name.
- Smith, K., Jones, L. J., & Brown, M. (2012). *Citation patterns of Asian scholars*. Sage.

Book chapter

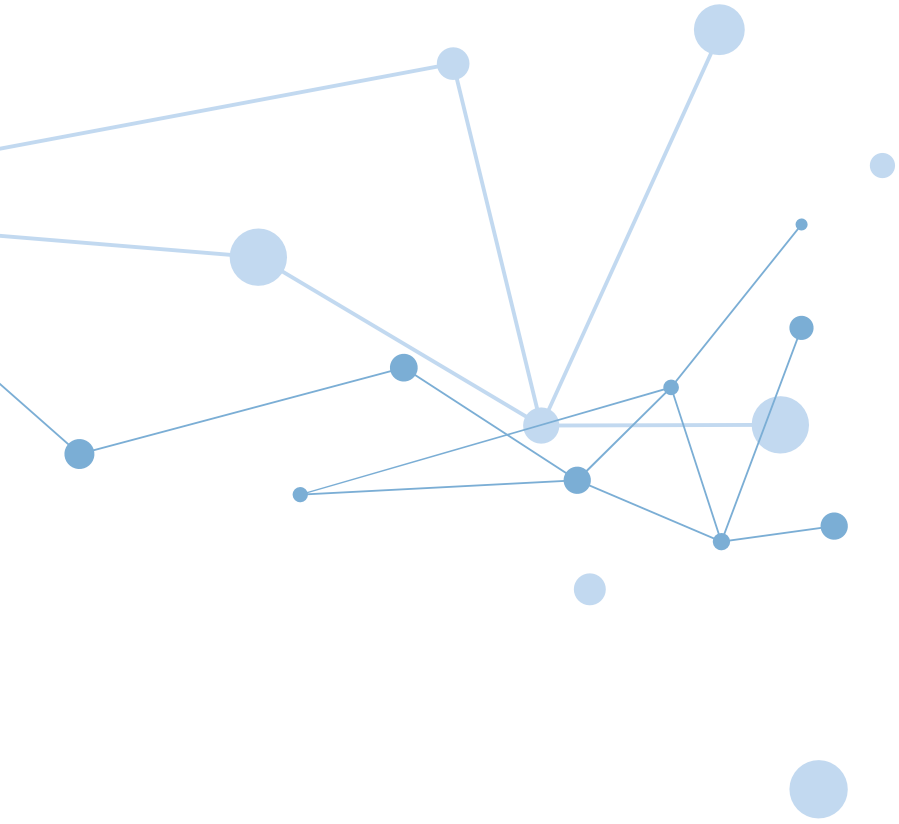
- Author, A., & Author, B. (Year). Chapter title. In A. Editor, B. Editor, & C. Editor (Eds.), *Book title* (pp. xx-xx). Publisher Name.
- Smith, K. & Brown, M. (2012). Author impact factor by weighted citation counts. In G. Martin (Ed.), *Bibliometric approach to quality assessment* (pp. 101-121). Springer.

Conference paper

- Author, A., & Author, B. (Year, Month Day). Article title. In A. Editor & B. Editor (Eds.), *Conference title* (pp. xx-xx). Publisher Name.
- Lin, H., Sun, B., Wu, J. & Xiong, H. (2016, June 24-26). Topic detection from short text: A term-based consensus clustering method. In B. Yang (Ed.), *2016 13th International Conference on Service Systems and Service Management (ICSSSM 2016)* (pp. 1-6). IEEE.

Online document

- Author, A., & Author, B. (Year). *Article title*. URL.
- Smith, K. & Brown, M. (2010). The future of digital library in Asia. *Digital Libraries*, 7,111-119. http://www.diglib.org/publist.htm.



JISTaP

Journal of Information
Science Theory and Practice