Book Section:

Kibwami, N and Tutesigensi, A orcid.org/0000-0002-5514-1594 (2021) Using literature-based discovery in built environment research. In: Manu, E and Akotia, J, (eds.) Secondary Research Methods in the Built Environment. Routledge , London270 . ISBN 9780367429881

https://doi.org/10.1201/9781003000532

CHAPTER (X)

USING LITERATURE-BASED DISCOVERY IN BUILT ENVIRONMENT RESEARCH

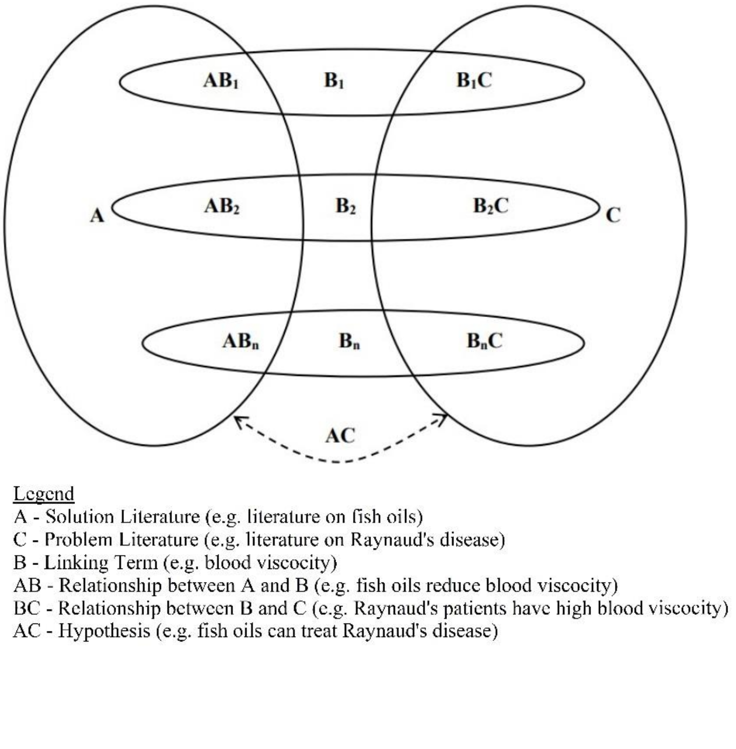
Nathan Kibwami and Apollo Tutesigensi

SUMMARY

Literature-based discovery (LBD) involves identifying novel relationships/theories from two or more disparate contexts of literature. Particularly, LBD can be used to investigate or search for novel hypotheses; this method has successfully been used in biomedical research, from where it originated. However, use of LBD outside the biomedical domain is still scanty. In built environment (BE) research, available studies suggest there is limited detail about the use of LBD. In addition, there is apparent confusion between LBD and other literature-related approaches such as systematic literature review. In this chapter, we focus on facilitating development of a robust understanding of LBD among BE researchers in order to increase use of LBD in BE research. We propose a five-step approach to implementing LBD and, hence, demonstrate how the core principles of LBD can be upheld. The proposed approach facilitates rigorous development of plausible research hypotheses and/or justification of existing practices/knowledge based on secondary information from disparate contexts.

1. Introduction

Literature-based discovery (LBD), proposed by Don R. Swanson (Swanson, 1986), is a form of text interrogation of juxtaposed two (or more) disparate scientific literatures to identify "… nontrivial assertions that are implicit …" (Smalheiser, 2012, p.218). LBD presents a systematic two-component approach to bridging disparate research disciplines, through text mining (Kostoff, 2006, p.924). The text mining aspect involves extraction and analysis of information expressed in form of text (Kostoff, 2006, Smalheiser, 2012, p.218, Miyanishi et al., 2010, Ittipanuvat et al., 2013). If indirect linkages are found between the disparate literatures and no one has previously reported them, new knowledge is created (Weeber et al., 2001). As such, LBD follows a kind of syllogism (see Figure 1) prescribing that for disparate literatures A and C, if A reports a relationship (AB) with a term B, C reports a relationship (BC) with the same term B, hypotheses (AC) can be derived connecting A and C (Smalheiser, 2012).



Legend
A - Solution Literature (e.g. literature on fish oils)
C - Problem Literature (e.g. literature on Raynaud's disease)
B - Linking Term (e.g. blood viscocity)
AB - Relationship between A and B (e.g. fish oils reduce blood viscocity)
BC - Relationship between B and C (e.g. Raynaud's patients have high blood viscocity)
AC - Hypothesis (e.g. fish oils can treat Raynaud's disease)

Figure 1: *The ABC model of LBD; dotted arrow denotes potential hypotheses connecting A and C (Adapted from Artificial Intelligence, 91, Swanson, D. R. & Smalheiser, N. R., An interactive system for finding complementary literatures: A stimulus to scientific discovery, 183-203, Copyright (1997), with permission from Elsevier.)*

Pioneering use of LBD is credited to Swanson's medical-related research work published in 1986. From two disparate scientific literatures, one related to fish oil and another on Raynaud's disease, Swanson "proposed [a] hypothesis that fish oil might ameliorate Raynaud's syndrome" (Swanson, 1986, p.12). This hypothesis was later empirically tested and found acceptable (DiGiacomo et al., 1989). Several subsequent LBD studies in medical disciplines have since been undertaken, resulting in several hypotheses which have been empirically tested and accepted (Srinivasan, 2004, Weeber et al., 2001, Lindsay and Gordon, 1999).

Outside medical research, LBD has been cited in addressing a variety of research problems (Cory, 1997, Ittipanuvat et al., 2014, Kostoff et al., 2008b), some of which belong to built environment (BE) research (Yung et al., 2013, Dixit et al., 2010). Proponents of LBD in BE research opine that this method has potential in addressing some BE research problems, if correctly applied (Kibwami and Tutesigensi, 2014). These sentiments have been supported in works related to research methods for construction (see Fellows and Liu, 2015), and systematic literature reviews related to LBD (see Thilakaratne et al., 2019).

However, the way LBD has, hitherto, been used in BE studies, suggests violation of its fundamental principles. If such violations are not checked, the purposes for which LBD was originally developed will be gradually lost in BE research. In this chapter, we focus on facilitating robust understanding of LBD among BE researchers in order to improve accuracy and frequency of LBD in BE research. In subsequent sections of this chapter we highlight the epistemology of LBD and standard approaches to LBD before discussing current use of LBD in BE research and associated problems. After this, we present an approach to facilitate authentic use of LBD in BE research before drawing conclusions.

2. Epistemology of LBD

The objective of any research is to advance knowledge within the respective field, topic, or context of inquiry. However, there is a possibility that a problem prevalent in a certain context might be unknowingly solved by another disparate field, oblivious to the problem (Hristovski et al., 2005). Such inadvertent solutions can remain undiscovered and consequently unpursued, if no inquiry ever considers the disparate fields together. Revelation of such undiscovered public knowledge is the focus of LBD.

The epistemological assumption of LBD is grounded in the idea that there is knowledge hidden in scientific literature. The sheer volume and growth of scientific literature makes it almost impossible for researchers to be aware or keep up-to-date with advances within, let alone outside, their own research disciplines (Yetisgen-Yildiz and Pratt, 2009, Srinivasan, 2004). Due to the quantity of information, researcher specialisation is necessary but this leads to increased fragmentation, thus, a vast number of mutually isolated specialities (Swanson, 1991, Ittipanuvat et al., 2014). The increased fragmentation gradually creates an infinite growth of indirect connections amongst specialities, some of which might unknowingly offer answers to important prevalent problems (Swanson, 1991).

Therefore, LBD takes the assumption that the sum of the world's knowledge is greater than the sum of knowledge embedded in scientific literature (Cory, 1997). Through a process of 'mining' scientific literature, these implicit linkages carrying implicit knowledge can be revealed, which can lead to creation of new knowledge (Lekka et al., 2011).

3. Approaches to LBD

There are two fundamental approaches to LBD: open discovery and closed discovery (Kostoff et al., 2008a, Weeber et al., 2001).

3.1 Open discovery

In open discovery (OD), the process starts with the problem literature (C), and using a correlation-mining technique, such as 'linking-term count' (Yetisgen-Yildiz and Pratt, 2006), 'linking terms' (B) related to the problem are identified, and then through a similar technique, 'target terms', related to the linking terms, are identified in a disparate 'solution literature' (A) (Yetisgen-Yildiz and Pratt, 2009, Weeber et al., 2001) (See Figure 2). Hypotheses are then derived connecting the problem literature C and solution literature A. Swanson's (Swanson, 1986) early work referred to OD as procedure 1 of LBD (Swanson and Smalheiser, 1997), while others refer to it as 'one node' LBD (Smalheiser et al., 2009) or one directional procedure (Srinivasan, 2004) characterised by generation of hypotheses (Weeber et al., 2001).

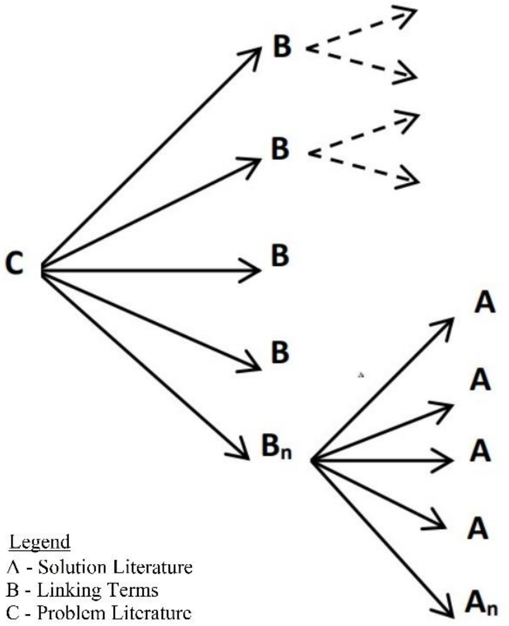


Figure 2: *The open discovery LBD process; dotted lines denote unsuccessful pursuits (Adapted from Journal of the American Society for Information Science and Technology, 52, Weeber, M., Klein, H., De Jong-Van Den Berg, L. T. W. & Vos, R., Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries, 548-557, Copyright (2001), with permission from John Wiley and Sons.)*

3.2 Closed discovery

Closed discovery (CD) starts with the problem literature (C) and solution literature (A) simultaneously as illustrated in Figure 3. The assumption of CD is that some connections between A and C are already known, either derived from the OD process, or stated by conjecture and therefore, the aim of CD is to identify new ones (Srinivasan, 2004). Consequently, common linking B-terms are identified, working towards identifying new linkages between the two literatures (Kostoff et al., 2008a). Various terminology is often used to refer to the CD process e.g. procedure 2 of LBD (Swanson and Smalheiser, 1997), two-node approach (Smalheiser et al., 2009). CD primarily offers a mechanism of testing or confirming hypotheses (Weeber et al., 2001) but in the process, there is potential to generate new hypotheses (Smalheiser et al., 2009).

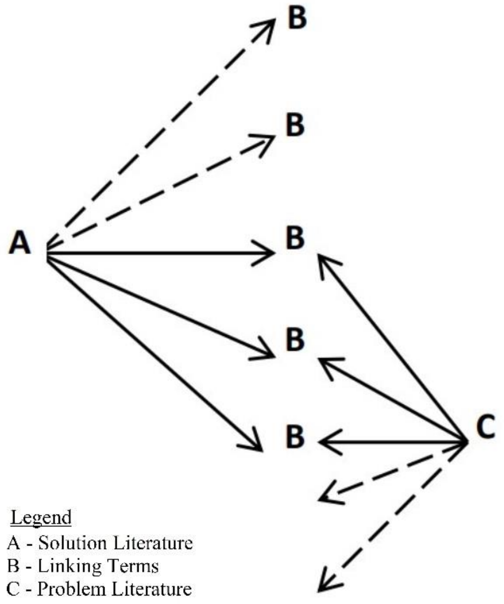


Figure 3: *The closed discovery LBD process; dotted lines denote unsuccessful pursuits (Adapted from Journal of the American Society for Information Science and Technology, 52, Weeber, M., Klein, H., De Jong-Van Den Berg, L. T. W. & Vos, R., Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries, 548-557, Copyright (2001), with permission from John Wiley and Sons.)*

CD is the most predominant approach to LBD and was adopted in the illustrative example presented in this chapter. Indeed, some researchers (e.g. Smalheiser et al., 2009) have developed tools such as Arrowsmith (see Arrowsmith, 2007) that provide guidance on carrying out CD. Arrowsmith uses sophisticated text-mining algorithms and analyses two disparate literatures and returns potential B-terms that the investigator can follow up to discover novel linkages.

4. A Critique of LBD in BE research

Outside medical research, LBD has been cited in a number of studies addressing a variety of research problems including a couple in BE research (Yung et al., 2013, Dixit et al., 2010).

In a study to identify parameters that lead to differing embodied energy measurements among buildings, Dixit and colleagues used LBD as the “research method” (Dixit et al., 2010). They stated that a concept of triangulation, involving “cross-referencing various sources of information about the same phenomenon” was used to identify parameters that caused variation in embodied energy figures of buildings. The study reported that a literature search revealed 10 parameters that influence embodied energy measurements. These results were presented in form of a matrix showing the 10 parameters identified and the respective 23 literature sources underpinning the findings. The study concluded that addressing the identified parameters could lead to consistent measurement of embodied energy of building materials. However, no explicit conclusions were given on the efficacy of using LBD. A similar approach of using LBD was employed in another related study by the same first author (Dixit et al., 2013).

A study seeking to provide an audit of life cycle energy analysis of buildings stated that "a literature-based discovery method has been adopted" (Yung et al., 2013). As justification of research method, the study, citing Dixit et al. (2010), stated that LBD has previously been applied to energy studies of buildings. In the study, a literature review on lifecycle energy studies of buildings was done, consequently identifying 38 research works consisting of 206 cases (i.e. buildings), across 16 countries. The embodied energy in each of the cases was identified and summaries were presented in scatter plots showing differences in the total embodied and operational energy of various types of buildings. The study concluded that, by using LBD, a database of lifecycle energy of buildings was created.

Based on evidence gathered from medical LBD studies, where LBD originated, LBD requires more than one context of literature that are disparate. Literature search has to be performed on two disparate contexts. However, single context literature was used in the two BE studies (Yung et al., 2013, Dixit et al., 2010). The literature considered in Dixit et al. (2010) was limited to embodied energy analysis, while that in Yung et al. (2013) was limited to lifecycle energy analysis of buildings. Using a single context literature does not facilitate LBD since it is expected that the solution or problem would have been discovered or well known within that context (Kostoff et al., 2008a). In addition, the literature used in these BE studies was not disparate since it had several cross citations. For instance, the article (Hammond and Jones, 2008) cited another (Menzies et al., 2007) yet both were used in Dixit et al. (2010). Similarly, in Yung et al. (2013), articles (e.g. Monahan and Powell, 2011, Asif et al., 2007) are used yet one of them cites the other. By violating LBD's fundamental principle of 'using multiple disparate literatures', these BE studies' claim of using LBD is not well founded.

Studies outside BE utilising LBD (Cory, 1997, Ittipanuvat et al., 2014, Kostoff et al., 2008b) provide clear articulation of the application of authentic LBD. In such studies, features of LBD (e.g. linking of terms, hypotheses, etc.) were explicitly articulated. However, besides mere citing that LBD was used, and use of literature databases, there is no explicit articulation of how and why LBD was used in the BE studies (Dixit et al., 2010, Dixit et al., 2013, Yung et al., 2013). While work in Dixit et al. (2010) acknowledges using LBD, nothing is mentioned on how LBD fits into the aim of the work. Similarly, in Yung et al. (2013, p.45) there are no explanations as to why LBD was appropriate to use other than claiming that it "matches perfectly the aim of [that] paper". Since there is little evidence and details to confirm that LBD featured in addressing the problem, it is difficult to conclude that LBD was appropriate for the research questions addressed in these BE studies.

Furthermore, confusing LBD with systematic literature review is evident since barely any features of LBD can be traced in these BE studies. Hence, there is little evidence to reject a hypothesis that they simply carried out (systematic) literature reviews. For instance, their results do not provide any linkages or hypotheses typical of LBD findings, but are instead summarised in a fashion typical of systematic literature reviews. Certainly, these BE studies are not different from other similar literature review studies (e.g. Ibn-Mohammed et al., 2013, Dakwale et al., 2011, Ramesh et al., 2010, Casals, 2006, Menzies et al., 2007) that do not mention LBD.

Overall, it appears that some BE researchers assume LBD to be another form of literature review, a situation which potentially propagates confusion with respect to the differences between LBD and literature review. LBD and literature review are not the same and if LBD BE studies do not articulate this, confusion of LBD and literature review will persist. This will unfortunately hinder full exploitation of the potential of LBD in BE research. This situation must be avoided. This is our motivation for providing guidance for BE researchers using LBD in section 5 below.

5 The proposed LBD approach

Following a CD process, an LBD approach composed of five steps is proposed in this chapter. Being a CD based LBD approach, it presupposes that the investigator has specified the two disparate literatures/contexts, with some preliminary propositions to pursue the investigation.

5.1 Step 1: Literature data retrieval

A comprehensive literature search is performed on two identified disparate 'contexts' of inquiry (i.e. A and C) to generate the corpora for performing LBD; preference should be made to peer-reviewed journal articles. This step is in many ways similar to that of systematic searches performed in conducting systematic literature reviews.

5.2 Step 2: Term extraction

The term-extraction procedure used needs to consider both ‘recall’ and ‘precision’ (Naumann and Herschel, 2010). ‘Recall’ relates to the number of terms that can be retrieved, whereas ‘precision’ is related to the relevance or plausibility of the extracted terms. Higher ‘precision’ can only be guaranteed at the expense of lower ‘recall’, and vice versa (Ganti and Sarma, 2013). Therefore, term extraction procedures (or software) that solely rely on statistical information (e.g. frequency of occurrence) are not preferable since many implausible terms (e.g. is, of, the, etc.) will be extracted. An approach that can balance both precision (based on linguistics) and recall (based on statistics) is necessary. The C-value/NC-value method of text mining (Frantzi et al., 1998) is recommended since it considers both statistical and linguistic information.

On statistical information

Term extraction in LBD involves using statistical procedures (Yetisgen-Yildiz and Pratt, 2006, Frantzi et al., 1998). These statistical procedures usually treat a particular string of characters solely as an instance of a word or phrase without reference to its deeper linguistic significance (Lindsay and Gordon, 1999, p.575), in a way similar to text extraction processes prevalent in other techniques of textual analysis such as manifest content analysis (Hsieh and Shannon, 2005). The LBD term-extraction statistics commonly used, and also included in the suggested approach are: term frequency (Tf), document frequency (Df), inverse document frequency (iDf), and term frequency - inverse document frequency (Tf-iDf) (Lindsay and Gordon, 1999, Ittipanuvat et al., 2014, Frantzi et al., 1998, Gordon et al., 2002).

Initially, terms should be sorted or ranked by their Tf (i.e. number of times a term appears in a corpus). However, using Tf alone for further evaluation means that terms appearing less frequently might be missed out (i.e. since they are low ranked), yet they may be plausible. To circumvent this, the concept of iDf developed in Jones (1972) is suggested. The iDf weighting boosts terms with low frequency, yet concentrated in few specific documents/articles. It is computed as log (D/Df), where D is the total number of documents in the corpus considered and Df is the number of documents in which a term appears. This consequently yields a Tf-iDf measure, which is the product of Tf and iDf (Salton and Buckley, 1988). Tf-iDf is usually a preferred measure and has been cited in several LBD studies (Lindsay and Gordon, 1999, Ittipanuvat et al., 2013, Srinivasan, 2004) as a better measure of relevance of a term than Tf. Therefore, terms should be ranked by Tf-iDf and low-ranking terms may be discarded.

On specifying linguistic information

Without applying any linguistic information in term extraction, statistical measures alone can generate many terms most of which might be implausible (i.e. with low precision). However, using linguistic information such as tagging, linguistic filtering, and stop word list (see Frantzi et al., 1998) improves on precision of terms to be extracted from each context. In applying linguistic information, tagging is used to attach grammatical tags (e.g. noun, adjective) to each term in the corpus. The appropriate linguistic filter applied then uses the grammatical tags to extract specified terms (e.g. nouns only, verbs only, adjectives only etc.). Illustrations of linguistic filtering reveal that most terms are usually composed of nouns, verbs or adjectives and for multiword terms, they are usually constituted of at least a noun (Frantzi et al., 1998). Thus terms extracted should be linguistically filtered to nouns, verbs and adjectives in that order of preference. For automated filters like those in medical databases, linguistic filtering can be automatically set. The approach suggested herein is semi-automated since linguistic filtering is manually done by inspecting each extracted term. The stop word list, which has previously been employed in many LBD studies (see Lindsay and Gordon, 1999, Swanson and Smalheiser, 1997) is then used to distinguish between potentially useful and non-useful terms such as those that frequently occur (e.g. terms like 'is', 'to', 'what' etc.) (Weeber et al., 2001). The stop list can be precompiled based on the predicted suitability of terms (Swanson and Smalheiser, 1997) or compiled concurrently with the term extraction process (Lindsay and Gordon, 1999). As part of the synonymy and stemming rules (Lindsay and Gordon, 1999), it is suggested that only exactly (i.e. not synonyms) matching words should be considered in order to control unnecessary recall and noise. However, singular-plural stemming rules (Lindsay and Gordon, 1999) can be applied and in such cases, the terms (e.g. house and houses) should be combined into one.

Depending on the procedure used for extracting terms (i.e. manual, automatic), a manageable number and length of terms should be considered. In well-structured and online corpora (e.g. in MEDLINE), it is possible to know the approximate number of terms to work with (see Weeber et al., 2001, p.551). However, for a semi-automated process suggested in this work, where articles are manually gathered from different databases, only an estimate can be possible. For instance, for literature consisting 20 articles, assuming an average full-article length of 7000 words, this would constitute working with 140,000 terms. To manage the winnowing process towards precision, an initial working number of terms from each context should be set. Meanwhile, the decision of setting the minimum length of terms (i.e. number of characters per term) depends on the desired precision and recall. Shorter terms are better on recall but not precision. Also, terms can be unigrams (i.e. one word terms), bigrams (i.e. two word terms) or n-grams (Frantzi et al., 1998, Ittipanuvat et al., 2013). Because of some limitations highlighted later, the current approach considers unigrams. The 'recall' for unigrams is usually high since, unigrams can exist either on their own, or as nested terms (i.e.

sub-terms of bigrams/n-grams). In Ittipanuvat et al. (2014), unigrams accounted for over three quarters of the total terms extracted.

5.3 Step 3: Category development

Unlike in biomedical databases where terms can be automatically classified into their respective predetermined semantic categories (see Smalheiser et al., 2009), manual categorisation is suggested, which rather demands 'human intervention' and acquaintance with qualitative data analysis techniques such as latent content analysis (Hsieh and Shannon, 2005) and the paradigm model (Strauss and Corbin, 1998). However, the intensity of human intervention does not entirely manifest as a disadvantage since it gets the analyst up-close with 'what the literature is saying'. To guide the categorisation process, a paradigm model, initially proposed in Strauss and Corbin (1998) and subsequent texts (Corbin and Strauss, 2008), is suggested. It consists of phenomena (i.e. what is going on?), conditions (i.e. what are the causes), actions/interactions (i.e. what is the response?) and consequences (i.e. what are the results?). Following such questions posed in the components of the paradigm model, for each of the key terms, the literature where it appears is read by line, paragraph, page or entire article, in order to elucidate the context of how the term was used. Essentially, the approach involves 'coding', "… an analytic process through which data are fractured, conceptualised, and integrated …" (Strauss and Corbin, 1998, p.3). Coding is done by sentence and paragraph, following the key search terms only, and appropriate software can be used to aid the process. Consequently, categories are developed from the key terms and it is possible for a given term to belong to several categories.

5.4 Step 4: Semantic similarity

A semantic similarity measure is a "… function that, given two […] sets of terms annotating two entries, returns a numerical value reflecting the closeness in meaning between them" (Pesquita et al., 2009 p.1). Literature (Ganti and Sarma, 2013, Naumann and Herschel, 2010, Pesquita et al., 2009) discusses several similarity measures (e.g. Jaccard Index, Dice coefficient, and cosine) which are also often used in LBD studies (see Ittipanuvat et al., 2013, Miyanishi et al., 2010). This work suggests the cosine similarity measure. At this stage, categories based on the key terms only would have been developed from the corpora. From appropriate coding software, it is possible to map the 'extracted terms' that intersect with a given category and therefore, several term-combinations (e.g. terms in both A and C, in A only, and in C only) associated with a developed category can be worked out. Like the key terms, it is possible for a given extracted term to belong in several categories.

The categories are then transformed into vectors and the similarity between them is computed. Since vectors only work with integers, each term is therefore represented by its Tf-iDf measure. Put another way, a category composed of terms is represented as a vector composed of Tf-iDfs. This idea, initially suggested in Salton and Buckley (1988), is usually used in works related to textual analyses, document indexing, and document retrieval. The similarity between two vectors is a property of the cosine of the angle between them (i.e. 1 if the vectors are identical and zero if they are not). The cosine values are computed using the cosine vector similarity formula (Salton and Buckley, 1988, p.514) as per Equation (1) below:

$$Similarity\ (A_v, C_v) = \sum\left(w_{A,t} \times w_{C,t}\right) / \left(\sqrt{\sum w_{A,t}^2} \times \sqrt{\sum w_{C,t}^2}\right) \qquad (1)$$

where A_v and C_v are Tf-iDf vectors representing literature contexts of A and C respectively; $w_{A,t}$ and $w_{C,t}$ are weights (i.e. Tf-iDf) of a term t with regard to literature A and C respectively.

5.5 Step 5: Deducing relationships

This step is the climax of a typical LBD study and involves deduction of relationships (i.e. plausible hypotheses) or confirmation of the same. In this work, deducing relationship is based on the cosine similarity measure and the Tf-iDf measure. It is assumed that vectors (i.e. categories) with cosine similarity values closer to 1 will be more related and thus plausible for generating plausible hypotheses. This assumption is rather not new (see Miyanishi et al., 2010, p.1554), though needs cautious interpretation. Although it would be considered that the lower cosine values offer fewer linkages to explore, they may, ipso facto, be potential sources for novel relationships. Nonetheless, the key guidance to pursue any plausible hypothesis/relationship regarding any term in the vectors is based on the cosine similarity score and the term's rank/weighting (i.e. Tf-iDf). In other words, it is inferred that the plausibility of a hypothesis linking an A-Term to a C-Term is related to the cosine similarity between the two vectors that describe how that terms manifest in A and C.

6. Application of the proposed LBD approach

Table 1 provides a step-by-step summary of the outcomes from the application of the proposed LBD approach undertaken in 2013 aimed at identifying lessons about managing carbon emissions Uganda could learn from United Kingdom (UK). This work was based on the idea that the necessary disparity of literatures for LBD application can be provided by country context (Gordon and Awad, 2008). Personal experience and anecdotal evidence suggested that there was little, if anything, implemented in the Ugandan building sector to address carbon emissions. This was confirmed by a nil return when a systematic search for literature involving the key words of 'building(s)' or 'construction' and 'carbon emissions' was implemented. A similar search involving UK returned a rich collection of publications. One of the testable hypotheses that emerged brought together the idea of Clean Development Mechanism (CDM) and embodied carbon emissions of buildings in Uganda. Following this, a CDM was proposed (see Kibwami and Tutesigensi (2015)) and may be tested in future.

Table 1: *Illustration of the new LBD approach*

Step	**Action and Outcomes**
1. Literature data retrieval	A comprehensive literature search about carbon emissions in Uganda (comprising problem literature C) and carbon emissions in UK (comprising solution literature A) was undertaken; a total of 105 articles were identified (29 and 76 for C and A respectively).
2. Term extraction	To balance precision and recall, while presenting a manageable number of terms, 1000 terms were extracted and ranked from each context. A number of terms present in both A and C (i.e. the B-terms) including buildings, climate, renewable, energy, costs and technology emerged.
3. Category development	The most prominent categories identified and considered were strategies to address emissions, causes of emissions, barriers to reducing emissions, and regulations related to emissions. Terms that belonged to each of the categories, with respect to A and C, were extracted and ranked by their Tf-iDf measure.
4. Semantic similarity	The cosine similarity of the category 'strategies to address emissions' between A and C in relation to B terms was 0.8616 (see Appendix A). This result implied a good relationship between A and C, suggesting that perhaps similar initiatives of addressing emissions existed in both contexts. However, when terms not common to both contexts were included in the computations, the similarity reduced to 0.3005 (see Appendix B). This reduction demonstrated that the manifestation of 'strategies to address emissions' in the two contexts was different and there could be ideas about addressing carbon emissions in context A which could benefit context C.
5. Deducing relationships	Upon critical consideration of the A only and C only literature, plausible hypotheses were posed regarding addressing emissions associated with buildings in context A. One such hypothesis was as follows: *CDM* (Clean Development Mechanism) will reduce embodied carbon emissions of *residential* buildings in Uganda.

7. Conclusions

In this chapter, we demonstrate that the utility of LBD lies in its capacity to help researchers generate hypotheses about a given problem by deducing connections between two disparate contexts. Authenticity of application of LBD should always be judged against this principle.

Although LBD is increasingly gaining recognition, it is mostly still limited to addressing medical problems. Nonetheless, its uptake in BE research appears to be emerging, albeit, with some deficiencies. Hitherto, LBD in BE research has been characterised by misconception which has manifested in the form of: focus on single instead of two or more disparate literatures; insufficient consideration to justify choice; and confusing LBD with mainstream systematic literature review. There is urgent need to move from this status quo towards authentic application of LBD in BE research.

To facilitate the move to authentic application of LBD, we have presented a five-step approach of LBD which adheres to the fundamental principles of LBD and illustrated its application with an example. The approach consists of literature data retrieval, term extraction, category development, determination of semantic similarity and deducing of relationships. The proposed approach demonstrates robust use of qualitative information to derive quantitative indicators that enable researchers to link phenomena that appear in disparate contexts and put forward testable hypotheses. We implore BE researchers to tap the demonstrated potential of LBD whilst adhering to the principles and assumptions of the method.

Whereas this work has argued for, and underscored the efficacy of LBD, it would not be complete without underscoring the associated limitations in BE research. Unlike in biomedical research, the approach presented in this work is semi-automated, involving a great effort of human assessment, implying that some tasks are limited to only what could be reasonably handled. As such, there could be potential for bias since human perceptions and opinions differ. However, any of these biases will be found out when the hypothesis testing takes place, so, instead of this being a cause for concern, it should be seen as providing opportunity to discard fantasies. Researchers must, however, be vigilant to avoid missing hypotheses because of researcher bias.

References

Arrowsmith. (2007). *Arrowsmith Project: linking documents, disciplines, investigators and databases.* [Online]. Available: http://goo.gl/JQe6X6 [Accessed 10 November 2019].

Asif, M., Muneer, T. & Kelley, R. (2007). Life cycle assessment: A case study of a dwelling home in Scotland. *Building and Environment,* 42, 1391-1394.

Casals, X. G. (2006). Analysis of building energy regulation and certification in Europe: Their role, limitations and differences. *Energy and Buildings,* 38, 381-392.

Corbin, J. M. & Strauss, A. L. (2008). *Basics of qualitative research : techniques and procedures for developing grounded theory,* Los Angeles, Calif. ; London, SAGE.

Cory, K. (1997). Discovering Hidden Analogies in an Online Humanities Database. *Computers and the Humanities,* 31, 1-12.

Dakwale, V. A., Ralegaonkar, R. V. & Mandavgane, S. (2011). Improving environmental performance of building through increased energy efficiency: A review. *Sustainable Cities and Society,* 1, 211-218.

Digiacomo, R. A., Kremer, J. M. & Shah, D. M. (1989). Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *American Journal of Medicine,* 86, 158-64.

Dixit, M. K., Culp, C. H. & Fernández-Solís, J. L. (2013). System boundary for embodied energy in buildings: A conceptual model for definition. *Renewable and Sustainable Energy Reviews,* 21, 153-164.

Dixit, M. K., Fernández-Solís, J. L., Lavy, S. & Culp, C. H. (2010). Identification of parameters for embodied energy measurement: A literature review. *Energy and Buildings,* 42, 1238-1247.

Fellows, R. F. & Liu, A. (2015). *Research methods for construction,* Oxford, Wiley-Blackwell.

Frantzi, K., Ananiadou, S. & Tsujii, J. 1998. (The C-value/NC-value Method of Automatic Recognition for Multi-word Terms). *Research and Advanced Technology for Digital Libraries.* Springer Berlin Heidelberg.

Ganti, V. & Sarma, A. D. (2013). Data Cleaning: A Practical Perspective. *Synthesis Lectures on Data Management,* 5, 1-85.

Gordon, M., Lindsay, R. K. & Fan, W. (2002). Literature-based discovery on the World Wide Web. *ACM Trans. Internet Technol.,* 2, 261-275.

Gordon, M. D. & Awad, N. F. (2008). The tip of the iceberg: The quest for innovation at the base of the pyramid. *In:* Bruza, P and Weeber, M (Eds.), *Literature-based discovery*, Vol. 15, pp. 23-37: Springer Berlin Heidelberg.

Hammond, G. P. & Jones, C. I. (2008). Embodied energy and carbon in construction materials. *Proceedings of the ICE - Energy,* 161, 87-98.

Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *Int J Med Inform,* 74, 289-98.

Hsieh, H.-F. & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research,* 15, 1277-1288.

Ibn-Mohammed, T., Greenough, R., Taylor, S., Ozawa-Meida, L. & Acquaye, A. (2013). Operational vs. embodied emissions in buildings—A review of current trends. *Energy and Buildings,* 66, 232-245.

Ittipanuvat, V., Fujita, K., Sakata, I. & Kajikawa, Y. (2013). Finding linkage between technology and social issue: A Literature Based Discovery approach. *Journal of Engineering and Technology Management,* http://dx.doi.org/10.1016/j.jengtecman.2013.05.006.

Ittipanuvat, V., Fujita, K., Sakata, I. & Kajikawa, Y. (2014). Finding linkage between technology and social issue: A Literature Based Discovery approach. *Journal of Engineering and Technology Management,* 32, 160-184.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation,* 28, 11-21.

Kibwami, N. & Tutesigensi, A. (2014). Using the literature based discovery research method in a context of built Environment research. *In:* RAIDEN, A. B. & ABOAGYE-NIMO, E. (eds.) *Procs 30th Annual ARCOM Conference, 1-3 September 2014.* Portsmouth, UK: Association of Researchers in Construction Management.

Kibwami, N. & Tutesigensi, A. (2015). Integrating clean development mechanism into the development approval process of buildings: A case of urban housing in Uganda. *Habitat International,* 53, 331-341.

Kostoff, R. N. (2006). Systematic acceleration of radical discovery and innovation in science and technology. *Technological Forecasting and Social Change,* 73, 923-936.

Kostoff, R. N., Briggs, M. B., Solka, J. L. & Rushenberg, R. L. (2008a). Literature-related discovery (LRD): Methodology. *Technological Forecasting and Social Change,* 75, 186-202.

Kostoff, R. N., Solka, J. L., Rushenberg, R. L. & Wyatt, J. A. (2008b). Literature-related discovery (LRD): Water purification. *Technological Forecasting and Social Change,* 75, 256-275.

Lekka, E., Deftereos, S. N., Persidis, A., Persidis, A. & Andronis, C. (2011). Literature analysis for systematic drug repurposing: a case study from Biovista. *Drug Discovery Today: Therapeutic Strategies,* 8, 103-108.

Lindsay, R. K. & Gordon, M. D. (1999). Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science,* 50, 574-587.

Menzies, G. F., Turan, S. & Banfill, P. F. G. 2007. (Life-cycle assessment and embodied energy: a review). *Proceedings of the ICE - Construction Materials* [Online], 160. Available: http://www.icevirtuallibrary.com/content/article/10.1680/coma.2007.160.4.135.

Miyanishi, T., Seki, K. & Uehara, K. (2010). Hypothesis generation and ranking based on event similarities. *Proceedings of the 2010 ACM Symposium on Applied Computing.* Sierre, Switzerland: ACM.

Monahan, J. & Powell, J. C. (2011). An embodied carbon and energy analysis of modern methods of construction in housing: A case study using a lifecycle assessment framework. *Energy and Buildings,* 43, 179-188.

Naumann, F. & Herschel, M. (2010). An Introduction to Duplicate Detection. *Synthesis Lectures on Data Management,* 2, 1-87.

Pesquita, C., Faria, D., Falcao, A. O., Lord, P. & Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology,* 5.

Ramesh, T., Prakash, R. & Shukla, K. K. (2010). Life cycle energy analysis of buildings: An overview. *Energy and Buildings,* 42, 1592-1600.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieaval. *Information Processing and management,* 24, 513-523.

Smalheiser, N. R. (2012). Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology,* 63, 218-224.

Smalheiser, N. R., Torvik, V. I. & Zhou, W. (2009). Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs in Biomedicine,* 94, 190-197.

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology,* 55, 396-413.

Strauss, A. L. & Corbin, J. M. (1998). *Basics of qualitative research : techniques and procedures for developing grounded theory,* Thousand Oaks, Sage Publications.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine,* 30, 7-18.

Swanson, D. R. (1991). Complementary structures in disjoint science literatures. *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval.* Chicago, Illinois, USA: ACM.

Swanson, D. R. & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence,* 91, 183-203.

Thilakaratne, M., Falkner, K. & Atapattu, T. (2019). A Systematic Review on Literature-based Discovery: General Overview, Methodology, &#x0026; Statistical Analysis. *ACM Comput. Surv.,* 52, 1-34.

Weeber, M., Klein, H., De Jong-Van Den Berg, L. T. W. & Vos, R. (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, 548-557.

Yetisgen-Yildiz, M. & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Inform,* 39, 600-11.

Yetisgen-Yildiz, M. & Pratt, W. (2009). A new evaluation methodology for literature-based discovery systems. *J Biomed Inform,* 42, 633-43.

Yung, P., Lam, K. C. & Yu, C. (2013). An audit of life cycle energy analyses of buildings. *Habitat International,* 39, 43-54.

Appendix A Computation of cosine similarity for shared terms

Term (t)	$w_{A,t}$	$w_{C,t}$	$w_{A,t} \times w_{C,t}$	$w_{A,t}^2$	$w_{C,t}^2$
Energy	25.5150	8.4960	216.7755	651.0152	72.1821
Renewable	18.9020	4.5845	86.6563	357.2845	21.0178
Emissions	17.5746	8.2349	144.7255	308.8648	67.8143
Carbon	17.0672	13.8532	236.4352	291.2903	191.9103
New	16.7210	2.8943	48.3958	279.5908	8.3771
Solar	16.7210	2.8943	48.3958	279.5908	8.3771
Low	15.4379	2.2923	35.3877	238.3300	5.2544
Electricity	14.4494	4.4891	64.8654	208.7863	20.1523
Sustainable	12.6433	1.4472	18.2968	159.8520	2.0943
Change	11.7680	5.0706	59.6708	138.4863	25.7109
Fuel	11.7402	6.0211	70.6883	137.8316	36.2532
Power	11.7402	2.9101	34.1652	137.8316	8.4687
Sector	11.7402	2.9101	34.1652	137.8316	8.4687
Policy	10.4803	3.8801	40.6653	109.8375	15.0555
Potential	10.4803	3.3804	35.4277	109.8375	11.4271
Construction	9.6108	2.2923	22.0305	92.3678	5.2544
Consumption	9.0309	2.9101	26.2809	81.5572	8.4687
Demand	9.0309	1.4472	13.0691	81.5572	2.0943
Government	8.6497	3.4384	29.7411	74.8180	11.8225
Technology	8.6497	1.4472	12.5175	74.8180	2.0943
Study	8.1278	2.9101	23.6528	66.0613	8.4687
Environmental	7.7505	4.5845	35.5321	60.0698	21.0178
Costs	7.6887	2.9101	22.3748	59.1154	8.4687
Climate	7.2247	6.7337	48.6490	52.1966	45.3426
Increase	7.1962	2.2923	16.4955	51.7853	5.2544
High	6.7276	1.4472	9.7359	45.2603	2.0943
Research	6.6433	2.2923	15.2281	44.1329	5.2544
System	6.1682	6.7608	41.7017	38.0463	45.7082
Available	4.8165	1.4472	6.9702	23.1985	2.0943
Data	4.8165	1.4472	6.9702	23.1985	2.0943
Level	4.8165	3.4384	16.5609	23.1985	11.8225
National	3.0103	4.4891	13.5136	9.0619	20.1523
Average	1.8062	2.9101	5.2562	3.2623	8.4687
Total			1540.9966	4449.9665	718.5391

From the above:

$\sum(w_{A,t} \times w_{C,t})$ = 1540.9966; $\sum w_{A,t}^2$= 4449.9665; and $\sum w_{C,t}^2$ = 718.5391

Hence:

$\sqrt{\sum w_{A,t}^2}$ = 66.7081 and $\sqrt{\sum w_{C,t}^2}$ = 26.8056

Therefore: cosine similarity for shared terms $= \dfrac{\sum(w_{A,t} \times w_{C,t})}{\left(\sqrt{\sum w_{A,t}^2} \times \sqrt{\sum w_{C,t}^2}\right)}$

= 1540.9966 ÷ (66.7081 × 26.8056)

= 0.8618

Appendix B Computation of cosine similarity for all terms

Term (t)	$w_{A,t}$	$w_{C,t}$	$w_{A,t} \times w_{C,t}$	$w_{A,t}^2$	$w_{C,t}^2$
Energy	25.5150	8.4960	216.7755	651.0152	72.1821
Building	24.4449	0.0000	0.0000	597.5517	0.0000
CO2	23.3364	0.0000	0.0000	544.5890	0.0000
Heat	19.9416	0.0000	0.0000	397.6658	0.0000
Heating	18.9874	0.0000	0.0000	360.5195	0.0000
Reduction	18.9874	0.0000	0.0000	360.5195	0.0000
Renewable	18.9020	4.5845	86.6563	357.2845	21.0178
Emissions	17.5746	8.2349	144.7255	308.8648	67.8143
Carbon	17.0672	13.8532	236.4352	291.2903	191.9103
New	16.7210	2.8943	48.3958	279.5908	8.3771
Solar	16.7210	2.8943	48.3958	279.5908	8.3771
Technologies	16.5272	0.0000	0.0000	273.1490	0.0000
Cooling	16.1236	0.0000	0.0000	259.9705	0.0000
Low	15.4379	2.2923	35.3877	238.3300	5.2544
Savings	14.5310	0.0000	0.0000	211.1488	0.0000
Gas	14.5211	0.0000	0.0000	210.8635	0.0000
Measures	14.5211	0.0000	0.0000	210.8635	0.0000
Design	14.5112	0.0000	0.0000	210.5761	0.0000
Homes	14.5112	0.0000	0.0000	210.5761	0.0000
Electricity	14.4494	4.4891	64.8654	208.7863	20.1523
Buildings	13.8611	0.0000	0.0000	192.1298	0.0000
Efficiency	13.0860	0.0000	0.0000	171.2429	0.0000
Zero	12.8989	0.0000	0.0000	166.3811	0.0000
Levels	12.7791	0.0000	0.0000	163.3044	0.0000
Performance	12.7791	0.0000	0.0000	163.3044	0.0000
Sustainable	12.6433	1.4472	18.2968	159.8520	2.0943
Domestic	12.2366	0.0000	0.0000	149.7344	0.0000
Change	11.7680	5.0706	59.6708	138.4863	25.7109
Fuel	11.7402	6.0211	70.6883	137.8316	36.2532
Power	11.7402	2.9101	34.1652	137.8316	8.4687
Sector	11.7402	2.9101	34.1652	137.8316	8.4687
Water	11.7402	0.0000	0.0000	137.8316	0.0000
Significant	11.4391	0.0000	0.0000	130.8539	0.0000
Thermal	11.0752	0.0000	0.0000	122.6598	0.0000
Dwellings	11.0721	0.0000	0.0000	122.5914	0.0000
Space	10.5719	0.0000	0.0000	111.7651	0.0000
Policy	10.4803	3.8801	40.6653	109.8375	15.0555
Potential	10.4803	3.3804	35.4277	109.8375	11.4271
Dwelling	10.2803	0.0000	0.0000	105.6844	0.0000
Emission	10.2803	0.0000	0.0000	105.6844	0.0000
Generation	9.9340	0.0000	0.0000	98.6842	0.0000

House	9.9340	0.0000	0.0000	98.6842	0.0000
Wind	9.9340	0.0000	0.0000	98.6842	0.0000
Construction	9.6108	2.2923	22.0305	92.3678	5.2544
Air	9.2523	0.0000	0.0000	85.6043	0.0000
Housing	9.2523	0.0000	0.0000	85.6043	0.0000
Site	9.2523	0.0000	0.0000	85.6043	0.0000
Standard	9.2523	0.0000	0.0000	85.6043	0.0000
Ventilation	9.2523	0.0000	0.0000	85.6043	0.0000
Consumption	9.0309	2.9101	26.2809	81.5572	8.4687
Cycle	9.0309	0.0000	0.0000	81.5572	0.0000
Demand	9.0309	1.4472	13.0691	81.5572	2.0943
Embodied	9.0309	0.0000	0.0000	81.5572	0.0000
Model	8.8577	0.0000	0.0000	78.4585	0.0000
Government	8.6497	3.4384	29.7411	74.8180	11.8225
Technology	8.6497	1.4472	12.5175	74.8180	2.0943
Compared	8.2242	0.0000	0.0000	67.6380	0.0000
Build	8.1278	0.0000	0.0000	66.0613	0.0000
Study	8.1278	2.9101	23.6528	66.0613	8.4687
Residential	7.9744	0.0000	0.0000	63.5903	0.0000
Environmental	7.7505	4.5845	35.5321	60.0698	21.0178
Future	7.7505	0.0000	0.0000	60.0698	0.0000
Stock	7.7505	0.0000	0.0000	60.0698	0.0000
Wall	7.7505	0.0000	0.0000	60.0698	0.0000
Costs	7.6887	2.9101	22.3748	59.1154	8.4687
Climate	7.2247	6.7337	48.6490	52.1966	45.3426
Impact	7.2247	0.0000	0.0000	52.1966	0.0000
Increase	7.1962	2.2923	16.4955	51.7853	5.2544
Results	7.1962	0.0000	0.0000	51.7853	0.0000
Existing	6.7276	0.0000	0.0000	45.2603	0.0000
High	6.7276	1.4472	9.7359	45.2603	2.0943
Household	6.6453	0.0000	0.0000	44.1599	0.0000
Hot	6.6433	0.0000	0.0000	44.1329	0.0000
Period	6.6433	0.0000	0.0000	44.1329	0.0000
Research	6.6433	2.2923	15.2281	44.1329	5.2544
Higher	6.1682	0.0000	0.0000	38.0463	0.0000
System	6.1682	6.7608	41.7017	38.0463	45.7082
Current	6.0206	0.0000	0.0000	36.2476	0.0000
Built	5.5361	0.0000	0.0000	30.6478	0.0000
Control	5.5361	0.0000	0.0000	30.6478	0.0000
Internal	5.3162	0.0000	0.0000	28.2624	0.0000
Required	5.3162	0.0000	0.0000	28.2624	0.0000
Available	4.8165	1.4472	6.9702	23.1985	2.0943
Data	4.8165	1.4472	6.9702	23.1985	2.0943
Effect	4.8165	0.0000	0.0000	23.1985	0.0000
Level	4.8165	3.4384	16.5609	23.1985	11.8225
SAP	4.8165	0.0000	0.0000	23.1985	0.0000
Assessment	3.0103	0.0000	0.0000	9.0619	0.0000
National	3.0103	4.4891	13.5136	9.0619	20.1523

Scenario	3.0103	0.0000	0.0000	9.0619	0.0000
Scenarios	3.0103	0.0000	0.0000	9.0619	0.0000
Annual	1.8062	0.0000	0.0000	3.2623	0.0000
Average	1.8062	2.9101	5.2562	3.2623	8.4687
London	1.8062	0.0000	0.0000	3.2623	0.0000
Models	1.8062	0.0000	0.0000	3.2623	0.0000
Office	1.8062	0.0000	0.0000	3.2623	0.0000
Temperatures	1.8062	0.0000	0.0000	3.2623	0.0000
Weather	1.8062	0.0000	0.0000	3.2623	0.0000
Project	0.0000	20.2011	0.0000	0.0000	408.0836
Uganda	0.0000	14.7182	0.0000	0.0000	216.6239
CDM	0.0000	11.3731	0.0000	0.0000	129.3479
Forest	0.0000	7.3591	0.0000	0.0000	54.1560
Development	0.0000	7.2247	0.0000	0.0000	52.1966
Improved	0.0000	6.8768	0.0000	0.0000	47.2899
Diesel	0.0000	6.6227	0.0000	0.0000	43.8596
Land	0.0000	6.0211	0.0000	0.0000	36.2532
Rural	0.0000	6.0211	0.0000	0.0000	36.2532
Biomass	0.0000	5.9855	0.0000	0.0000	35.8263
Market	0.0000	5.8202	0.0000	0.0000	33.8750
Stoves	0.0000	5.8202	0.0000	0.0000	33.8750
Wood	0.0000	4.8502	0.0000	0.0000	23.5243
Charcoal	0.0000	4.5845	0.0000	0.0000	21.0178
Briquettes	0.0000	4.3415	0.0000	0.0000	18.8484
International	0.0000	4.3415	0.0000	0.0000	18.8484
Local	0.0000	4.3415	0.0000	0.0000	18.8484
Global	0.0000	4.2255	0.0000	0.0000	17.8548
Many	0.0000	4.2255	0.0000	0.0000	17.8548
Small	0.0000	4.2255	0.0000	0.0000	17.8548
Countries	0.0000	3.8801	0.0000	0.0000	15.0555
Activities	0.0000	3.7409	0.0000	0.0000	13.9946
Benefits	0.0000	3.4384	0.0000	0.0000	11.8225
Private	0.0000	3.4384	0.0000	0.0000	11.8225
Support	0.0000	3.3804	0.0000	0.0000	11.4271
Capacity	0.0000	2.9101	0.0000	0.0000	8.4687
Generator	0.0000	2.9101	0.0000	0.0000	8.4687
Grid	0.0000	2.9101	0.0000	0.0000	8.4687
Implementation	0.0000	2.9101	0.0000	0.0000	8.4687
Supply	0.0000	2.9101	0.0000	0.0000	8.4687
Cooking	0.0000	2.8943	0.0000	0.0000	8.3771
Country	0.0000	2.8943	0.0000	0.0000	8.3771
Health	0.0000	2.8943	0.0000	0.0000	8.3771
Africa	0.0000	2.2923	0.0000	0.0000	5.2544
Community	0.0000	2.2923	0.0000	0.0000	5.2544
Fuelwood	0.0000	2.2923	0.0000	0.0000	5.2544
Human	0.0000	2.2923	0.0000	0.0000	5.2544
Impacts	0.0000	2.2923	0.0000	0.0000	5.2544
Production	0.0000	2.2923	0.0000	0.0000	5.2544

City	0.0000	1.4472	0.0000	0.0000	2.0943
Economic	0.0000	1.4472	0.0000	0.0000	2.0943
Kampala	0.0000	1.4472	0.0000	0.0000	2.0943
Needs	0.0000	1.4472	0.0000	0.0000	2.0943
Plan	0.0000	1.4472	0.0000	0.0000	2.0943
Planning	0.0000	1.4472	0.0000	0.0000	2.0943
Resources	0.0000	1.4472	0.0000	0.0000	2.0943
Social	0.0000	1.4472	0.0000	0.0000	2.0943
Technical	0.0000	1.4472	0.0000	0.0000	2.0943
		Total	1540.9966	12051.2111	2182.8028

From the above:
$\sum(w_{A,t} \times w_{C,t})$ = 1540.9966; $\sum w_{A,t}^2$= 12051.2111; and $\sum w_{C,t}^2$ = 2182.8028

Hence:
$\sqrt{\sum w_{A,t}^2}$ = 109.7780 and $\sqrt{\sum w_{C,t}^2}$ = 46.7205

Therefore: cosine similarity for all terms $= \frac{\sum(w_{A,t}\times w_{C,t})}{\left(\sqrt{\sum w_{A,t}^2}\times\sqrt{\sum w_{C,t}^2}\right)}$

= 1540.9966 ÷ (109.7780 × 46.7205)

= 0.3005