Journal of Language, Technology & Entrepreneurship in Africa (JOLTE)
Vol. 12 No. 2 2021

A Bisociated Domain-Based Serendipitous Novelty-Recommendation Technique for Recommender Systems

Benard Magara Maake
bmaake@kisiiuniversity.ac.ke

Fredrick Mzee Awuor
fawuor@kisiiuniversity.ac.ke

School of Information Science and Technology
Kisii University, Kenya.

Abstract
Traditional recommendation paradigms such as content-based filtering (CBF) tend to recommend items that are very similar to user profile characteristics and item input, resulting in the classical twin problem of overspecialization and concentration bias of recommendations. This twin problem is prevalent with CBF recommender systems due to the utilisation of accuracy metrics to retrieve similar items, and, limiting recommendation computations to single recognized user-centered domains, rather than cross-domains. This paper proposes a Bisociated domain-based serendipitous novelty recommendation techniques using Bisolinkers exploratory creativity discovery technique. The use of Bisolinkers enables establishing unique links between two seemingly unrelated domains, to enhance recommendation accuracy and user satisfaction. The presence of similar terms in two habitually incompatible domains demonstrates that two seemingly unrelated domains contain elements that are related and may act as a link to connect these two domains.

Keywords: *recommender systems, novelty, machine learning, outlier detection, bisociation*

1. Introduction

We live in a society that is digitising all its operations owing to the advancement of computer systems. For over four decades, computers have been used in various processes, but it was not until the mid-1990s that a profound wave of digital revolution took place. Three important aspects transformed computer use: Introduction of user-friendly operating systems and interfaces; Internet and the World Wide Web rapidly spread to people; computer, software, telecommunications and entertainment companies that previously worked independently converged (Bhimani & Willcocks, 2014). Thus, the modus operandi of many organisations and institutions from then changed significantly.

Data from various ever-increasing Internet-based gadgets such as mobile phone applications, social media, et cetera, is being collected, stored and analysed to find relevant, applicable, and usable knowledge for organisations. In addition, the world is becoming a type of information system through the possibility of the 'Internet of Things' (IoT). In this system, everything is being linked and connected to everyone, triggering the 'Big Data' paradigm, which calls for new techniques of collection, analysing, and interpretation (Benard, Sunday, & Tranos, 2019) (Maake, Ojo, & Zuva, 2019) and (M. B. Magara, Ojo, & Zuva, 2017). The process of extracting knowledge and information from such 'Big Data' has become a

major research topic which requires for its analysis methods from statistics, computer science, databases, machine learning, inter alia (Bhimani & Willcocks, 2014; Feldman & Sanger, 2007).

Mass availability of information leaves users sifting through millions of available choices (Dong, Tokarchuk, & Ma, 2009). This situation becomes a very expensive endeavour, which makes it even more difficult to make better- and well-informed choices. Therefore, systems are needed which reduce user search efforts, while providing mutual benefits to both users and organisations deploying these systems (Kim, Ghiasi, Spear, Laskowski, & Li, 2017).

Recommender systems are tools being developed to navigate these complex information spaces, to facilitate efficiency, productivity, and health of all their users. They form part of information-filtering systems that eliminate unwanted information, while automatically presenting to users useful and relevant data (Sridharan, 2014). These systems recommend music (for example, Spotify & Pandora), entertainment (for example, Netflix & YouTube), travel and leisure services (for example, TripAdvisor), consumer products (for example, Amazon & eBay), and more (Kim et al., 2017; Maake Benard Magara, Ojo, Ngwira, & Zuva, 2016). The following advantages have been realised: personalised recommendations, relevant content recommended, increase in business revenue, discovery of new items, increase in diversity in the category of items recommended, useful and effective recommendations, quality purchases, and so on (Kim et al., 2017; Ricci, Rokach, & Shapira, 2015).

Recommender systems managing academic literature utilise features such as keywords, citations, and text content, in determining the most apposite content to recommend to users. This category of recommender systems is known as content-based filtering (CBF) recommender systems (Adamopoulos & Tuzhilin, 2014). Unfortunately, content-based filtering research-paper recommender systems keep suggesting research-paper articles that are similar to what the target user has indicated as interesting, leading to boring, obvious, and uninteresting recommendations − a twin problem known as overspecialization and concentration bias of recommendations (Adamopoulos & Tuzhilin, 2014). This primarily happens because most recommender systems are evaluated centered on accuracy metrics, which do not relate well to user desires such as novelty, relevance, and unexpectedness (Kotkov, Wang, & Veijalainen, 2016). The effect of the twin problem of overspecialization and concentration bias is a poor, inconsistent user experience having limited focus on user goals and knowledge discovery. In order to alleviate this twin problem and improve CBF recommender systems, we propose a technique that produces recommendations by establishing links from domains seemingly unrelated from one another.

2. Bisociation

According to Koestler (1964), two concepts are said to be bisociated if and only if there is no direct obvious evidence linking them. Besides, and one concept has to cross contexts to find the link and the new link provides some novel insights (Koestler, 1964). Association is the discovery of patterns inside a single domain, whereas bisociation is the discovery of patterns across two or more unrelated domains (Kötter, Thiel, & Berthold, 2010).

Koestler further explains the meaning of bisociation as linking unrelated, often incompatible, information in a new innovative way (Ahmed & Fuge, 2018). Bisociative knowledge discovery techniques can be used to investigate the presence of latent relationships between seemingly unrelated domains. It is highly probable that if these techniques are integrated with recommender systems, then new, interesting and perhaps serendipitous items would likely be suggested between these disconnected domains. For these reasons, this research was aimed at modelling a research-paper recommender system, which recommends novel items from two large, typically known incompatible information spaces using exploratory creativity discovery (Dubitzky, Kötter, Schmidt, & Berthold, 2012; Maake & Tranos, 2019; Maake Benard Magara,

Ojo, & Zuva, 2018). Link between two seemingly unrelated domains is established using a combination of topic modelling, bisociation and serendipity, as we proposed in (Maake & Tranos, 2019; Maake Benard Magara et al., 2018).

3. Novelty in Recommender Systems

A novel item is one that is previously not known to a user (Kaminskas & Bridge, 2017), and non-redundant without allowing known but unexpected items, or a fraction of relevant documents that are unknown to a user (Baeza-Yates & Ribeiro-Neto, 1999). In recommender systems, novelty is the ability of a recommender system to introduce to users items that have not been previously experienced, these items coming from outside the user system (Y. C. Zhang, Séaghdha, Quercia, & Jambor, 2012). Novelty has been defined as the difference between the present and past experience (Castells, Hurley, & Vargas, 2015). Evidently, many definitions of novelty in recommender systems exist; however, this research proposes a modified definition that places bisociation in perspective:

"Novelty is receiving fortuitous and surprisingly relevant items from domains that are considered to be completely unrelated, but due to bisociatedness between these domains, items recommended interestingly turn out to be valuable".

Let $M1$ and $M2$ represent two habitually incompatible domains. Let C_V represent a user in domain $M2$ who has "viewed" items or concepts in this domain. Let the viewed concepts be represented as $v_{c1}, v_{c2}, v_{c3}, \dots, v_{cn}$ as portrayed in Figure 1. Hence, we can say that all the view concepts are contained in C_V as expressed below:

$C_V \ni v_{c1}, v_{c2}, v_{c3}, \dots, v_{cn}$ which can be summarised as:

$$C_V = v_{c1}, \dots, v_{cn} \tag{1}$$

The total viewed concepts by user C_V can also be represented as the sum of all viewed concepts:

$$C_V = \sum_{i=1}^{n} v_{ci} \tag{2}$$

All distinct viewed concepts have distances from the user. Let all distances between the user C_V and all viewed concepts in domain $M2$ such that $\{C_V;\ v_{c1}, v_{c2}, v_{c3}, \dots, v_{cn}\} \in M2$ be represented as $d_{v1}, d_{v2}, d_{v3}, \dots, d_{vk}$, respectively. This can be further presented as:

$$= d_{v1}, \dots, d_{vk} = \sum_{i=1}^{k} d_{vi} \tag{3}$$

Let I_r represent an item recommendation from domain $M1$ that is supposed to be tested for novelty. Let the distance between I_r and C_V be represented as $\delta_{v,r}$.

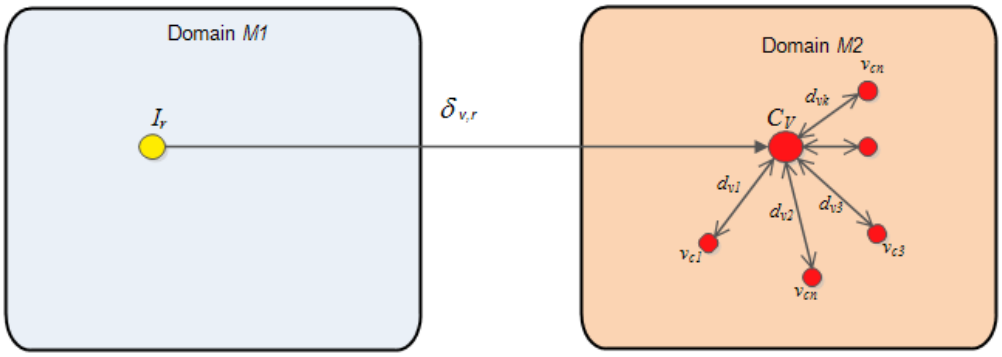


Figure 1: Novelty in Bisociative Research Paper Recommender System

$$\delta_{v,r} = \text{dist}(C_V; I_r) \tag{4}$$

For a user to identify an item as novel, the item must be received fortuitously (accidently/ unexpectedly), surprisingly (unpredictably), and it must be relevant (important/valuable). Hence, novelty can be described as the distance between a recommended item I_r and the previously accessed or viewed items, C_V:

$$novelty = \text{dist}(C_V;\ I_r) \tag{5}$$

where dist ( ) is any similarity distance measure that can be used to measure the differences between any two or more items. From Figure 1 above, a novel recommendation is not equal to or similar (distance measure) to the concept C_V, and it has to be one that originates from another domain $M1$ (a recommendation from a domain that is usually considered as mismatched). It should also be relevant to the user of the system. If we describe novelty as λ, we will then have the following novelty expression:

$$\lambda = \begin{cases} 1, & \text{if } (I_r \in M1) \wedge (\delta_{v,r} > d_{v1}, \dots, d_{vk}) \\ 0, & \text{Otherwise} \end{cases} \tag{6}$$

A. Method

Algorithm 5.2: Novelty in bisociated domains

Input:

- Concepts (items) from two bisociated domains $M1$ and $M2$
- Viewed concepts represented as a user in domain $M2$, $\{C_V = v_{c1}, v_{c2}, \dots, v_{cn}\}$
- Recommended item(s) I_r from domain $M1$
- Set a given threshold α
- $\delta_{v,r}$ is the distance of the recommended item I_r from the user C_V

Output:

- Novelty (λ)

1: **load** concepts from bisociated domains $\{I_r \leftarrow M1; C_V \leftarrow M2\}$
2: **for** $C_V \in M2$ **do**
3: $d_{v1}, \dots, d_{vk} \leftarrow$ `findDistance` $(C_V; v_{c1}, \dots, v_{cn})$
4: `array_distances` [ ] $= d_{v1}, \dots, d_{vk}$;
5: **end for**
6: **if** $I_r \in M1$ **then**
7: $\delta_{v,r}$ = `findDistance` $(C_V;\ I_r)$;

8: **else**
9: **exit;**
10: **if** $\delta_{v,r} >$ `array_distances[ ]` && $\delta_{v,r} > \alpha$ **then**
11: $\lambda = 1$;
12: **else**
13 $\lambda = 0$;
14: **exit;**
15: **return** ($\boldsymbol{\lambda}$)

B. Pre-processing

The process of detecting novelty in textual documents involved three main steps:

- A pre-processing step that is used to clean the data by removing all the stop words (since stop words influence novelty prediction), stem all words and perform Parts-Of-Speech tagging.
- The next step is categorization: documents are classified whether or not they are relevant.
- The final step is novelty mining: documents are quantitatively measured using novelty metrics such as the cosine similarity measure (Maake et al., 2019). Figure 2.

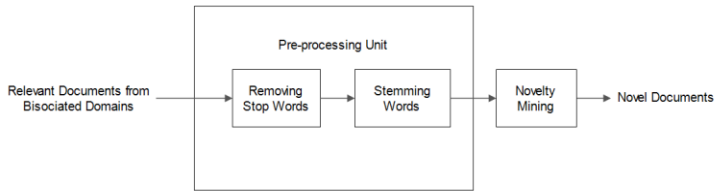


Figure 2: Steps in implementing novelty in bisociated domains.

C. Pre-processing Techniques

Once relevant documents from bisociated domains had been fed into the pre-processing channel, changes on the textual composition of our documents took place. Two main pre-processing techniques were employed in this section: stop word removal and stemming. The output of this section was a bag of words (BOW) to make a term-document matrix (TDM) (a term-sentence matrix (TSM) can also be created). The vector space previously constructed predicted whether incoming documents were novel, using the cosine similarity measure. The metric calculated the similarity of current documents d_t, and all its history documents (documents that were in the same domains, in our case, M_2), $d_i (1 \leq i \leq t-1)$. It then calculated the novelty of documents by subtracting the maximum of these cosine similarities from one (1), providing the following metric:

$$\text{Novelty Score } (d_t) = 1 - \max_{1 \leq i \leq t-1} \cos(d_t, d_i) \tag{6}$$

where the cosine similarity took the following form:

$$\cos(d_t, d_i) = \frac{\sum_{k=1}^{n} w_k(d_t).w_k(d_i)}{\|d_t\|.\|d_i\|}$$

where w is the weight of the k^{th} elements

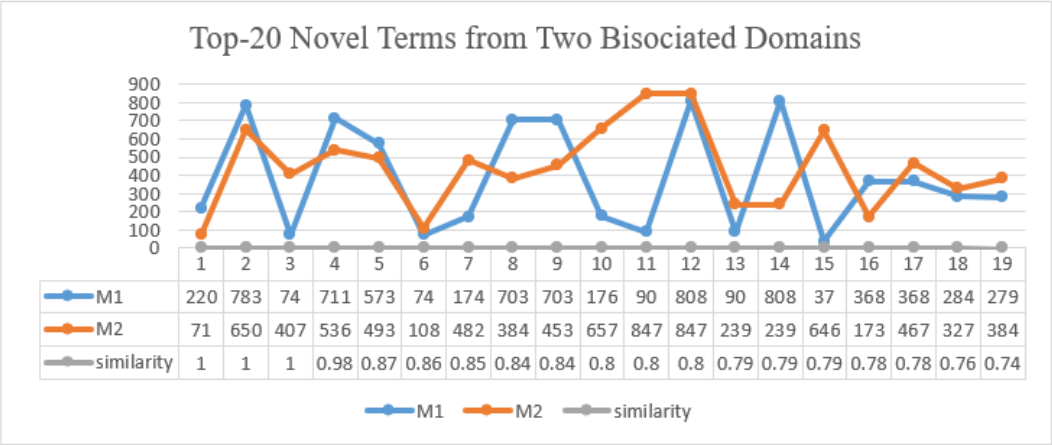
Finally, a document was declared novel or not depending on whether the novelty score obtained from incoming documents fell above or below a certain preferred threshold (in this case, a similarity of 0.7 was considered). (Y. Zhang, Callan, & Minka, 2002) indicated that irrelevant recommendations had the possibility of being new to a user yet lacking in novel attributes. Therefore, this research only considers relevant items for its experiments.

4. Experimental Results.

Data sets were collected from the two separate domains of magnesium and migraine and loaded into our programming environment. The first step was to pre-process, removing all punctuation marks, numbers, English stop words, and white spaces, and to lower all the terms into lowercase. We obtained a bag of words with which to construct two document term matrixes (DTMs), one for migraine, and another for magnesium. All the unnecessary terms were removed from the DTMs, while the remaining terms were weighted using the TF-IDF. To calculate the similarity measure, we applied the cosine similarity measure, owing to its efficiency and symmetry. For novelty mining, the cosine similarity measure TF-IDF, and a novelty minimum threshold, were selected for both the domains. All the values that were less than the threshold (that is, 0.5) were deleted and not utilised in the experiment. Finally, we compared the two DTMs for novelty, and the top-20 most similar terms between the two domains were established. Distance measures taken after the experiment were given in descending order. Table 1 displays how similar two terms were in two unrelated domains. The first column was a unique identifier provided after the similarity value was determined, while the column marked $M1$ and $M2$ are the two DTMs representing the unrelated domains of migraine and magnesium, respectively. The last column represents the similarity value in descending order arising from the two terms in domain $M1$ and $M2$. Figure 3 displays the first 19 terms from both domains that were novel (similar to) for one another.

Table 1: Top-20 Similar Terms between Domain – M1 and Domain – M2

Unique Identifier	$M1$	$M2$	Similarity
2335	220	71	0.9988660
18461	783	650	0.9987659
11565	74	407	0.9981646
15925	711	536	0.9760487
14522	573	493	0.8749978
3374	74	108	0.8567798
13705	174	482	0.8519813
10964	703	384	0.8364352
12727	703	453	0.8364352
18509	176	657	0.7995800
24337	90	847	0.7961725
24343	808	847	0.7961725
7301	90	239	0.7892390
7330	808	239	0.7892390
18324	37	646	0.7888917
5116	368	173	0.7841055
13120	368	467	0.7841055
9539	284	327	0.7551985



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
M1	220	783	74	711	573	74	174	703	703	176	90	808	90	808	37	368	368	284	279
M2	71	650	407	536	493	108	482	384	453	657	847	847	239	239	646	173	467	327	384
similarity	1	1	1	0.98	0.87	0.86	0.85	0.84	0.84	0.8	0.8	0.8	0.79	0.79	0.79	0.78	0.78	0.76	0.74

Figure 3: Top-20 novel terms from bisociated domains

5. Conclusion

The work presented in this paper is designed to support recommender systems in the field of research-paper recommendations, moving towards identifying, creating, and recommending serendipitous new-knowledge by means of moving beyond the normal TF-IDF and accuracy models to other new metrics that capture other user impression qualities. The study further supports moving beyond single-domain recommendations to cross-domain knowledge discovery and recommendations through appropriate novelty and beyond-similarity techniques. Having realized that seemingly distant domains (bisociation) may contain items that are similar as shown in Figure 3, it is recommended that further research is conducted when trying to establish the relationships between two or more unrelated domains, and more especially in the field of knowledge discovery.

REFERENCES:

Adamopoulos, P., & Tuzhilin, A. (2014). *On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems.* Paper presented at the Proceedings of the 8th ACM Conference on Recommender systems.

Ahmed, F., & Fuge, M. (2018). Creative Exploration Using Topic Based Bisociative Networks. *arXiv preprint arXiv:1801.10084*.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463): ACM press New York.

Benard, M. M., Sunday, O. O., & Tranos, Z. (2019). Information Processing in Research Paper Recommender System Classes. In B. Raj Kumar & B. Paul (Eds.), *Research Data Access and Management in Modern Libraries* (pp. 90-118). Hershey, PA, USA: IGI Global.

Bhimani, A., & Willcocks, L. (2014). Digitisation,'Big Data'and the transformation of accounting information. *Accounting and Business Research, 44*(4), 469-490.

Castells, P., Hurley, N. J., & Vargas, S. (2015). Novelty and diversity in recommender systems. In *Recommender systems handbook* (pp. 881-918): Springer.

Dong, R., Tokarchuk, L., & Ma, A. (2009). *Digging friendship: paper recommendation in social network.* Paper presented at the Proceedings of Networking & Electronic Commerce Research Conference (NAEC 2009).

Dubitzky, W., Kötter, T., Schmidt, O., & Berthold, M. R. (2012). Towards creative information exploration based on Koestler's concept of bisociation. In *Bisociative Knowledge Discovery* (pp. 11-32): Springer.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*: Cambridge university press.

Kaminskas, M., & Bridge, D. (2017). Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS), 7*(1), 2.

Kim, H. M., Ghiasi, B., Spear, M., Laskowski, M., & Li, J. (2017). Online serendipity: The case for curated recommender systems. *Business Horizons, 60*(5), 613-620.

Koestler, A. (1964). The act of creation. London: Arkana. In: Penguin Books.

Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems, 111*, 180-192.

Kötter, T., Thiel, K., & Berthold, M. R. (2010). *Domain bridging associations support creativity*.

Maake, B. M., Ojo, S. O., & Zuva, T. (2019). A Survey on Data Mining Techniques in Research Paper Recommender Systems. In *Research Data Access and Management in Modern Libraries* (pp. 119-143): IGI Global.

Maake, B. M., & Tranos, Z. (2019). A SERENDIPITOUS RESEARCH PAPER RECOMMENDER SYSTEM. *International Journal of Business and Management Studies, 11*(1), 39-53.

Magara, M. B., Ojo, S., Ngwira, S., & Zuva, T. (2016). *MPlist: Context aware music playlist.* Paper presented at the 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech).

Magara, M. B., Ojo, S., & Zuva, T. (2017, 4-7 Dec. 2017). *Toward Altmetric-Driven Research-Paper Recommender System Framework.* Paper presented at the 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS).

Magara, M. B., Ojo, S. O., & Zuva, T. (2018). *Towards a Serendipitous Research Paper Recommender System Using Bisociative Information Networks (BisoNets).* Paper presented at the 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD).

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. *Recommender systems handbook, 54*, 1-34.

Sridharan, S. (2014). Introducing serendipity in recommender systems through collaborative methods.

Zhang, Y., Callan, J., & Minka, T. (2002). *Novelty and redundancy detection in adaptive filtering.* Paper presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

Zhang, Y. C., Séaghdha, D. Ó., Quercia, D., & Jambor, T. (2012). *Auralist: introducing serendipity into music recommendation.* Paper presented at the Proceedings of the fifth ACM international conference on Web search and data mining.