# Automatic Detection of Cross-Disciplinary Knowledge Associations

**Menasha Thilakaratne, Katrina Falkner, Thushari Atapattu**
School of Computer Science
University of Adelaide
Adelaide, Australia
{firstname.lastname}@adelaide.edu.au

## Abstract

Detecting interesting, cross-disciplinary knowledge associations hidden in scientific publications can greatly assist scientists to formulate and validate scientifically sensible novel research hypotheses. This will also introduce new areas of research that can be successfully linked with their research discipline. Currently, this process is mostly performed manually by exploring the scientific publications, requiring a substantial amount of time and effort. Due to the exponential growth of scientific literature, it has become almost impossible for a scientist to keep track of all research advances. As a result, scientists tend to deal with fragments of the literature according to their specialisation. Consequently, important and hidden associations among these fragmented knowledge that can be linked to produce significant scientific discoveries remain unnoticed. This doctoral work aims to develop a novel knowledge discovery approach that suggests most promising research pathways by analysing the existing scientific literature.

## 1 Problem Statement

Formulation of scientifically sensible novel research hypotheses requires a comprehensive analysis of the existing literature. However, the voluminous nature of literature (Cheadle et al., 2016) makes the hypothesis generation process extremely difficult and time-consuming even in the narrow specialisation of a scientist. In this regard, *Literature-Based Discovery (LBD)* research is highly beneficial as it aims to detect non-trivial implicit associations by analysing a massive number of documents that have the potential to generate novel research hypotheses (Ganiz et al., 2005). Moreover, LBD outcomes encourage the progress of cross-disciplinary research by suggesting promising cross domain research pathways, which are typically unnoticed during manual analysis (Sebastian et al., 2017b).

Independent of the domain, LBD is highly valuable to accelerate knowledge acquisition and research development process. However, the existing LBD approaches are mostly limited to *medical domain* that attempt to find associations among *genes*, *proteins*, *drugs*, and *diseases*. The main reason for this can be the highly specific and descriptive nature of medical literature that is suitable for LBD research (Ittipanuvat et al., 2014). The application of LBD process in domains such as *Computer Science (CS)* is challenging due to the rapidly evolving nature of terms in the content of research publications. The medical related LBD approaches are strongly coupled with the medical domain knowledge by utilising resources such as Unified Medical Language System (UMLS), MetaMap, and Medical Subject Headings (MeSH) descriptors (Sebastian et al., 2017a). This makes the applicability of these approaches to other domains challenging.

LBD research outside of medical domain is still in an immature state. There are only a few LBD studies performed outside of medical domain (e.g., *Water Purification* (Kostoff et al., 2008), *Technology & Social Issues* (Ittipanuvat et al., 2014), *Humanities* (Cory, 1997)). To date, a work by Gordon and Lindsay (2002) is the only available CS-related LBD research. Hence, in this doctoral research, we attempt to contribute to LBD discipline outside of medical domain by automating cross-disciplinary knowledge discovery process. As a proof of concept, the proposed solution will be applied to different CS-related concepts.

*Proceedings of ACL 2018, Student Research Workshop*, pages 45–51
Melbourne, Australia, July 15 - 20, 2018.

## 2 LBD Discovery Models

Most of the LBD literature are based on the fundamental premise introduced by Swanson namely, *ABC Model* (Swanson, 1986). It employs a simple syllogism to identify the potential knowledge associations. i.e. given two concepts A and C in two disjoint scientific literature, if A is associated with concept B, and the same concept B is associated with C, the model deduces that A is associated with C. Swanson demonstrated how these combined knowledge pairs contribute to reach solutions by manually making several medical discoveries (e.g., *Raynaud's disease* ↔ *Fish Oil* (Swanson, 1986) and *Migraine* ↔ *Magnesium* (Swanson, 1988)). These medical discoveries are the basis of the LBD discipline.

The ABC model has two variants named as *open* and *closed* discovery models (Figure 1) (Henry and McInnes, 2017). Open discovery starts with an initial user-defined concept (e.g., *Learning Analytics (LA)*) where the LBD process automatically analyses the literature related to the initial concept to detect the potential interesting and implicit associations. This model is generally used when there is a single problem (A-concept) with limited knowledge on what concepts can be involved (B and C concepts). This model greatly assists in *hypothesis generation* process. On the contrary, closed discovery process requires two user-defined concepts (e.g., *LA* and *Deep Learning*) as the input to output potential hidden associations (B-concepts) between these specified two concepts A and C. This model is generally used for *hypotheses testing and validation*. However, the derived associations of the model can also be considered to generate more granular hypotheses. The granularity of the user-defined concepts of the two discovery models can vary depending on the researcher's interest (Ganiz et al., 2005).

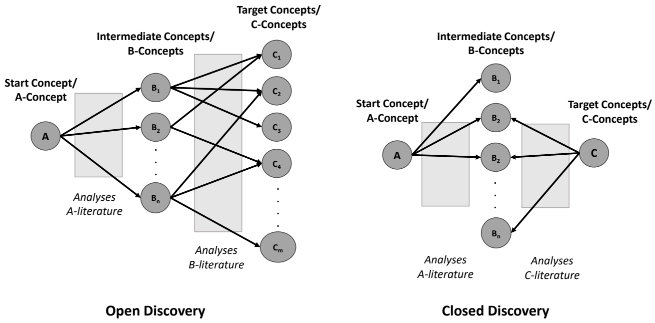

Figure 1: Open and closed discovery models.

## 3 Related Work

Even though the early work in LBD was performed manually, over the time, different computational techniques were adopted to automate the LBD process. Most of the existing LBD research are semi-automated and requires a human expert to make decisions during the LBD process (Sebastian et al., 2017a).

Much of the early computational approaches utilise lexical statistics (Swanson and Smalheiser, 1997) such as term frequency-inverse document frequency, token frequencies, which can be considered as the most primitive LBD approach. Later, Distributional Semantics approaches (Gordon and Dumais, 1998) such as Latent Semantic Indexing, Reflective Random Indexing were introduced. Subsequently, knowledge-based approaches (Weeber et al., 2001) were adopted in the LBD process that heavily rely on the existence of external structured knowledge-based resources to acquire domain-specific knowledge.

Relations-based approaches (Hristovski et al., 2006) make use of user-defined explicit predicates to convey the meaning of the associations. However, these approaches are restricted to problems where semantic types and predicates are known in advance. Another category is Graph-based approaches (Cameron et al., 2015) that generate a number of bridging terms to define the associations. Bibliometrics-based approaches (Kostoff, 2014) utilise bibliographic link structures in the LBD process to identify potential knowledge associations. Several attempts have been taken in LBD literature to employ link prediction techniques (Sebastian et al., 2015). i.e. attributes of the concepts and observed links are used to predict the existence of new links between the concepts.

As discussed earlier, majority of the LBD research are in medical domain and dependent on medical domain knowledge. As a result, it is not feasible to apply these approaches to other domains. To date, there are only a handful of LBD research studies performed outside of the medical domain. This points out the importance of contributing to non-medical LBD research which is still in an early stage.

## 4 Goals and Research Questions

Development of an automatic LBD system can significantly improve the typical research process

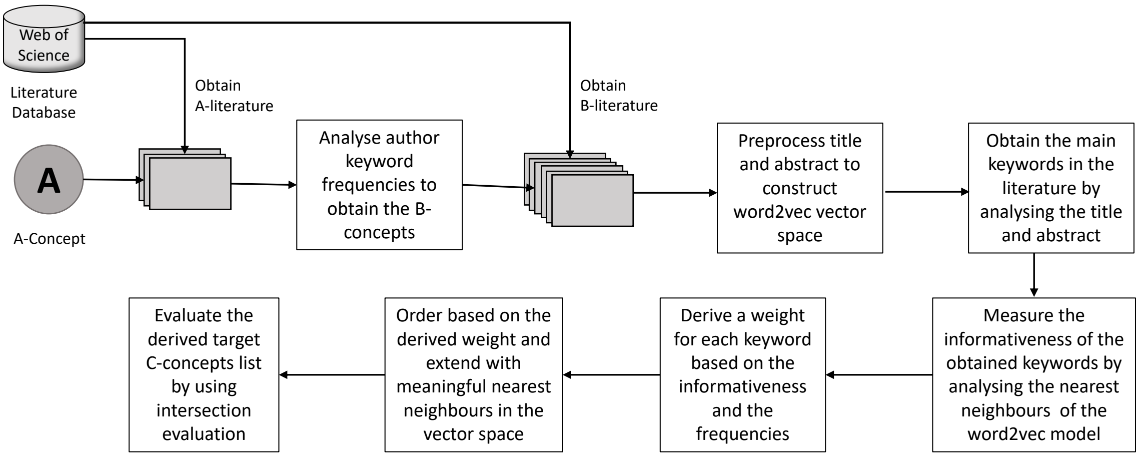
Web of Science
Literature Database
A
A-Concept
Obtain A-literature
Analyse author keyword frequencies to obtain the B-concepts
Obtain B-literature
Preprocess title and abstract to construct word2vec vector space
Obtain the main keywords in the literature by analysing the title and abstract
Measure the informativeness of the obtained keywords by analysing the nearest neighbours of the word2vec model
Derive a weight for each keyword based on the informativeness and the frequencies
Order based on the derived weight and extend with meaningful nearest neighbours in the vector space
Evaluate the derived target C-concepts list by using intersection evaluation

Figure 2: High-level overview of the proposed open discovery model.

followed by the scientists. With such system, scientists can generate scientifically sensible research hypotheses in a shorter time by considering the suggestions provided by the system. Moreover, the cross-domain knowledge discovery process of LBD facilitates the development of cross disciplinary research. Thus, in this study we are aiming to develop a novel knowledge discovery approach by utilising both the variants of LBD process namely, *open* and *closed* discovery.

Our main intention is to uplift the LBD process in non-medical domains. The ultimate goals of this doctoral research are; 1) Fully automate the LBD process: Most of the existing LBD approaches are semi-automatic and require human decisions to direct the knowledge discovery process at various stages. Thus, automating the entire LBD process will be highly beneficial for the users of the LBD model. 2) Provide a generic LBD solution that is independent of domain specific knowledge: Most of the existing LBD approaches rely on domain-specific knowledge to identify the knowledge associations. As a result, the applicability of these approaches to other domains are limited. Therefore, it is important to generate a LBD model that is suitable for any domain, without incorporating any domain specific knowledge.

While focusing on technical literature, in particular, CS domain, the research questions of this study are; 1) How to leverage NLP and Machine Learning techniques to enhance the understanding of content in the literature to accurately detect research areas with different levels of granularity? 2) How bibliometrics analysis can be integrated to enhance identification of implicit knowledge associations? 3) What are the scoring schemes that can be used to rank the identified associations? 4) How to improve the existing evaluation approaches to accurately validate the LBD outcomes? This doctoral work plans to answer these research questions to fulfill the aforementioned goals.

## 5 Current Work and Future Directions

We have conducted a comprehensive literature analysis and have defined our research questions based on the gaps identified in the literature. Currently, we are carrying out preliminary studies to identify potential techniques that would be useful to enhance the predictability of the open discovery LBD model. Figure 2 depicts the high-level overview of our methodology that addresses research questions 1 and 3. The proposed LBD approach is based on Swanson's ABC discovery model. In this approach we are attempting to identify the importance of *neural word embeddings* (Mikolov et al., 2013b) to accurately capture the context of the main keywords of the abstracts. As for the literature database, we are using *Web of Science Core Collection (WoS)*[1] to obtain the meta data of the literature such as title, abstract, and author keywords.

As shown in Figure 2, initially the frequencies of the author keywords were analysed to obtain the B-literature. The retrieved title and abstract in B-literature were cleaned using several carefully picked preprocessing techniques. In summary, the potential abbreviations in the text are identified by using multiple regex patterns. Afterwards, variable length n-grams in the text were

[1] https://clarivate.com/products/web-of-science/

identified by using a formula based on collocation patterns described in Mikolov et al. (2013a). We also removed numbers, punctuation marks and terms constituent of single letters before analysing the texts.

After the preprocessing phase, we identified the sentence in the abstract that describes the intention/purpose of the study by using multiple intention-based word patterns. We further processed the identified purpose sentence along with the title by removing stop words. The intention of using the purpose sentence and title is that they typically include the most important concepts that best describe the study.

For each post-processed n-grams ($w_i$) of the purpose sentence and title, we calculated a semantic importance (*informativeness*) score based on *word2vec* (Mikolov et al., 2013b) word embedding method. In other words, we measured the informativeness of $w_i$ based on the validity and semantic richness of *N* neighbouring terms derived using *cosine similarity*. To measure the validity and semantic richness of *N* neighbouring terms, we imposed the following three criterions for the three categories of the neighbouring terms; unigrams, abbreviations and n-grams respectively. 1) valid technical unigram 2) valid detected abbreviation 3) valid n-gram by eliminating partial n-grams with different Part of Speech (POS) tag patterns. If the neighbouring term ($n_i$) fulfills the relevant criteria based on its category, it will be considered as a *valid, quality neighbouring term* ($n_i \in \mathcal{N}$ & $\mathcal{N} \subset N$). We excluded $w_i$ if its informativeness is less than or equal to 50%. i.e. the excluded terms have majority of neighbours that does not fulfill the valid, quality neighbouring criterions.

$$informativeness(w_i) = \frac{1}{N}\sum_{i=1}^{N}[n_i \in \mathcal{N}]$$

The frequency of $w_i$ denotes the importance of the term within B-literature. Therefore, we multiplied informativeness($w_i$) by the number of occurrences $w_i$ appeared in the title and purpose sentences to obtain the final weighted score. This score was used to rank the derived $w_i$ terms which represent the important concepts in B-literature (i.e. *seed concepts*). Each seed concept was extended by linking valid quality neighbouring terms (using the same criteria used to measure the validity and semantic richness of the neighbouring terms) in *word2vec* vector space.

We performed the same steps of the experiment with *fastText* (Bojanowski et al., 2016) word embedding method. An important observation is that with respect to *word2vec*, we obtained a broad topic coverage in the same field showing what areas are connected with the seed concept whereas with *fastText*, we obtained topics in a narrow range.

We used *Learning Analytics (LA)* as the A-concept to evaluate our approach. The reason for choosing LA is that it is a relatively novel but rapidly growing area connected to many disciplines like education, psychology, machine learning etc. We utilised *intersection evaluation* of Gordon and Lindsay's work (2002) to evaluate the C-concepts obtained for LA. As for the evaluation literature database, we used WoS and Scopus[2] to obtain the intersection frequencies. We categorised the derived C-concepts as *existing*, *emerging* and *novel* based on the intersection frequencies. In total we obtained 564 knowledge associations [3].

Unlike Gordon and Lindsay's work (2002), the knowledge discovery process of our approach is fully automated and does not require any human intervention to make decisions during the process. Therefore, verifying the validity of the obtained concepts is important to ensure that the associated intersections are meaningful. We performed an expert-based concept validation by utilising *legitimate concept criteria* discussed in Hurtado et al. (2016). From the evaluation performed by two LA researchers with CS background, we obtained an average accuracy of 97.87%, 98.21%, 95.68% for existing, emerging, and novel associations respectively. The experts' agreement for the valid terms is 93.6%. Through our experiment, we could successfully identify many interesting C-concepts that have the potential to generate scientifically sensible novel research hypothesis. For example, our *existing* C-concepts included well-established research areas in LA such as *machine learning*, *data mining*, and *e-learning* whereas the *emerging* C-concepts included infrequently used potential research areas in LA such as *computer vision*, *linked open data*, and *cognitive science*. We could also obtain many interesting *novel* C-concepts such as *word embedding techniques*, *deep learning architectures such as LSTM, BLSTM, CNN*, that can be utilised in future

[2] https://www.scopus.com/

[3] https://bit.ly/2rApl7C

Table 1: Methodology comparison

| | (Gordon and Lindsay, 2002) | Our Approach |
|---|---|---|
| Process | Requires human intervention | Fully automatic |
| Concepts | Bi-grams only | Uni-grams, abbreviations and variable length n-grams |
| Techniques | Lexical Statistics | Lexical Statistics & Distributional Semantics |
| Output | List of bi-grams | Detailed contextualised semantics groupings |

LA research. Therefore, the suggestions provided through our approach will greatly influence to uplift the process of research in LA. In comparison with the only existing CS-related LBD approach (Gordon and Lindsay, 2002), our approach utilises an improved methodology (Table 1).

To the best of our knowledge, this is the first non-medical LBD study that utilises neural word embeddings to detect the target C-concepts. Our initial results demonstrate the importance of exploiting neural word embeddings to effectively identify potential cross-disciplinary knowledge associations buried in literature. We would like to further enhance our existing approach by considering the below-mentioned future directions that are categorised based on our four research questions (RQ) described in Section 4.

**RQ 1 (Content Analysis):** A subtle analysis of literature is needed to accurately capture the hidden knowledge associations. In our current experiment, we are considering concepts at keywords-level by identifying seed concepts. As an improvement, we would like to have an organised topic structure with different levels of granularity. In order to achieve that, we are intending to utilise semantic web technologies and pre-existing topical categories (e.g., Dewey Decimal Classification) to enhance the understanding of the content. Moreover, the identified topic structure will also be useful to provide a clearly structured, logical output to the user than merely listing the identified associations. Due to the lack of LBD research that analyse the effects of topic modeling (Sebastian et al., 2017b), it is also important to study how topical information propagate among research publications to detect interesting, implicit knowledge associations. Another interesting future direction would be to utilise deep language understanding techniques to infer ontologies from the scientific literature automatically which can be utilised to identify more granular knowledge associations.

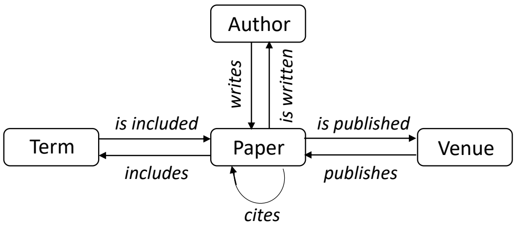

Figure 3: Entity & relation types (Shakibian and Charkari, 2017).

**RQ 2 (Bibliometrics Analysis):** In our current experiment, we are utilising the popular ABC model to discover the knowledge associations. However, the inference steps introduced through ABC model is simple and not foolproof. Therefore, in our future research studies, we are intending to analyse more complex inference steps to identify complex knowledge associations that cannot be identified through ABC model. To achieve that, we are aiming to integrate a graph-based approach by analysing the relationships among the four entity types (i.e. *author*, *term*, *paper*, *venue*) illustrated in Figure 3. In other words, we are intending to utilise different bibliographics-based link structures such as co-author relationships, direct citation links, co-word analysis, bibliographics coupling, and co-citation links to uncover complex knowledge associations. For example, when authors from disjoint research fields collaborate for a research, it implies a potential association between the two knowledge areas. This simple co-author relationship can be further expanded to more complex associations by analysing shared authors in the citations of the source and target literature, analysing authors in source literature that are cited by the target literature etc. Same as for the *author* entity, this procedure can be followed for the remaining entities (i.e. *paper*, *term*, *venue*) of the network schema in Figure 3 to derive more complex and implicit associations. With regards to *term* entity, the identified associations can be further expanded by leveraging topic modeling and topical categories. We are intending to automatically generate all the aforementioned associations (up to four degree meta path associations) for the four entity types by traversing through the network schema in Figure 3. From the derived associations we would like to identify the most effective association links by comparing the output results. The identified effective association links will provide an improved understanding of an implicit association than the simple ABC model.

**RQ 3 (Associations Ranking):** The generated target concepts list should be ordered in a way where the most significant knowledge association should be listed in the top. Therefore, it is important to identify the factors that can be utilised to rank the derived associations. In our current approach, we are incorporating semantic importance and frequency to rank the associations. In frequently evolving domains like CS, it is also important to consider the temporal factors to accurately identify the new research advancements. Therefore, we would like to propose different temporal-related weighting mechanisms to rank the target C-concepts. To accomplish this, we need to analyse the concepts in chronologically ordered time slices. For example, when deriving the temporal weight of a concept, factors such as the significance of the concept in its corresponding time interval, and changes of the concept's trend over the time using a sliding-window mechanism need to be considered. Moreover, we can also utilise pre-existing algorithms that analyse the evolution of topics (e.g., Dynamic Topic Model (Wang and McCallum, 2006), Topics over Time Model (Blei and Lafferty, 2006)) in this regard.

**RQ 4 (Evaluation):** Evaluating the validity of the identified knowledge associations outside of medical domain is very challenging due to the unavailability of gold standard datasets. Medical LBD literature mostly attempted to replicate Swanson's manually detected medical discoveries (e.g., Raynaud's Disease ↔ Fish Oil) to evaluate their results (Sebastian et al., 2017a). However, when dealing with other domains, the possible evaluation approaches that can be utilised are intersection evaluation (Gordon and Lindsay, 2002), time-sliced evaluation (Yetisgen-Yildiz and Pratt, 2009) and expert based evaluation (Gordon and Lindsay, 2002) that have number of inherent limitations in accurately validating the results. Therefore, we would like to improve the existing LBD evaluation approaches to accurately evaluate our results. In our current evaluation, we are using intersection evaluation along with expert-based concept validation. In our future experiments, we would like to quantitatively evaluate the knowledge associations by utilising information retrieval metrics such as *precision* and *recall* (to evaluate the complete set of target C-concepts) and *11-point average interpolated precision curves*, *precision at k*, and *mean average precision* (to evaluate the rankings of the target C-concepts). To quantitatively evaluate the overall quality of the results a ground truth is required. For this purpose, we are intending to create a time-sliced dataset described in the work of Yetisgen-Yildiz and Pratt (2009). In other words, the literature is divided into two sets namely, pre-cut-off set (includes literature before a cut-off date) and post-cut-off set (includes literature after the cut-off date). Afterwards, the LBD methodology is applied to pre-cut-off set to obtain the implicit knowledge associations. Later, the existence of these identified associations (that do not exist explicitly in pre-cut-off set) is checked in the post-cut-off set. A major limitation of this approach is that a knowledge association considered as a false positive can become a true positive once a new research is published. This limitation can be overcome upto some extent by incorporating human experts to further evaluate validity of the false positives. Another interesting avenue for evaluation would be *user performance evaluation* by incorporating users with diversified range of expertise such as users with limited prior knowledge and experts in the field (Qi and Ohsawa, 2016). Through this approach, we can evaluate the extent to which the proposed LBD approach assist different levels of users to generate hypotheses by utilising the suggested knowledge associations.

Thus, this doctoral work can be expanded in numerous ways since LBD outside of medical domain is still in an early stage. Our next phase is to address the above discussed four focus points. Moreover, in our future work, we are also aiming to test our proposed LBD methodology on the better studied medical domain as well as on other domains such as humanities and social sciences.

## References

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd interna-*

*tional conference on Machine learning - ICML '06*, pages 113–120.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Delroy Cameron, Ramakanth Kavuluru, Thomas C. Rindflesch, Amit P. Sheth, Krishnaprasad Thirunarayan, and Olivier Bodenreider. 2015. Context-driven automatic subgraph creation for literature-based discovery. *Journal of Biomedical Informatics*, 54:141–157.

Chris Cheadle, Hongbao Cao, Andrey Kalinin, and Jaqui Hodgkinson. 2016. Advanced literature analysis in a Big Data world. *Annals of the New York Academy of Sciences*, 1387(1):25–33.

Kenneth A. Cory. 1997. Discovering hidden analogies in an online humanities database. *Computers and the Humanities*, 31:1–12.

Mc Ganiz, Wm Pottenger, and Cd Janneck. 2005. Recent advances in literature based discovery. Technical report.

M. Gordon and R. K. Lindsay. 2002. Literature-based discovery on the world wide web. *ACM Transactions on Internet Technology*, 2:261–275.

Michael D Gordon and Susan Dumais. 1998. Using latent semantic indexing for literature based discovery. *J. Am. Soc. Inf. Sci. Technol.*, 49(8):674–685.

Sam Henry and Bridget T. McInnes. 2017. Literature Based Discovery: Models, methods, and trends.

Dimitar Hristovski, Carol Friedman, Thomas C Rindflesch, and Borut Peterlin. 2006. Exploiting semantic relations for literature-based discovery. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 349–353.

Jose L. Hurtado, Ankur Agarwal, and Xingquan Zhu. 2016. Topic discovery and future trend forecasting for texts. *Journal of Big Data*, 3(1).

V. Ittipanuvat, K. Fujita, I. Sakata, and Y. Kajikawa. 2014. Finding linkage between technology and social issue: a literature based discovery approach. *Journal of Engineering and Technology Management*, pages 160–184.

Ronald N. Kostoff. 2014. Literature-related discovery: Common factors for Parkinson's Disease and Crohn's Disease. *Scientometrics*, 100(3):623–657.

Ronald N. Kostoff, Jeffrey L. Solka, Robert L. Rushenberg, and Jeffrey A. Wyatt. 2008. Literature-related discovery (LRD): Water purification. *Technological Forecasting and Social Change*, 75(2):256–275.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and hrases and their compositionality. In *NIPS*, pages 1–9.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, pages 1–12.

Ji Qi and Yukio Ohsawa. 2016. Matrix-like visualization based on topic modeling for discovering connections between disjoint disciplines. *Intelligent Decision Technologies*, 10(3):273–283.

Yakub Sebastian, Eu Gene Siew, and Sylvester O. Orimaye. 2017a. Emerging approaches in literature-based discovery: techniques and performance review.

Yakub Sebastian, Eu Gene Siew, and Sylvester Olubolu Orimaye. 2015. Predicting future links between disjoint research areas using heterogeneous bibliographic information network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9078, pages 610–621.

Yakub Sebastian, Eu Gene Siew, and Sylvester Olubolu Orimaye. 2017b. Learning the heterogeneous bibliographic information network for literature-based discovery. *Knowledge-Based Systems*, 115:66–79.

Hadi Shakibian and Nasrollah Moghadam Charkari. 2017. Mutual information model for link prediction in heterogeneous complex networks. *Scientific Reports*, 7.

D. R. Swanson. 1988. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31:526–557.

Don R. Swanson. 1986. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1):1–18.

Don R. Swanson and Neil R. Smalheiser. 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433.

Marc Weeber, Henny Klein, Lolkje T.W. De Jong-Van Den Berg, and Rein Vos. 2001. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557.

Meliha Yetisgen-Yildiz and Wanda Pratt. 2009. A new evaluation methodology for literature-based discovery systems. *Journal of Biomedical Informatics*, 42(4):633–643.