International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 1
ISSN: 2088-5334

A Bisociated Research Paper Recommendation Model using BiSOLinkers

Benard M. Maake [a,d,*], Sunday O. Ojo [b], Keneilwe Zuva [c], Fredrick A. Mzee [d]

[a] *Department of Computer Systems Engineering, Tshwane University of Technology, Staatsartillerie Rd, Pretoria, 0183, South Africa*
[b] *Department of Information Technology, Durban University of Technology, Durban, South Africa*
[c] *Department of Computer Science, University of Botswana, Private Bag 0022, Gaborone, Botswana*
[d] *Department of Computing Sciences, Kisii University, P.O Box 408-40200, Kisii, Kenya*

*Corresponding author: *bmaake@kisiiuniversity.ac.ke*

***Abstract*- In the current days of information overload, it is nearly impossible to obtain a form of relevant knowledge from massive information repositories without using information retrieval and filtering tools. The academic field daily receives lots of research articles, thus making it virtually impossible for researchers to trace and retrieve important articles for their research work. Unfortunately, the tools used to search, retrieve and recommend relevant research papers suggest similar articles based on the user profile characteristic, resulting in the overspecialization problem whereby recommendations are boring, similar, and uninteresting. We attempt to address this problem by recommending research papers from domains considered unrelated and unconnected. This is achieved through identifying bridging concepts that can bridge these two unrelated domains through their outlying concepts – BiSOLinkers. We modeled a bisociation framework using graph theory and text mining technologies. Machine learning algorithms were utilized to identify outliers within the dataset, and the accuracy achieved by most algorithms was between 96.30% and 99.49%, suggesting that the classifiers accurately classified and identified the outliers. We additionally utilized the Latent Dirichlet Allocation (LDA) algorithm to identify the topics bridging the two unrelated domains at their point of intersection. BisoNets were finally generated, conceptually demonstrating how the two unrelated domains were linked, necessitating cross-domain recommendations. Hence, it is established that recommender systems' overspecialization can be addressed by combining bisociation, topic modeling, and text mining approaches.**

***Keywords*— Bisociation; data mining; knowledge discovery; recommender system; serendipity; text mining; topic modeling.**

Manuscript received 3 Jan. 2021; revised 28 Apr. 2021; accepted 21 May 2021. Date of publication 28 Feb. 2022.

I. INTRODUCTION

The massive digitization of several aspects of our lives is attributed to the rapid development of information and communication technologies (ICT). Hence, the digital information space is exponentially growing by the day, and access to knowledge and information is being hindered due to information overload [1], [2]. It was estimated in the year 2015 that an average American was consuming 15.5 hours of media content and nearly 74 Gigabytes of information every single day. In academia, recent bibliometrics shows that the number of published scientific papers has been climbing by 8% - 9% each year over the past several decades. For instance, the PubMed database alone receives more than 1 million papers a year in the biomedical field, which is approximately two papers every minute. Hence becoming overwhelmingly hard to navigate the growing deluge of data [3]. Consequently, a critical demand for superior tools which can support web and academic users (researchers) cope with this data deluge that is on the rise [4], [5].

Recommender systems are part of information filtering systems that eliminate unwanted information, presenting users with only useful and relevant data automatically [6], [7]. In today's digital world, it is impossible to stay abreast by reading a few published articles like journals, and at the same time, it is impossible to read all the journals that have been released. Therefore, researchers must identify the right tools to help them overcome the problem of information overload in academia while receiving satisfactory recommendations from trusted sources [3], [8].

Recommender systems are tools that have been developed to navigate complex information spaces facilitating efficiency, productivity, and health of all its users. These systems have been deployed in various fields such as music [9]-[11], video [12]-[14], mobile [15], [16], research papers [17]-[21]. In academia, research-paper recommender systems (RPRS) are

the tools developed to know more about the users to recommend relevant articles [22] better. RPRS algorithms are designed to recommend articles similar to the target paper or user profiles, leading to a problem known as the portfolio effect [24], [25].

This undesirable effect is generated by recommendation algorithms that concentrate on prediction accuracy while sacrificing other important aspects that improve the user experience of receiving recommendations, such as serendipity, variety, and coverage [26], [27]. McNee *et al.* [28] indicated that the most accurate recommendations are not necessarily the most useful recommendations [28]. Therefore, we design to make RPRS more useful by addressing the overspecialization problem by linking domains that are considered unrelated, thereby facilitating serendipitous recommendations.

The concept of bisociative knowledge discovery is attributed to Koestler [29], who stated that "Two concepts are bisociated if and only if there is no direct link, obvious evidence linking them, and, one concept has to cross contexts to find the link and the new link provides some novel insights." For that reason, this research is aimed at modeling a research paper recommendation model of a system that seeks to establish latent links and relations that might exist between two or more unrelated domains. If there exist links, then a graphical network representation of the bisociation will be depicted through BisoNets. Then finally, relevant concepts and research papers will be recommended across these two large and normally unconnected information spaces using exploratory creativity discovery methods [30].

Topic models are based on the idea that there lies a mixture of topics in a text corpus whereby a topic is a multinomial distribution over words. Due to the textual nature of research papers, topic modeling was utilized by Pan and Li [31] as a means of recommending research papers, and thematic similarity measured how the text was interrelated. The cold start problem was addressed by generating recommendations using topic analysis.

In many RPRS, the number of researchers (users) compared to the number of papers (items) does not balance, and this results in the data sparsity problem where no item is rated or recommended as useful. Moreover, to solve that problem, Amami *et al.* [32] proposed a model where research papers and users were subjected to language and topic modeling respectively to determine the relationship between users and papers and based on the determined closeness of the language used in research papers, unseen research papers were retrieved. Ahmad and Fuge [33] used topic modeling to ensure that the right set of words and topics were utilized to bridge two unrelated domains, achieving contiguity of creative solutions through mediation, similarity, and serendipity [34]-[36]. Lastly, Ahmad and Fuge [33] used topic models to discover topical links that bridge two unrelated domains. Their proposed model was a computational framework for discovering new connections while supporting creativity and the discovery of novel and new ideas. We extend this framework by identifying the topics in a novel way, then we link and recommend serendipitous research paper concepts between these unrelated domains.

Serendipity is a new dimension in recommender systems used to address overspecialization by improving user satisfaction by recommending novel, interesting and unexpected items. Serendipity algorithms expand user tastes by adding a "surprise me" option to the various recommender systems running the algorithm [37], [38]. Therefore, this research attempts to implement the serendipity concept in the field of RPRS by recommending research papers from unrelated domains.

Overspecialization in RPRS is receiving similar research papers as user profile characteristics indicate, or what the user points out as interesting [39], [40]. This method of computing recommendations limits the possibility of linking more domains to expand the coverage of recommended articles. Unfortunately, explicitly defining the user's domain of interest during recommendations promotes recommending highly similar items (research paper articles) from a single domain [41]. Hence, we postulate that the integration of information from different domains into a single network will enable cross-domain recommendations and associations - bisociation [20], [42], [43], which will consequently facilitate cross-domain recommendations, thereby addressing the overspecialization problem.

II. Materials and Methods

A. Dataset Used for Experimentation

The dataset used in this research is a well-researched migraine-magnesium domain pair introduced by Swanson [44] and utilized in evaluating developed metrics that identified bridging terms between two unrelated domains. In Juršič *et al.* [45], 60 pairs of articles were discovered in a literature-based discovery process when the standard 43 bridging terms were used. Research paper titles were retrieved from the PubMed database using two keyword queries for the separate domains to get the dataset. The first keyword used was "migraine" for the migraine domain, while the second keyword used was "magnesium" for the magnesium domain. An additional condition to the query was the restriction used on the range of publishing dates for the articles. To reproduce the bridging terms that were discovered by Swanson [44], one needed to download from PubMed research paper titles published not later than 1988.

In our research, we queried the PubMed research paper repository using the magnesium and migraine keywords, respectively, and a total of 3360 migraine and 8843 magnesium titles were returned. Cumulatively, 12203 research paper titles were retrieved from both domains. These titles were used in this research to determine whether the two domains were related in any way.

B. Bisociation Methodology

Koestler [29] stated that bisociation is a combinational problem that joins unrelated and often conflicting information differently. The questions were asked to identify good bisociation relations, including what is bridging a domain? What are the ways utilized to determine creative bridges? How can creative bridges be represented conceptually? The domain bridging is linking domains that were otherwise considered unconnected [45]. Creative bridges can be determined using text mining technologies, represented conceptually as graphs.

A network representation of bisociation is a BisoNet (Bisociative Information Network) which supports the integration of semantically meaningful information and loosely coupled information fragments [42]. Given topics from a domain D, we proposed a method of ranking concepts from two habitually separated domains. The ranked concepts were then utilized to showcase their proficiency of generating BisoNets. The steps utilized to solve these problems included: (i) Learning each term representation in a document d and domain D. (ii) Discovering candidates for bridging terms. (iii) Constructing BisoNets from highly probable bridging terms or concepts.

Let $M1$ and $M2$ represent two unrelated and unconnected domains such that $(M1 \nsim M2)$ and thus cannot be linked together. Let concepts that are related and existing between these two domains be represented as a problem π. Let the concepts that make up this problem be referred to as bridging concepts/ terms, X, such that $\{c1, c2, c3, c4, c5\} \in X$. Bridging concepts that intersect between two domains are called BiSOLinkers (Bisociated Serendipitous Outlier Linkers) since they form part of outliers' concepts from the two unrelated domains that link the domains serendipitously. Let the intersection between these unrelated domains be represented by b, such that $b_{c_i,\dots,c_n} = c_{i,\dots,n} \in M1 \cap M2$, where $c_i, \dots, c_n \in b$ represents the BiSOLinker candidate concepts. Concepts $\{c2, c3 \in b\}$ are true BiSOLinkers since they belong to b and X.

To identify the terms and concepts in these two domains, we had to take research papers from both domains and preprocess them to individual terms [43]. Then we used machine learning algorithms to identify the outliers, and finally, topic modeling algorithms [46] were finally utilized to discover the concepts that were intersecting and linking the two domains. In summary, bisociation may be revealed through link discovery, graph mining, and computer-aided interactive navigation (which can be explained using graph structures).

C. BisoNet Formation

Research paper text were decomposed into concepts and terms which represented the vertices of a BisoNet. These vertices were terms that ranged from one word to n-gram terms. We utilized the term frequency-inverse document frequency (tf-idf) metric to identify and select important keywords.

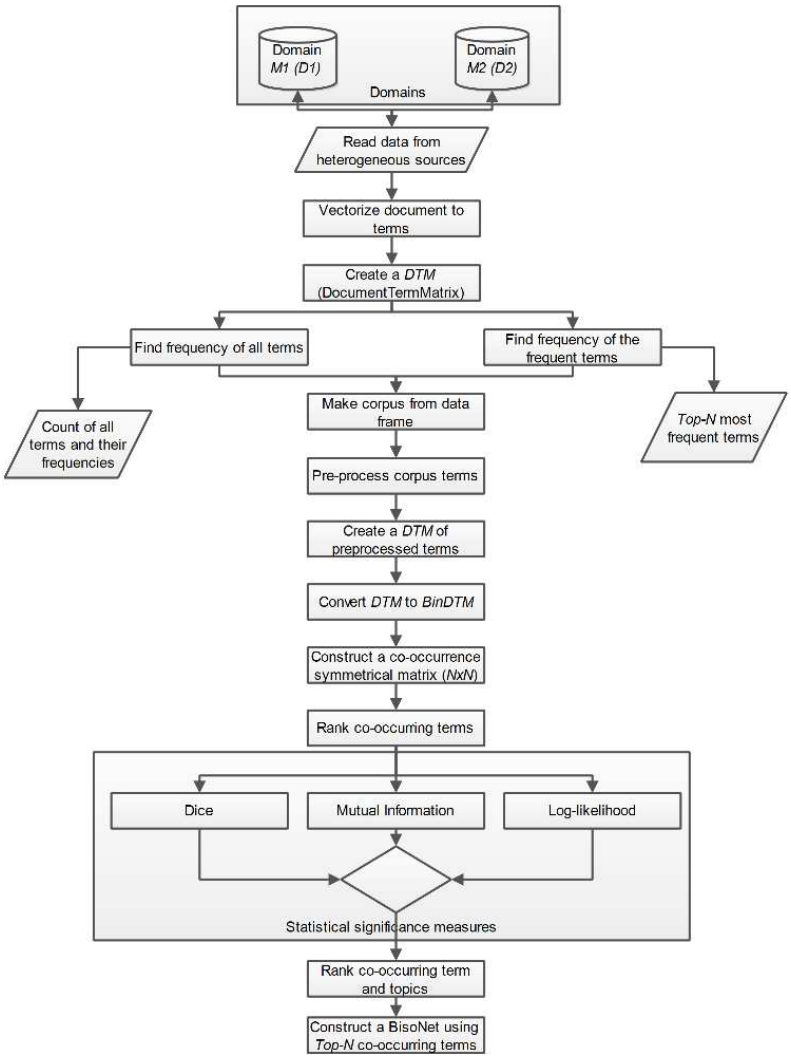

Fig. 1 BisoNet formation

A node contained a list of term frequency values of its associated term as an attribute in the different collections of documents. Vertices with a certain term frequency were associated with a keyword and a set of documents in which

they occurred, and to determine whether a link was to be created between any two vertices, weights were assigned based on co-occurrence significance measures between vertices (weights between vertices signify the existence of a relationship), as in Figure 1.

D. Outlier Identification and BiSOLinkers Construction

The principal approach to determining outliers in many domains is training a multi-class machine learning classifier to distinguish between the labels. If there are documents of one class and they are consistently misclassified as belonging to another class (False Negative), they form a part of documents (outliers) that are likely to bridge the two domains. To identify the outliers within the migraine-magnesium domains, our system read datasets from the two domains and then mixed them with their labels assigned to their correct classes. Cleaning and preprocessing the data was undertaken, then terms appearing more than ten times (>10) were only selected for further processes. A data frame of these terms was created, further weighting them with the tf-idf measure. The data was then separated into a training set (80%) and a testing set (20%). Five classification algorithms were employed to classify the documents using trained classifiers. The accuracy, computation time, and a number of misclassified research paper titles were all noted. All the falsely classified documents (False Negative) were further utilized in the formation of BiSOLinkers.

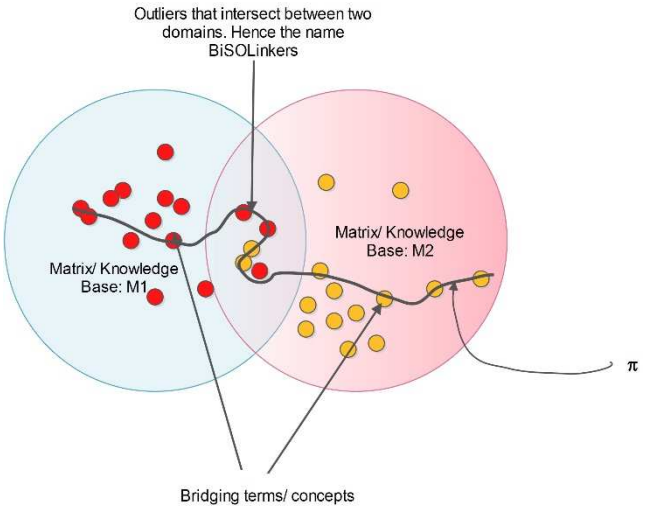

Fig. 2 Bisociated domains intersecting resulting in the identification of BisoLinkers

To produce topics from documents, one must use topic modeling techniques, which current topics as clusters of similar words. These techniques are expressed in mathematical frameworks that allow them to examine text documents (like research papers), ascertaining the statistics of all words and determining the type of topics that might be in a document and the supposed topical balance of each document. We derived key insight from topic modeling by using topics as "bridges" between two domains. This was easily achieved by clustering terms into topics containing some semantic relations, hence ensuring clearer frames and contexts, and ambiguity in related terms was significantly reduced. We were looking at identifying infrequent topics to bring discovery or understanding to a particular problem but again common enough to cross domains through their outliers. We defined these outlier topics as *BiSOLinkers* (**Bi**sociated **S**erendipitous **O**utlier **Linkers**), see Fig. 2.

BiSOLinkers are considered outliers in two habitually incompatible domains, but they fortuitously linked the two dissimilar domains (migraine-magnesium). Sluban *et al.* [47] stated that documents of a paired domain (union of two different domain literature) misclassified by a classifier can be considered domain outliers. They further stated that outlier documents frequently embody new information and can potentially lead to new knowledge. Ahmed and Fuge [33] generated BisoNets using topics employing textual data and pointed out that topics among outliers had a high chance of having a very high bisociation score. In our research, we also exploited outliers that intersect the migraine and magnesium domains, and we did so to identify topic proportions within domains, documents, and outliers to exploit relevant concepts for recommendations between both domains.

E. Topic Modeling in Outliers (BiSOLinkers) of Bisociated Domains

Outliers identified by the five classifiers were further processed with topic modeling algorithms to establish whether similar concepts were lying in those outliers' items, plus whether they would act as links between the two unrelated domains. We utilized the Latent Dirichlet Allocation (LDA) algorithm [46], [48], which is a topic modeling technique that assumes that documents (text corpora) are nothing but a mixture of topics, and it further speculates that these topics overlap within a document even though they are not known beforehand. Hence, the desired number of topics to be mined from the document or text corpus has to be specified beforehand so that the model can generate the latent topics that exist within the document. Each document will have a distribution of topics, and each topic will also have a distribution of words, all generated by simple probabilistic procedures.

LDA is, therefore, an algorithm that was used to describe and identify the mixture of topics that were contained in a research paper document such that $P(z|d)$, and each topic being also described by a distribution of words, such that $P(t|z)$. To formalize this representation, we utilized the following expression:

$$P(t_i|d) = \sum_{j=1}^{z} P(t_i|z_i = j)\, P(z_i = j|d) \qquad (1)$$

where $P(t_i|d)$ represented the probability of the ith term or word within a document d, whereas z_i represented the latent topics that were yet to be discovered. $P(t_i|d)$ represented the probability of a particular term t_i within a topic j. Lastly, $P(z_i = j|d)$ represented the probability of a word being generated from topic j within a document d.

LDA has been described to be a bag of words model, implying that the word order within the document does not count or affect the document representation. Therefore, this means that all unimportant words and rare terms like stop words should be removed from the text corpus to avoid the model from overcompensating for every frequent term and word, which again do not contribute to the generation of topics. In order to also control the granularity of differences that might be included within latent topics, the number of latent topics is well-defined in advance [48], [49]. The

distribution used to draw the per-document topic distribution is known as a Dirichlet distribution, which allocates words of a document to different topics, and it took the following form:

$$P(\theta|\vec{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \quad (2)$$

which is a distribution over probability distributions. After training the corpus, each respective document had a distribution of overall topics, and each topic had a distribution over all the terms.

There are several ways of implementing LDA, but in this research, we utilized the learning algorithm that is based on Gibbs sampling [50], [51], which works as follows: The model first initializes by assigning every word/ term in the document to a random topic. Iterations are then performed on every word, un-assigning its current topic, decrements the topic count corpus wide and then reassigns the word to a new topic based on the local probability of the topic assigned to the current document, and global (corpus wide) probability of the work assigned to the current topic. These multiple iterations over the word t_i in document d_i, will sample a new topic j based on the distribution represented in Equation (1), and this will continue until the LDA model parameters converge. This process was represented as Equation (3).

$$P(z_i = j|t_i, d_i, z_{-i}) \propto \frac{C_{t_i j}^{TZ} + \beta}{\sum_t C_{tj}^{TZ} + T\beta} \frac{C_{d_i j}^{DZ} + \alpha}{\sum_z C_{d_i j}^{DZ} + Z\alpha} \quad (3)$$

where C^{TZ} sustained the computation of all topic-word assignments while C^{DZ} maintained a count of all the document-topic assignments. z_{-i} symbolized all assignments for topic word and document topic excluding the current topic z_i for term t_i. Then the symbols α and β represented a smoothing parameter that ensured that the probability never got to 0. The posterior probabilities in Equation (1) were estimated using the following expressions.

$$P(t_i|z_i = j) = \frac{C_{t_i j}^{TZ} + \beta}{\sum_t C_{tj}^{TZ} + T\beta} \quad (4)$$

$$P(z_i = j|d_i) = \frac{C_{d_i j}^{DZ} + \alpha}{\sum_z C_{d_i j}^{DZ} + Z\alpha} \quad (5)$$

After that, to establish the topical similarity between two research paper a and b over their vector coefficients, we utilized an adapted form of the cosine similarity shown below:

$$Topical_{sim(a,b)} = \frac{\overrightarrow{\theta[a]}.\overrightarrow{\theta[b]}}{||\overrightarrow{\theta[a]}|| \, ||\overrightarrow{\theta[b]}||} \quad (6)$$

To identify topics in our dataset, we created a corpus from research papers that were detected as outliers. The preprocessing chain followed with cleaning and transforming the text into a document term matrix (DTM) which had terms with a frequency greater than ten (>10). We loaded the topic models' package in the R programming environment to compute the topic probability distribution over the entire vocabulary, and the five most likely topics in each title were inferred.

Algorithm 1., assimilated from Ahmed and Fuge [33], was utilized to identify the outliers and rank bisociative topics found within the outliers. Let $\mathcal{I}$ be a set of all N research paper documents from $M1$ &$M2$ domains. Let O_d represent outliers for domain d. Let X be a data frame N X T representing a document topic matrix, such that a row i represents the i^{th} document's T dimensional topic proportion vector. For topic t in domain d: Topic bisociation score

$$(t, d) = \frac{\sum_{j \in O_d} X_{j,t}}{\sum_{i \in \mathcal{I}} X_{i,t}} \quad (7)$$

We further utilized five classifiers, namely: Support vector machines, Naïve Bayes, Neural Network, Random Forest, and Logistic Boosting, to help us identify the outliers within the dataset.

Algorithm 1: OrderBisociativeTopicsThroughRanking

Input: A group of domains $\mathcal{D}$
A group of ideas $\mathcal{I}$
A Vector $d_{\mathcal{I}}$ of which domain $d \in \mathcal{D}$ each idea $i \in \mathcal{I}$ belongs to
A domain query $q \in \mathcal{D}$
Output: ranked list of bisociated topic scores w.r.t q

```
topics, X ← IDEASVECTORIZED (I)
∅ ← FINDOUTLIERS (X, d_I)
topicalScores = ∑_{j∈O_d} X_{j,t} / ∑_{i∈I} X_{i,t}
return topics.rankedBy(topicalScores)

def IDEASVECTORIZED (I):
    topics, X = executeTopicModelingLDA (I)
    return topics, X
def DETECTOUTLIERS(X, d_I)
    classifier = trainDomainClassifier (X, d_I)
    D_predicted = classifier.predictProbabilitiesInDomains (X)
    outliers ← ∅
    for i ∈ X do
        d_true ← d_I[i]
        d_predicted ← argmax_{d∈D} D_predicted [i, d]
        If d_true ≠ d_predicted then
            outliers ← outliers ∪ i
    return outliers
```

F. BisoNets Generation

A BisoNet (Bisociation Network) is a graphical representation of the bisociative relationships that exist between unrelated domains. This graph structure was represented as follows:

Let V represent the vertices of a graph (information units), and E represent the edges of the graph (relationships between the information units). Let λ represent the label on each node and ω the weight between the vertices.

$$B = (V_1, \ldots, V_k, E, \lambda, \omega) \quad (8)$$

Therefore, a BisoNet is a graphical representation of bisociation between unrelated domains with attributes greater than two partitions, $k \geq 2$. Unlike BisoNets created from document terms, it is possible to have vertices generated from topics, with edges representing how strong one topic is from another. BisoNets in this research were then generated from the topics. Fig. 3 displays a BisoNet generated from our dataset.

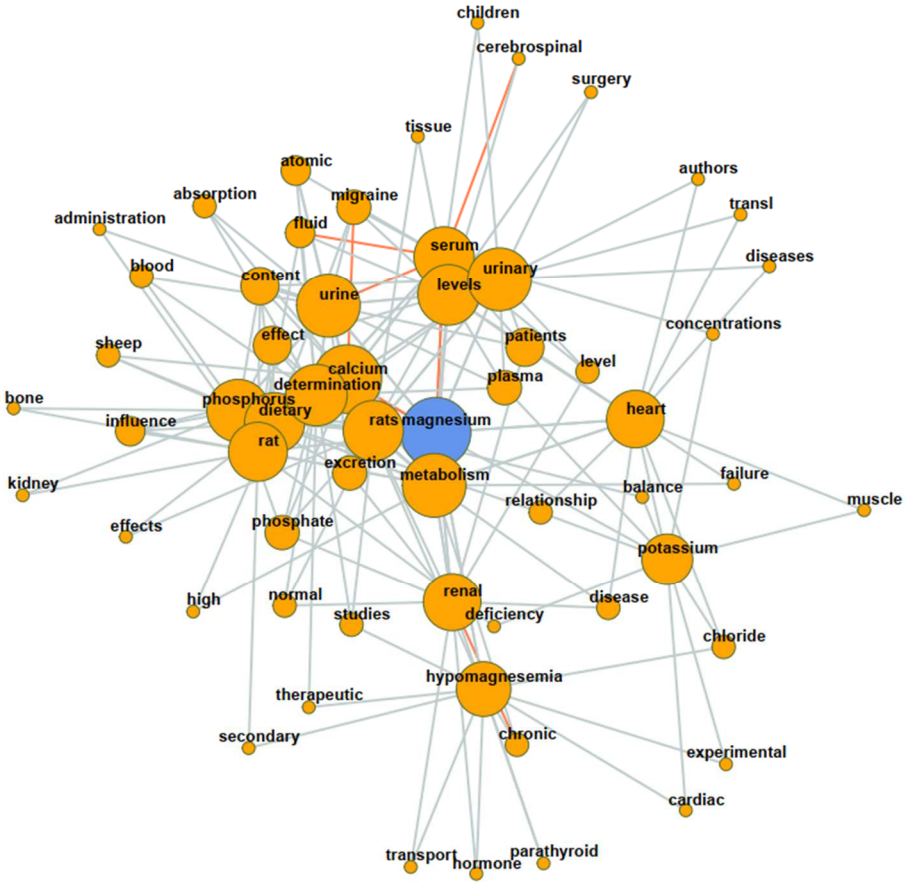

Fig. 3 BisoNet developed from two habitually unrelated domains

G. Recommendations Between Bisociated Domains

Once two previously unrelated and unconnected domains are connected together using BisoNets, it becomes possible to exploit this newly discovered association (Bisociation) to recommend the domains. Let I_r represent an item or concept from one domain labeled M_1 that is qualified (among concepts that make the problem π) to be sent as a recommendation to another completely unrelated and unconnected domain M_2 .

For this item to be qualified as a novel and serendipitous recommendation, then it must meet the following requirements: (i) The concept item to be recommended must be a bridging concept between two unrelated domains, such that the concept $c_i \in \pi$. (ii) The concept must be a BiSOLinker such that $c_i \in (M_1 \cap M_2)$. (iii) The concept must be selected from novel items and concepts in the originating domain [52]. (iv) The recommended concept should not be recommended to its domain of origin.

III. Results and Discussions

A. Bisociation Methodology

To address the problem of overspecialization in RPRS, we proposed a bisociation model that linked two unrelated domains using bridging concepts that formed a common problem between the two unrelated domains. Research paper titles were decomposed into terms (vertices), which again were analyzed to express what topic they represented. Weight between concepts (vertices) represented how related two concepts were, illustrating the concept of bisociation through graphs. Further, outliers were extracted from the two unrelated domains, and topics that could act as links to the unrelated domains were identified. We demonstrated how to bridge and link two unrelated domains to provide possibilities of cross-domain serendipitous recommendations. The developed model was represented as a BisoNets-enabled recommendation model.

B. BisoNet Node Formation and Linking

BisoNet vertices represented concepts that were generated from decomposing research paper titles. These concepts usually range from one word to n-gram term/concept, depending upon the investigation's nature. In this research, text mining techniques were utilized to preprocess the dataset. Similarity measurements between vertices denoted the existence of relationships, and higher weights signified the certainty of an existing strong relationship, whereas lower weights inferred the presence of weak or no relationship.

C. Outlier Identification and BiSOLinker Construction

Machine learning classifiers were used to distinguish between labels in our dataset. All the research paper titles that were consistently misclassified as belonging to another domain were identified to be outliers and possible bridging concepts. The overall accuracy of all the machine learning algorithms except for Naïve Bayes algorithm was relatively good, ranging between 96.30% to 99.49%, see Table 1.

TABLE I
ALGORITHMS UTILIZED TO IDENTIFY OUTLIERS – MISCLASSIFIED TITLES

Algorithm	Accuracy	Time	Titles misclassified
SVM	98.88%	26.90947 min	28
Naïve Bayes	56.19%	10.10075 sec	367
Logits Boost	99.49%	54.86728 sec	8
Random Forest	99.17%	43.79682 sec	8
Neural Network	96.30%	5.33358 sec	36

The poorest machine learning algorithm was the Naïve Bayes classifier that attained an accuracy of 56.19 %, evidencing that it was unsuitable to be used in identifying outliers present in the dataset.

We, therefore, remained with four classifiers for the experiment, and the range in accuracy between the highest and lowest values attained was 3.19%, which indicated that most algorithms had a good classification accuracy, and as a result, the outliers identified by the classifiers were of high quality. The best performing algorithm in terms of accuracy was Logit Boost algorithm having an accuracy of 99.49% and wrongly misclassified 8 research paper titles. Random Forest closely followed this, then SVM, and finally, Neural Network. The SVM classifier took the longest computational time of approximately 30 minutes, whereas Neural Network took the shortest time of approximately 5 seconds, hence implying that different classifiers have different computational performance characteristics with textual data. For that reason, classifiers should be selected for experimentations depending on what is being investigated. Figure 4. displays a comparison of all the accuracies that were obtained from the classifiers. The Naïve Bayes performed dismally, and hence it could not be utilized for outlier detection.

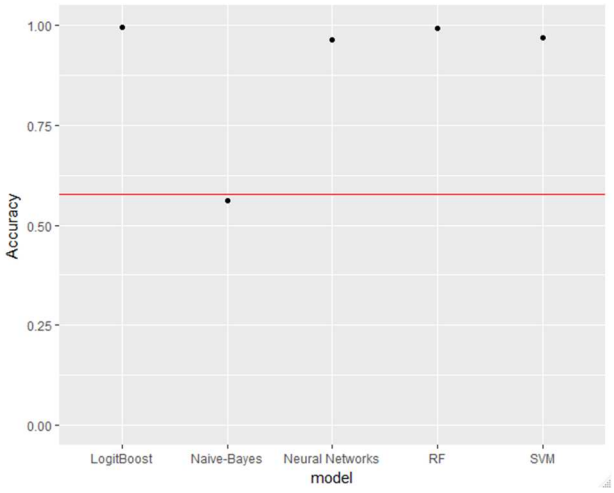

Fig. 4 Accuracy of outlier detection classifiers

D. Topic Modeling in Outliers (BiSOLinkers) of Bisociated Domains

The LDA algorithm was able to infer the five most likely topics in each research paper title in our dataset. We then concatenated these topics to have a pseudo-name representation for each topic, and they can be utilized to query either domain for serendipitous recommendations to the contrasting domain. The term-topic probability distribution of each title was later sorted in decreasing topic proportion within the entire collection, revealing the relevant topics common within the two domains, Fig. 5. Finally, we counted how often a topic appeared as a primary topic within a document, Fig. 6., depicting how significant topics within both domains were. Algorithm 1 effectively executed the process of identifying and ranking relevant topics that were found within outliers.

```
                                migrain children childhood review case  0.05692423
                      migrain treatment headach therapi pathogenesi 0.05667106
                           calcium phosphorus sheep excret influenc 0.05444024
                            migrain headach classic patient attack 0.05405584
                                determin serum urin absorpt biolog 0.05162663
                                studi experiment magnesium diet cow 0.05146366
                  renal secret parathyroid hormon hypermagnesemia 0.05145820
                          level serum urinari children malnutrit 0.05135142
                             rat effect magnesium tissu parenter 0.04967442
                      migrain clinic aspect cardiovascular practic 0.04951580
                             metabol dietari effect rat magnesium 0.04927146
                          renal chronic magnesium metabol failur 0.04902587
                           blood effect magnesium concentr infus 0.04853092
                      metabol diseas magnesium electrolyt patient 0.04813775
                             fluid cerebrospin blood serum brain 0.04763684
                           acut magnesium myocardi muscl infarct 0.04756816
                         plasma effect concentr level erythrocyt 0.04751302
                           defici magnesium test administr diseas 0.04602888
hypomagnesemia primari administr hypocalcemia hypomagnesaemia 0.04485836
                          author transl therapi clinic electrolyt 0.04424722
```

Fig. 5 Topic proportion in decreasing order

```
migrain treatment headach therapi pathogenesis  104
        migrain children childhood review case   94
renal secret parathyroid hormon hypermagnesemia  79
          studi experiment magnesium diet cow    76
           determin serum urin absorpt biolog    73
      calcium phosphorus sheep excret influenc   71
       migrain headach classic patient attack    71
         metabol dietari effect rat magnesium    59
       level serum urinari children malnutrit    55
migrain clinic aspect cardiovascular practic     49
        acut magnesium myocardi muscl infarct    47
     plasma effect concentr level erythrocyt     40
 metabol diseas magnesium electrolyt patient     40
        blood effect magnesium concentr infus    39
      renal chronic magnesium metabol failur     39
      defici magnesium test administr diseas     38
         fluid cerebrospin blood serum brain     37
         rat effect magnesium tissu parenter     35
```

Fig. 6 How often a topic appears as a primary topic within a document

E. BisoNet Generation

BisoNets were generated with an algorithm assimilated from Ahmed and Fuge [33]. This BisoNet revealed how concepts were related to one another and how concepts from other domains were related strongly to concepts in another unrelated domain. Fig. 6 reveals vertices (concepts) of different sizes and different colored edges (relations). The size of the vertex was determined by the frequency of a topic within the domain. Orange links (edges) from one concept to another depicted very strong relations, whereas gray links depicted very low or no significant relation between concepts. The BisoNet graph was constructed from a triple data frame that encoded the source topic, target topic, and relationship weight as an edge, as in Table 2.

TABLE II
A SAMPLE TRIPLE USED TO CREATE A BISONET

	From	To	Significance	Relationship
52	serum	Magnesium	47.64756	Strong
104	Levels	Cerebrospinal	6.686601	Weak
22	Serum	Cerebrospinal	54.593673	Strong
812	Potassium	Plasma	4.269386	Weak
96	Rats	Phosphorus	3.956897	Weak
154	Levels	excretion	3.688140	Weak

To fully construct the BisoNet, the following R programming packages were utilized: tm package for text processing and the igraph package for constructing network visualization representation - BisoNet. On Fig. 7., concepts such as serum, calcium, potassium, phosphorus, and effects had very strong relationships which were utilized for cross-domain (magnesium domain to migraine domain) recommendations.

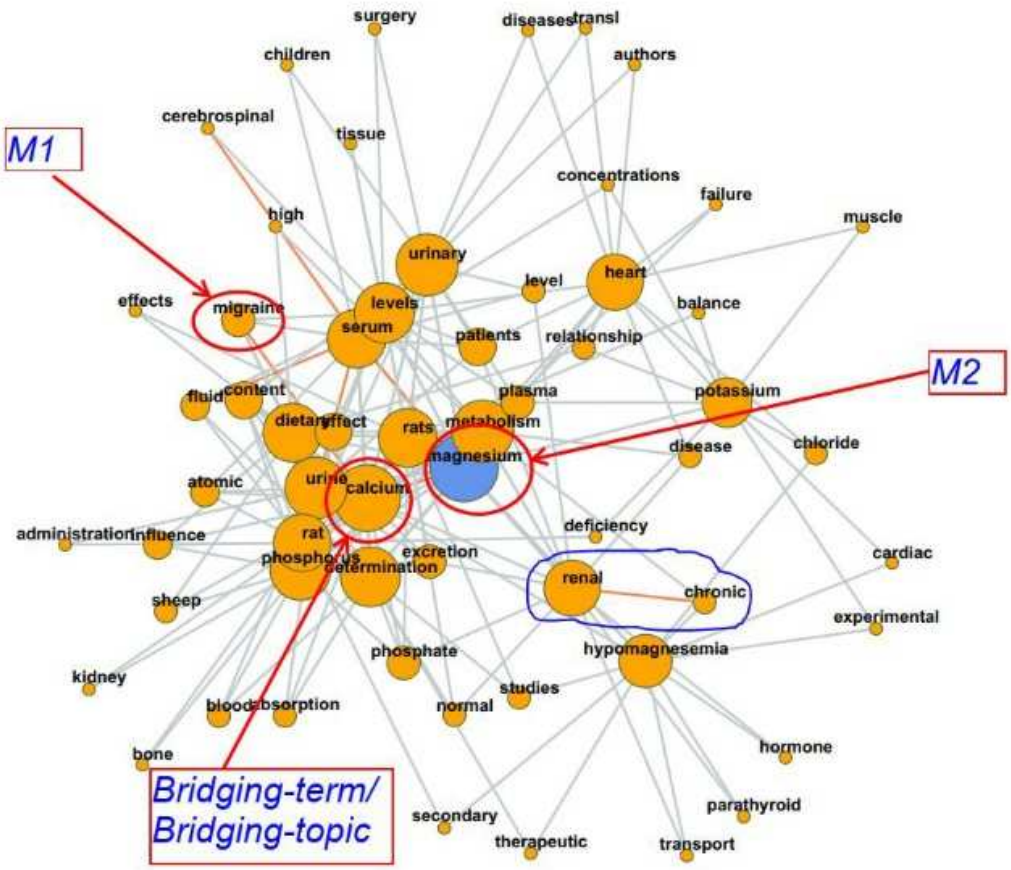

Fig. 7 A BisoNet displaying how concepts are related between two unrelated domains

IV. Conclusions

Bisociation was utilized in this research to address the overspecialization problem when recommendations were suggested from domains considered completely unrelated. Machine learning algorithms were utilized to retrieve outliers, which in turn provided links to the unrelated domains through the presence of BiSOLinkers. Concepts found between the unrelated domains were utilized to create graphical network representations known as BisoNets. Through these networks, it was demonstrated that recommendations were going to be sent across the unrelated domains through BiSOLinkers, resulting in serendipitous articles recommended across the domains. Topic modeling guaranteed that relevant concepts and terms were utilized in the process of recommending titles. Therefore, we have demonstrated that the problem of overspecialization can be addressed by combining bisociation, topic modelling, and text mining techniques, resulting in serendipitous recommendations as illustrated and summarized in Figure 8.

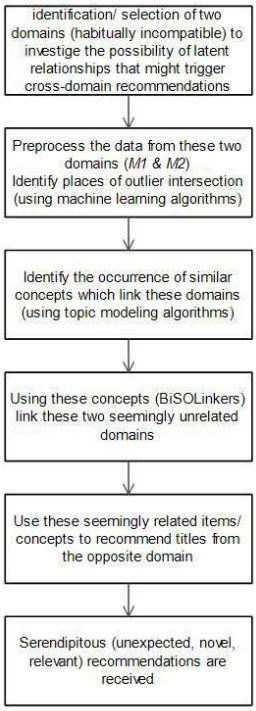

Fig. 8 Summary of major steps in recommending serendipitous items through BiSOLinkers

References

[1] S. Renjith, A. Sreekumar, And M. Jathavedan, "An Extensive Study on The Evolution of Context-Aware Personalized Travel Recommender Systems," *Information Processing & Management,* Vol. 57, P. 102078, 2020.

[2] F. Ferrara, N. Pudota, And C. Tasso, "A Keyphrase-Based Paper Recommender System," In *Digital Libraries and Archives*, Ed: Springer, 2011, Pp. 14-25.

[3] E. Landhuis, "Scientific Literature: Information Overload," *Nature,* Vol. 535, Pp. 457-458, 2016.

[4] Nodus Labs. (2015). *Divinatory Recommender Systems: Between Similarity and Serendipity*. Available: Https://Noduslabs.Com/Research/Divinatory-Recommender-Systems-Similarity-Serendipity/.

[5] B. Cai, X. Zhu, And Y. Qin, "Parameters Optimization of Hybrid Strategy Recommendation Based on Particle Swarm Algorithm," *Expert Systems with Applications,* Vol. 168, P. 114388, 2021/04/15/ 2021.

[6] L. Quijano-Sánchez, I. Cantador, M. E. Cortés-Cediel, And O. Gil, "Recommender Systems for Smart Cities," *Information Systems,* P. 101545, 2020.

[7] S. Sridharan, "Introducing Serendipity in Recommender Systems Through Collaborative Methods," 2014.

[8] F. Amato, V. Moscato, A. Picariello, And F. Piccialli, "Sos: A Multimedia Recommender System for Online Social Networks," *Future Generation Computer Systems,* Vol. 93, Pp. 914-923, 2019/04/01/ 2019.

[9] B. M. Maake, S. O. Ojo, S. Ngwira, And T. Zuva, "Mplist: Context Aware Music Playlist," In *Emerging Technologies and Innovative Business Practices for The Transformation of Societies (Emergitech), Ieee International Conference On*, 2016, Pp. 309-316.

[10] A. Biswal, M. D. Borah, And Z. Hussain, "Music Recommender System Using Restricted Boltzmann Machine with Implicit Feedback," 2021.

[11] M. Schedl, P. Knees, And F. Gouyon, "New Paths in Music Recommender Systems Research," Presented at The Proceedings of The Eleventh Acm Conference on Recommender Systems, Como, Italy, 2017.

[12] C. Bhatt, M. Cooper, And J. Zhao, "Seqsense: Video Recommendation Using Topic Sequence Mining," In *International Conference on Multimedia Modeling*, 2018, Pp. 252-263.

[13] D. Abul-Fottouh, M. Y. Song, And A. Gruzd, "Examining Algorithmic Biases in Youtube's Recommendations of Vaccine Videos," *International Journal of Medical Informatics,* Vol. 140, P. 104175, 2020.

[14] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi*, Et Al.*, "The Youtube Video Recommendation System," In *Proceedings of The Fourth Acm Conference on Recommender Systems*, 2010, Pp. 293-296.

[15] T. Zuva, "Image Content in Shopping Recommender System for Mobile Users," 2012.

[16] D. Horowitz, D. Contreras, And M. Salamó, "Eventaware: A Mobile Recommender System for Events," *Pattern Recognition Letters,* Vol. 105, Pp. 121-134, 2018/04/01/ 2018.

[17] J. Beel, B. Gipp, S. Langer, And C. Breitinger, "Research-Paper Recommender Systems: A Literature Survey," *International Journal on Digital Libraries,* Pp. 1-34, 2015.

[18] C. Nishioka and H. Ogata, "Research Paper Recommender System for University Students on The E-Book System," 2018.

[19] B. M. Maake, S. O. Ojo, And T. Zuva, "Information Processing in Research Paper Recommender System Classes," In *Research Data Access and Management in Modern Libraries*, B. Raj Kumar and B. Paul, Eds., Ed Hershey, Pa, Usa: Igi Global, 2019, Pp. 90-118.

[20] B. M. Maake, S. O. Ojo, And T. Zuva, "A Serendipitous Research Paper Recommender System," *International Journal of Business and Management Studies,* Vol. 11, Pp. 38-53, 2019.

[21] S. Sowmiya and P. Hamsagayathri, "A Collaborative Approach for Course Recommendation System," In *Advances in Smart Grid Technology*, Ed: Springer, 2021, Pp. 527-536.

[22] J. Son and S. B. Kim, "Academic Paper Recommender System Using Multilevel Simultaneous Citation Networks," *Decision Support Systems,* Vol. 105, Pp. 24-33, 2018.

[23] J. Beel, S. Langer, M. Genzmehr, And A. Nürnberger, "Introducing Docear's Research Paper Recommender System," In *Proceedings of the 13th Acm/Ieee-Cs Joint Conference on Digital Libraries*, 2013, Pp. 459-460.

[24] A. Marchand And P. Marx, "Automated Product Recommendations with Preference-Based Explanations," *Journal of Retailing,* Vol. 96, Pp. 328-343, 2020.

[25] L. Steinert, "Beyond Similarity and Accuracy: A New Take on Automating Scientific Paper Recommendations," Phd, University of Duisburg-Essen, 2017.

[26] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, And T. Jambor, "Auralist: Introducing Serendipity into Music Recommendation," In *Proceedings of The Fifth Acm International Conference on Web Search and Data Mining*, 2012, Pp. 13-22.

[27] A. H. Afridi, "Transparency for Beyond-Accuracy Experiences: A Novel User Interface for Recommender Systems," *Procedia Computer Science,* Vol. 151, Pp. 335-344, 2019.

[28] S. M. Mcnee, J. Riedl, And J. A. Konstan, "Being Accurate Is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems," In *Chi'06 Extended Abstracts on Human Factors in Computing Systems*, 2006, Pp. 1097-1101.

[29] A. Koestler, "The Act of Creation," 1964.

[30] M. R. Berthold, "Towards Bisociative Knowledge Discovery," In *Bisociative Knowledge Discovery*, R. B. Michael, Ed., Ed: Springer-Verlag, 2012, Pp. 1-10.

[31] C. Pan and W. Li, "Research Paper Recommendation with Topic Analysis," In *Computer Design and Applications (Iccda), 2010 International Conference On*, 2010, Pp. V4-264-V4-268.

[32] M. Amami, G. Pasi, F. Stella, And R. Faiz, "An Lda-Based Approach to Scientific Paper Recommendation," In *International Conference on Applications of Natural Language to Information Systems*, 2016, Pp. 200-210.

[33] F. Ahmed and M. Fuge, "Creative Exploration Using Topic Based Bisociative Networks," *Arxiv Preprint Arxiv:1801.10084,* 2018.

[34] S. Mednick, "The Associative Basis of The Creative Process," *Psychological Review,* Vol. 69, P. 220, 1962.

[35] A. E. Fultz and K. M. Hmieleski, "The Art of Discovering and Exploiting Unexpected Opportunities: The Roles Of Organizational Improvisation And Serendipity In New Venture Performance," *Journal Of Business Venturing,* Vol. 36, P. 106121, 2021.

[36] C.-L. Wu, "Discriminating the Measurement Attributes of The Three Versions of Chinese Remote Associates Test," *Thinking Skills and Creativity,* Vol. 33, P. 100586, 2019.

[37] D. Kotkov, S. Wang, And J. Veijalainen, "A Survey of Serendipity in Recommender Systems," *Knowledge-Based Systems,* Vol. 111, Pp. 180-192, 2016.

[38] G. M. Lunardi, G. M. Machado, V. Maran, And J. P. M. De Oliveira, "A Metric for Filter Bubble Measurement in Recommender Algorithms Considering the News Domain," *Applied Soft Computing,* Vol. 97, P. 106771, 2020.

[39] B. Rawat, J. K. Samriya, N. Pandey, And S. C. Wariyal, "A Comprehensive Study on Recommendation Systems Their Issues and Future Research Direction In E-Learning Domain," *Materials Today: Proceedings,* 2020/12/01/ 2020.

[40] I. C. Paraschiv, M. Dascalu, P. Dessus, S. Trausan-Matu, And D. S. Mcnamara, "A Paper Recommendation System with Readerbench: The Graphical Visualization of Semantically Related Papers and Concepts," In *State-Of-The-Art and Future Directions of Smart Learning*, Ed: Springer, 2016, Pp. 445-451.

[41] S. L. Tomassen, "Research on Ontology-Driven Information Retrieval," In *Otm Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2006, Pp. 1460-1468.

[42] T. Kötter, K. Thiel, And M. R. Berthold, *Domain Bridging Associations Support Creativity*, 2010.

[43] B. M. Maake, S. O. Ojo, And T. Zuva, "Towards A Serendipitous Research Paper Recommender System Using Bisociative Information Networks (Bisonets)," In *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (Icabcd)*, 2018, Pp. 1-6.

[44] D. R. Swanson, "Migraine and Magnesium: Eleven Neglected Connections," *Perspectives in Biology and Medicine,* Vol. 31, Pp. 526-557, 1988.

[45] M. Juršič, B. Sluban, B. Cestnik, M. Grčar, And N. Lavrač, "Bridging Concept Identification for Constructing Information Networks from Text Documents," In *Bisociative Knowledge Discovery*, Ed: Springer, 2012, Pp. 66-90.

[46] D. M. Blei, A. Y. Ng, And M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* Vol. 3, Pp. 993-1022, 2003.

[47] B. Sluban, M. Juršič, B. Cestnik, And N. Lavrač, "Exploring the Power of Outliers for Cross-Domain Literature Mining," In *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools,*

And Applications, M. R. Berthold, Ed., Ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, Pp. 325-337.
[48] D. Kim, D. Seo, S. Cho, And P. Kang, "Multi-Co-Training for Document Classification Using Various Document Representations: Tf–Idf, Lda, And Doc2vec," *Information Sciences,* Vol. 477, Pp. 15-29, 2019/03/01/ 2019.
[49] K. Hornik and B. Grün, "Topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software,* Vol. 40, Pp. 1-30, 2011.
[50] Y. Chen, H. Zhang, R. Liu, Z. Ye, And J. Lin, "Experimental Explorations on Short Text Topic Mining Between Lda and Nmf Based Schemes," *Knowledge-Based Systems,* Vol. 163, Pp. 1-13, 2019.
[51] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proceedings of The National Academy of Sciences,* Vol. 101, Pp. 5228-5235, 2004.
[52] Y. Zhang, J. Callan, And T. Minka, "Novelty and Redundancy Detection in Adaptive Filtering," In *Proceedings of the 25th Annual International Acm Sigir Conference on Research and Development In Information Retrieval*, 2002, Pp. 81-88.