Scientometrics (2016) 108:745–776
DOI 10.1007/s11192-016-2003-5

Time-aware link prediction to explore network effects on temporal knowledge evolution

Nazim Choudhury[1] · **Shahadat Uddin**[1]

Received: 12 December 2015 / Published online: 8 June 2016

Abstract Quantitative measurements of bibliometrics based on knowledge entities (i.e., keywords) improve competencies in tracking the structure and dynamic development of various scientific domains. Co-word networks (a content analysis technique and type of knowledge network) are often employed to discern relationships among various scientific concepts in scholarly publications to reveal the development and evolution of scientific knowledge. In relation to evolutionary network analysis, different link prediction methods in network science can assist in the prediction of missing links and modelling of network dynamics. These traditional methods (based on topological similarity scores and time series methods of link prediction) can be used to predict future co-occurrence trends among scientific concepts. This study attempted to build supervised learning models for link prediction in co-word networks using network topological similarity metrics and their temporal evolutionary information. In addition to exploring the underlying mechanism of temporal co-word network evolution, classification datasets containing links with both positive and negative labels were also built. A set of topological metrics and their temporal evolutionary information were produced to describe instances of classification datasets. Supervised classifications methods were then applied to classify the links and accurately predict future associations among keywords. Time series based forecasting methods were used to predict the future values of topological evolution. Results in relation to supervised link prediction by different classifiers showed that both static and dynamic information are valuable in predicting new links between literary concepts extracted from scientific literature.

Keywords Link prediction · Knowledge evolution · Keyword network · Network effect

✉ Shahadat Uddin
shahadat.uddin@sydney.edu.au

[1] The Centre for Complex Systems, The University of Sydney, Sydney, NSW, Australia

Introduction

The creation of scientific knowledge has become more dynamic and interdisciplinary, as new avenues of scientific research emerge from new connections forged among disjoint and existing areas of science (Pan et al. 2012). Synthesising existing pools of scientific concepts often aids the creation of new hypotheses in scientific research (Choi et al. 2011). Numerous digital libraries and the exponential growth of scientific literature have liberated and massively enriched these ever-expanding concept pools and account for the rapid shift in scientific research trends from individual to multi-disciplinary domains. Today, scientists postulate new hypotheses encompassing concepts from multiple domains; for example, diverse concepts from the domains of physics, computer science and social science have contributed to the rise of 'network science'. The eminent source of any scientific concept is the scientific article in which the author(s) first conceptualise their work using appropriate keywords. These descriptor keywords represent the thematic context of science and are also known as knowledge entities (Ding et al. 2013). Many of these knowledge entities are created with a degree of independence; however, due to the amalgamation of various scientific domains, interrelationships among these entities are often unknown, even to their originators. Further, innovative scientific questions can now be answered through re-combinations and the association of concepts from multiple articles across interdisciplinary domains (van der Eijk et al. 2004).

Bibliometrics, a term first coined by Belgian librarian Paul Otlet in 1934 (Rousseau 2014), uses quantitative methods to map the structure of this interdisciplinary research trend (Van Raan 2003; Abbasi et al. 2011; Uddin et al. 2012, 2013). One bibliometric method for finding associations among interdisciplinary scientific concepts is based on the notion of co-occurrence; that is, the simultaneous appearance of feature items. Of various co-occurrence based techniques in bibliometrics (Waltman et al. 2010), the co-word analysis (Callon et al. 1983)—a content analysis technique, is a contemporary and quantitative linguistic method that provides both an intuitive (Wu and Leu 2014) and cognitive picture (Rip and Courtial 1984) of literary content. The term 'intuitive' reflects the fact that co-word network ostensibly discerns the intrinsic and complex relationships among different concepts from literature, and 'cognitive' denotes the intellectual and subjective information in regards to these relationships. Further, it maps the strength of associations between knowledge items from textual data by pursuing the interrelationships of science (He 1999). Combination of relational bibliometrics and network science to build a network of knowledge entities resulted in co-word network which is also designated as the knowledge network of science (Wang et al. 2010; Uddin et al. 2015; Khan et al. 2016). Leydesdorff (2002) conjectured that the dynamics of science is reflected through scientific literature and co-word networks are capable of modelling the dynamics of scientific knowledge structures (Leydesdorff 1996). Thus, from the aforementioned perspectives, co-word networks offer an effective approach not only to discern correlations among various research themes and trends but also map the evolution of knowledge across various domains (Su and Lee 2010). Due to the network encoding of relationships among keywords/key-concepts from scientific texts, it is also applicable to network and graph related measurements. Further, Wang et al. (2010) also found that co-word networks conform to and have an affinity with 'small world' and 'scale free' effects (two important properties of complex networks).

Similar to science itself and like many other real-world complex networks, co-word networks are highly dynamic. The evolution and development of science and technology

create new knowledge from previously accumulated and ubiquitous information (Lee et al. 2009). New relationships concurrently emerge among prevailing concepts with the knowledge growth over time. Related to this self-organizing development of science exhibiting growths, shrinkages and emergent behaviours (Sun et al. 2012), co-word networks also evolve both temporally (Ronda-Pupo and Guerras-Martin 2012) and spatially (e.g., power-law cluster-size distributions) (Van Raan 1997). New scientific concepts add new nodes and new hypotheses generate new links in co-word network. On the other hand, aggregated studies (conducted in respect of a single hypothesis that can be related to other hypothesis) reinforce the relationship among corresponding keywords/concepts, and new scientific domain engenders the rise of new subject categories within a network. From the perspective of modelling network dynamics and its evolution, evolutionary network analyses (e.g., link prediction) have been the subject of considerable discussion in the network science community (Zelinka et al. 2012). Link prediction is a time-evolving network analysis model that addresses the problem of predicting the likelihood of future associations among nodes that are missing from a network in its current state. Link prediction also attempts to identify the extent to which network evolution can be modelled using features intrinsic to the network topology itself (Liben-Nowell and Kleinberg 2007). An state-of-the-art categorisations of various link prediction techniques is discussed in survey article by (Wang et al. 2015). These techniques have mostly been based on topological similarity patterns between node pairs of static networks. Temporal link prediction in dynamic networks use time series based link prediction methods (Tylenda et al. 2009; Soares and Prudêncio 2012) that take account of the time-aware evolutionary history of network topologies and also employ different forecasting methods. Researchers have used both supervised and unsupervised methods to assess the viability of these techniques for link prediction.

The majority of the aforementioned link prediction techniques have successfully explored different social scholarly networks, including, co-authorship (Yu et al. 2014), scientific collaboration (Yan and Guns 2014), citation (Shibata et al. 2012) and information networks (Davis et al. 2011); however, little attention has been paid to co-word networks. Similarly, the use of link prediction across co-word networks to understand the evolutionary mechanism has attracted little interest from the network science community. Thus, this study sought to analyse supervised link prediction methods across co-word networks by using a set of features that arose from the network topology. The temporal evolution of network structural information was used to predict future links among different literary concepts. The performance of supervised link prediction was also analysed and compared across dynamic and static networks. This study sought to contribute to the literature by: (1) providing an understanding of the evolution of co-word networks in regards to different types of links constructed using author selected keywords from the scientific literature; (2) predicting topological evolution of future links by utilizing time series forecasting methods; (3) defining positive (i.e., links that truly occur in future) and negative (i.e., links that do not appear in future) classes of links in classification datasets considering topological features of both static and evolving networks; and (4) experimenting and evaluating a set of supervised learning methods to classify these links to predict future associations among keywords successfully. The primary contribution of this research relates to link prediction across both static and dynamic co-word networks using network topological information; however, this could also be valuable in generating new hypothesis for literature based discovery (Smalheiser and Swanson 1998), and forecasting emerging trends (Kontostathis et al. 2004). Further, this study identified a list of important topological features that contribute to co-word network evolution.

Link prediction and topological features

Link prediction is one of the most fundamental frameworks of evolutionary network analysis. Existing link prediction methods can be broadly categorised into similarity based strategies, maximum likelihood algorithms and probabilistic models. Despite most of these methods consider only a static snapshot of a network and ignore the dynamic evolutionary information, they are used in many real-world applications, including scientific collaborations (Wang and Sukthankar 2014), medical informatics (Kastrin et al. 2014) and information and network security (Huang and Zeng 2006). To overcome the issues associated with static networks, researchers have also considered the time-aware evolution of network topologies in dynamic networks (Li et al. 2014). Some analyses have also used the temporal evolution of link occurrences to predict the probability of occurrences for both new and repetitive links in the future (Huang and Lin 2009; Güneş et al. 2016). These methods use the time series approach for link prediction to emulate the dynamic behaviour of complex networks. They also use link creation time to analyse the effect of the elapsed time since a link first appeared and/or how recentness can affect the formation of new links around associated nodes.

Irrespective of the nature of analysis using the different methods described above, researchers have exploited both the node attributes and network structural information of observed links at a specified time T to predict the appearance of new links at the time $T + 1$. Network topology based structural similarity metrics represent the basic and widespread building block of link prediction models. In respect to link prediction, these metrics have shown significant performance improvements compared to other metrics, as they provide objective information about the actual connections in the network. Co-word networks that manifest the network representation of associations among different scientific concepts are also thought to be expedient to network topology based link prediction methods. Thus, in this study, a set of network topology based features was used to predict future associations among scientific concepts. The simplest topological metrics represent a structural similarity between node pairs that depend solely upon network structure. According to their characteristics, these metrics can be subdivided into two broad categories: (1) local or neighbourhood-based metrics; and (2) global metrics containing the ensemble of network path and random walk. The following section describes the metrics explored in this study and the corresponding equations used to compute the similarity between the nodes. The topological similarity score between non-connected nodes u and v is denoted by *score* (u, v) and calculated separately for each metric.

Local metrics

The local similarity metrics used in this study are listed below.

CommonNeighbours

The CommonNeighbours metric (Newman 2001) is the most widespread, basic and simplest type of metric used in link prediction. It measures the number of nodes with which two adjacent nodes have a direct association. If $\lceil(u)$ denotes the set of neighbours to node u, then, using this metric, the similarity score between node u and v is:

$$\text{score}\,(u, v) = |\lceil(u) \cap \lceil(v)|$$

AdamicAdar

The AdamicAdar (Adamic and Adar 2003) metric is similar to the CommonNeighbours metric, except that the high degree of common neighbours is less weighted. The reasoning behind this is that a popular common interest provides less evidence of a strong link between two nodes in a network. The similarity score for node u and v using AdamicAdar is defined as:

$$\text{score}(u,v) = \sum_{z \epsilon \lceil(u) \cap \lceil(v)} \frac{1}{\log|\lceil(z)|}$$

Resource Allocation

An empirical investigation by Zhou et al. (2009) revealed that many links are assigned similar scores for node similarity when only the information of the nearest neighbour is used. To address this issue, they presented a new measure named 'Resource Allocation' to exploit the next nearest neighbour information. Using this method, the similarity score between two nodes is calculated as:

$$\text{score}(u,v) = \sum_{z \epsilon \lceil(u) \cap \lceil(v)} \frac{1}{|\lceil(z)|}$$

Global metrics

In addition to node-neighbourhood information, the network paths between two nodes can also be used to measure the similarity/affinity between node pairs. Similarly, network interactions between node pairs can be computed by random walks on graphs that represent the diffusion of information from one node to the other. A random walk uses transition probabilities from a node to its neighbours to denote the destination of a random walker from a current node. A description of the global metrics used in this study follows.

Katz

The Katz (1953) metric is a variant of the shortest path distance and is also based on the ensemble of all paths. It directly sums all the paths that exist between a pair of vertices. However, to penalise the contribution of longer paths in the similarity computation, it exponentially damps the contribution of a path by a factor of β^l, where l is the path length and β is the damping factor. The similarity score between node u and v using this metric is computed as:

$$\text{score}(u,v) = \sum_{l=1}^{\infty} \beta^l \cdot |p_{u,v}^{(l)}| \; = \beta^1 p_{u,v}^{(1)} + \beta^2 p_{u,v}^{(2)} + \beta^3 p_{u,v}^{(3)} + \cdots$$

Here, $p_{u,v}^{l}$ denotes the path of length l between node u and v. This study used $\beta = 0.005$.

RootedPageRank

Rooted PageRank is a modification of PageRank (Chung and Zhao 2010), a core algorithm used by search engines to rank search results. For link prediction purposes, the random walk assumption of the original PageRank is altered by a similarity score between two vertices (u and v) that is measured as the stationary probability of v in a random walk that returns to u with a probability of $(1 - \alpha)$ in each step, moving to a random neighbour with probability α. The proximity score between node pairs u and v is calculated in this method as follows:

$$\text{score}\,(u, v) = -\mathrm{H}_{u,v} \cdot \pi_v$$

Here $\mathrm{H}_{u,v}$ is the expected time/step for random walk from u to reach v and π_v is the stationary distribution weight of v under the following random walk condition:

$$\begin{cases} \text{with probability } \alpha \text{ jump to } u \\ \text{with probability } 1 - \alpha \text{ jump to random neighbour of the current node} \end{cases}$$

This study used $\alpha = 0.005$.

SimRank

SimRank (Jeh and Widom 2002) adopts the idea that nodes linked to similar nodes are similar. SimRank method begins with the assumption that any node is maximally similar to itself and employs a decay factor of γ to determine how quickly similarities or weights of the connected nodes decrease as they get farther away from the original nodes. In link prediction, the proximity score using SimRank is computed as follows:

$$\text{score}\,(u, v) = \begin{cases} 1 & \text{if } u = v \\ \gamma \cdot \dfrac{\sum_{a \in \lceil(u)} \sum_{b \in \lceil(v)} \text{score}(u, v)}{|\lceil(\mathrm{u})||\lceil(\mathrm{v})|} & \text{otherwise} \end{cases}$$

In this study, the parameter γ was set to 0.8.

Aggregated features

In addition to topological proximity measures, individual network attributes can provide useful information to support link predictions. Centrality measures are the simplest of these attributes. As these attributes pertain to an individual node in a network, some aggregation functions need to be used to combine the attribute values of the corresponding nodes in a node pair. In this study, we used the sum function to aggregate the following centrality measures of each node in a node pair.

Degree centrality

The number of neighbours in the neighbourhood adjacent to a particular node defines its nodal degree or degree of connections and measures its participation in the network. Degree centrality assists in quantifying the momentum of knowledge convergence and diffusion by measuring knowledge flow from source nodes to target nodes. The intuition behind using degree centrality is that, the possibility of acquiring more connections by

nodes in a network will increase the likelihood of forming new links. The degree centrality C_u^d of a node u is calculated as follows.

$$C_u^d = \frac{\sum_{v:v\neq u} p_{uv}}{n-1}$$

where $p_{uv} = 1$ if there exists a link between node u and v or 0 otherwise and n = number of nodes.

Closeness centrality

The closeness centrality of a keyword in co-word network is defined by the inverse of the length of the shortest paths to/from all the other keywords. Higher closeness centrality indicates a higher influence on the other actors in the network. Thus, closeness centrality measures the momentum of influence or being influenced. The more influence a node has over other nodes, the more likely it is that the other nodes will be steered to form links. The degree centrality C_u^c of a node u is calculated as follows.

$$C_u^c = \frac{n-1}{\sum_{v=1}^{n-1} g(u,v)}$$

where $g(u,v)$ denotes the shortest-path distance between node u and v and n denotes the number of nodes in the network

Time series and forecasting models

Failure to acknowledge the dynamicity of a network resulting from changes in its past behaviour may lead to poor performance in regards to the accuracy of predicting future links among nodes. Researchers have used time series analyses and forecasting methods to overcome this issue and follow the frequently changing behaviour of network structure (Soares and Prudêncio 2012; Güneş et al. 2016). Using time series to acquire historical information in relation to the topological changes of non-connected node pairs can increase the accuracy of time series based link predictions. In time series forecasting, past observations of a time variable can be analysed to develop a model that describes the underlying relationship and extrapolation can be used to predict the future values of the variable. In this study, a univariate time series of topological similarity scores was built for non-connected node pairs in relation to individual co-word networks for each year of the training period. A separate time series was built for each similarity metric (see "Link prediction and topological features" section). Three well-known forecasting models with different underlying assumptions were then used to compute the final score for each metric. This became the input for the classification dataset. The three forecasting models used in this study are described below.

Exponential smoothing

Brown, Holt and Winter (De Gooijer and Hyndman 2006) developed the method of exponential smoothing in the late 1950s. Since then, it has motivated some successful forecasting methods. In this method, forecasts are the weighted averages of previous observations and the weights of the observations decay exponentially with time. Single

exponential smoothing (SES) with a weight of α is the simplest exponential smoothing method. The forecast equation can be defined as:

$$\hat{y}_t = \alpha y_{t-1} + (1-\alpha)\hat{y}_{t-1}$$

where $\hat{y}_t$ represents the forecasted value that depends on both the previous observations and previous forecasts. Linear exponential smoothing (LES) is a variation of this method that refines SES with a β component and considers any short trends in the series. The forecasting equation for LES can be described as:

$$\hat{y}_{t+h|t} = l_t + hb_t$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

where l_t is an estimate of the level of the series at time t, b_t denotes an estimate of the trend (i.e., the slope) of the series at time t, α is the smoothing parameter for the level and β is the smoothing parameter for the trend in which $0 \leq \alpha, \beta \leq 1$.

Notably, there are 15 variations of the exponential smoothing process and interested readers should refer to the work by Hyndman et al. (2008) for comprehensive details on this method.

Auto regressive integrated moving average (ARIMA)

First popularised by Box and Jenkins (1976), the auto regressive integrated moving average (ARIMA) model is a well-known linear forecasting technique in the area of short term forecasting. It projects future values of a time series based entirely on its own inertia and performs well where correlations between past patterns are stable. Under the ARIMA model, the future values of a variable are determined using a linear combination of past values and past errors. The model can be expressed as follows:

$$y_t = \theta_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q,}$$

where y_t = actual value, ε_t = random error at time t, φ_i and θ_j are the coefficients and p and q are the integers for the auto regressive (AR) and moving average (MA) polynomials. ARIMA (p, d, q) represents an ARIMA model where p equals the number of autoregressive terms, q equals the number of lagged forecast errors in the prediction equation and d equals the number of non-seasonal differences needed for stationarity.

Random forecasting

As the time series data was short in length and, in some instances, the topological similarity value was zero for keyword pairs due to their absence in corresponding networks, this study used random walk forecasting. Forecasts in this method are based on the last observed value of the time series. Random walk forecasting can be achieved if $n = 1$ is used in the moving average (MA) model. The forecast can be defined as:

$$y_t = y_{t-1}$$

Data acquisition

For data collection, 'Scopus' (Elsevier 1880) (the largest abstract and citation database of peer-reviewed literature, including scientific journals, books and conference proceedings) was used. Scopus delivers a comprehensive overview of the world's research output in science, technology, medicine, social science, the arts and humanities. The search strings for Scopus were 'Project Management' and 'Topic Model'. The latter denotes a technique that has been widely used to discover the main themes pervading large and unstructured texts and organise documents. The following constraints were imposed when searching the Scopus database: (i) the articles had to be published in English journals; (ii) the articles had to be published between the years 2010 and 2015 (inclusive); and (iii) the search terms had to be present in the articles' titles and abstracts. Titles, abstracts and associated author selected keywords were extracted for the 6 year period. The third dataset, 'Obesity', came from a previous research (Khan et al. 2016). However, the entire 20 years of this dataset was not used; rather, author selected keywords and articles published between the years 2008 and 2012 (inclusive) were extracted. In addition to data for this 5-year period, data for the year 2013 was also extracted from Scopus to achieve a similar duration of 6 years (the same period as that used in the other datasets). For the sake of brevity, these three research sections are referred to as OBS, PMG and TM for obesity, project management and topic model, respectively. Table 1, sets out the basic statistics in relation to the number of extracted articles and the author selected keywords for the three datasets.

Research methodology

This section introduces the research methodology and its functional parts (including keywords extraction, pre-processing, co-word network construction, classification dataset construction, forecasting and time series dataset construction) used to model network topological evolution and the supervised learning methodology used to classify positive and negative classes of link instances.

Keyword extraction

The main objective of the keyword extraction phase was to select a set of keywords that represent the characteristics of the selected articles. The relevant keywords act as constituents and can be used to generate co-word networks. Each extracted author selected keywords (as attached to each article) that represent how the author(s) understand their research work and its thematic contexts. To be computationally effective from the

Table 1 Statistics of three datasets

Dataset name	No. of articles	No. of keywords
OBS	44,482	15,022
PMGT	5764	11,253
TM	1613	3668

OBS obesity, *PMGT* project management, *TM* topic model

perspective of link prediction, keywords that appeared in at least more than one article were selected. Keywords that do not appear in more than one article are considered unsuccessful in gaining the attention of other scientific authors and will not significantly affect the prediction of future links.

Text pre-processing

Text transformation is important in retrieving information and discovering the structured non-trivial knowledge from texts. General data cleansing tasks were performed on the author selected keywords that include removing punctuation and accents from words and changing plural nouns to their singular form. Additionally, different representations of some keywords were standardised (e.g., Cocitation ↔ Co-Citation, Coword ↔ Co-word and Neighbor ↔ Neighbour) and different semantically similar concepts with similar meaning were normalised into one keyword (e.g., Longitudinal Study ↔ Longitudinal Analysis ↔ Longitudinal Method, Bayesian Analysis ↔ Bayesian Method ↔ Bayesian Approach). For some concepts, the abbreviated form was considered in place of the concept (e.g., SVM ↔ Support Vector Model/Machine, Body Mass Index ↔ BMI). The plural representations of some keywords were also changed to their single form (e.g., Strategies ↔ Strategy, Studies ↔ Study, Processes ↔ Process).

As stated by Liben-Nowell and Kleinberg (2007), it is impractical to seek predictions for edges whose source and destination nodes are not present both in the training and test intervals. In light of this and due to the aforementioned filtering processes (see "Keyword extraction" and "Text pre-processing" sections), a list of relevant keywords (as nodes for three datasets) was selected (see Table 2 for statistics).

Keyword co-occurrence network

Co-occurrence is the measurement that attempts to identify objects that tend to occur together. A keyword co-occurrence network (or simply a keyword network) is a co-word network in which the extracted author selected keywords act as nodes and their co-appearance within an article denote an edge in the network. Subsequent co-appearances in articles of particular keyword pairs increase the respective edge weight.

In the context of link prediction, the training phase for the first 5 years of each dataset and the final year as the test phase were defined. Consequently, the year range of 2010–2014 was used as the training phase for both the PMGT and TM datasets and the year 2015 was used as the test phase. In OBS, the period of 2008–2012 was defined as the training phase and the year 2013 as the test phase. Thus, the co-word networks for both the

Table 2 Node and edge statistics of co-word networks for the three datasets

Dataset name	E_T	E_{T+1}	Positive edges	Negative Edges	Nodes
OBS	100,623	34,382	18,543	10298 K	4545
PMGT	18,199	3885	2830	3234 K	2608
TM	2440	518	321	207 K	679

E_T represents the number of edges in the training phase and E_{T+1} represents the number of edges in test phase. Positive edges represent new co-occurrences appeared in the test phase but not in the training phase. Negative edges represent all other potential non-existent edges in both training and test phase

training and test phases were built separately for the three datasets. G_T (V_T, E_T) is used as the notation for the training network and G_{T+1} (V_T, E_{T+1}) as the notation for the test network for the rest of the study, where V_T represents the set of keywords appearing in both phases and E represents their co-occurrences in articles.

Supervised link prediction

The approach of Al Hasan et al. (2006) was adopted to set up supervised link prediction. In this experimental domain, the interaction between nodes (i.e., keywords) was defined as their co-appearances in research articles. As stated above, for the sake of link prediction, the training phase was defined as T and the test phase as $T + 1$. The classification dataset was prepared by choosing pairs of keywords from V_T that appeared together in articles during $T + 1$, but not in T. Each such pair could have either a positive or negative label [$I_{(u,v)} = 1/0$] depending on its presence in the test phase $T + 1$. Co-occurrence of keyword pairs within an article in $T + 1$ represented a positive instance in the classification dataset; for example, in the OBS dataset, there were 4545 keywords, 100,623 edges in T and 34,402 edges in $T + 1$. Of the 34,402 edges, 18,563 unique edges were found in $T + 1$ that were non-existent in T and represented positive instances in the classification dataset. The approximate number of links with negative class label in this case was 10298 K. Table 2 summaries the statistics of nodes and edges for both training and test phases along with the number of positive and negative class link instances for the three datasets.

Supervised methods of link prediction problem need to predict possible future links by successfully discriminating links with positive label [l(u,v) = 1] from links with negative label [l(u,v) = 0] within a classification dataset. Thus, supervised link prediction turns into a binary classification task that involves learning positive and negative instances by exploiting interesting features describing them. This study employed network intrinsic features (see "Link prediction and topological features" section) of the node pairs to describe individual data point in the classification dataset. Network structural features were used to generate values of topological similarity score, $score_i$ (u, v)—the ith feature for each instance in the dataset. Linkpred (Guns 2014) software was used to generate $score_i$ (u, v).

Similar to social networks, in the context of link prediction over co-word networks, the test network G_{T+1} was extremely sparse where only a small segment of nodes formed associations in $T + 1$. Consequently, a large number of potential links were non-existent. The surging number of negative class instances (compared to the number of positive class instances) resulted large class skewness in the classification dataset for the supervised link prediction. It is evident from Table 2 that the number of links with negative class label far outweighed the number of links with positive class label across each of the three datasets. This represents the most infamous problem of supervised link prediction; that is, the class imbalance problem. Lichtenwalter et al. (2010) asymptotically imposed a higher limit on the ratio between the positive and negative samples in a classification dataset. This study initially restricted its workload ratio of positive and negative class labels to 1:10 as followed by (Wang et al. 2007).

Static and evolving (dynamic) networks

This study used well-known topological metrics to analyse the temporal evolutions of co-word networks. The training phase of the three datasets was split into five different time windows ($n = 5$) with each window denoting a year. The co-word network G_T in the

training phase was split into smaller networks for each time window $t = 1, 2,\ldots, n$. The evolution of G_T in relation to different time window t was defined as $G_T = [G_1^T, G_2^T,\ldots,G_n^T]$. These individual splits of training network were referred to as ***subgraphs***. Next, a time series of topological similarity values was built in relation to each of these subgraphs for each of the non-connected node pairs from the classification datasets. Finally, three forecasting models (see "Time series and forecasting models" section) were applied to these time series to predict their final values. These values were used as input for the classification datasets to describe each instance. In relation to the three forecasting models (i.e., ARIMA, exponential smoothing and the random walk), three classification datasets were produced. For the sake of simplicity, these three datasets are named G_{ARIMA}, G_{ETS} and G_{RWF}, respectively in this study. An R forecast package (Hyndman and Khandakar

Table 3 Co-word network evolution with respect to nodes, edges, new keywords and different link formation mechanisms

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
TM						
Keywords (K)	311	192	238	259	295	257
New keywords (K_{new})	–	96	103	66	69	26
Old keywords (K_{old})	–	96	135	193	226	231
Edges (E)	743	383	500	583	690	601
Edges among K_{new} (E_{new})	–	57	60	14	19	6
Edges among K_{new} and K_{old} (E_{common})	–	193	244	169	181	77
New edges among K_{old} (E_{old})	–	143	217	419	490	518
PMGT						
Keywords (K)	1170	1207	1135	1237	1243	1185
New keywords (K_{new})	–	581	333	277	184	59
Old keywords (K_{old})	–	626	802	960	1059	1126
Edges (E)	4055	4145	4136	4276	4347	4195
Edges among K_{new} (E_{new})	–	400	118	101	54	10
Edges among K_{new} and K_{old} (E_{common})	–	1872	1218	923	1084	547
New edges among K_{old} (E_{old})	–	2041	2800	3253	3548	3885
OBS						
Keywords (K)	3131	3249	3548	3697	3847	3813
New keywords (K_{new})	–	834	401	136	41	2
Old keywords (K_{old})	–	2415	3147	3561	3806	3811
Edges (E)	21,948	23,879	27,005	32,999	35,628	34,402
Edges among K_{new} (E_{new})	–	185	63	7	0	0
Edges among K_{new} and K_{old} (E_{common})	–	4053	2208	908	339	20
New edges among K_{old} (E_{old})	–	19,641	24,734	32,084	35,289	34,382

E_{new} = number of links formed within the new keywords arrived each year, E_{common} = number of links formed between a new keyword in each year and an old keyword from the previous year(s), and E_{old} = number of links among the old keywords arrived in the preceding year(s) of a particular year. K = total number of keywords in a particular year; K_{new} = number of new keywords in each year and K_{old} = number of old keywords arrived before a particular year, E denotes the total number of links in a particular year where $E = E_{\text{new}} \cup E_{\text{common}} \cup E_{\text{old}}$. Year 1–5, denotes the training phase and year 6 denotes the test phase for each dataset

2008) was used for the forecasting purposes. This package automatically considers all variations before selecting the best ARIMA and exponential smoothing model for the data. Three datasets built in this way represent the classification datasets considering dynamic networks. A fourth classification dataset was also built in which topological similarity scores of non-connected node pairs were computed using the static snapshot of G_T. As it was built on a static version of the network, this dataset is named G_{static} for the rest of this study.

Classification algorithms

An abundance of classification algorithms exist in supervised learning. The performance of these algorithms may vary depending on the datasets, feature values and associated patterns present in the dataset. This study used simple logistic regression, Naïve Bayes, *K*-nearest neighbours (KNN), Random Forest and Bagging algorithms. The latter three algorithms use ensemble based methods. More sophisticated classification algorithms such as support vector machine and neural network algorithms can also be used in future study. WEKA—the well-known machine learning software, was used for the classification purpose (Hall et al. 2009). The performances of these classifiers were then compared using different performance measurement metrics.

Results

This section describes the results of this study. First, the co-word networks of the three main datasets were analysed to understand their evolutionary nature. Second, the results of the forecasting models were explored. Third, the performances of the classification methods were analysed.

Co-word network evolution

Table 3 sets out the analysis of the evolving nature of the co-word networks. This table shows the growth and shrinkage of the co-word networks in relation to the nodes (i.e., keywords), edges (i.e., keyword co-occurrences), new keyword arrivals and the different mechanisms for forming links among keyword pairs. The nature of associations among keywords revealed three main variations of link formations: (1) a new link can be established between a pair of new keywords that arrived in each year; (2) a recently arrived keyword can form a link with an old keyword introduced in the previous year(s); and (3) a new link can emerge between two old keywords from the pool of old keywords that are yet to be linked. The term 'old keyword' K_{old} represents the set of keywords introduced in the preceding year(s) of a particular year; for example, in PMGT, in the year 2014, the total number of old keywords was 1059. These keywords arrived within the period of 2010–2013 (inclusive). In Table 3, the numbers of edges formed according to the aforementioned three different mechanisms were named E_{new}, E_{common} and E_{old}, respectively. The latter mechanism, denoted by E_{old}, dominated the other two across the three datasets as the networks evolved over time. E_{old} contained two types of links: (1) new links among enduring keywords from the past; and (3) repetitive links that had already appeared in the preceding years. The number of links among the new keywords in each year was trivial and it was evident that majority of the new arrivals predominantly formed links with old

keywords (i.e., E_{common} in Table 3). The introduction of more new keywords K_{new} does not necessarily generate more new hypotheses for scientific research. The new premises of scientific researches emerged from new combinations of existing scientific concepts. Research growths relied on efficiently synthesising multi-disciplinary and prevalent concepts. In relation to medical research (e.g., obesity), both the number of new concepts and associations among them (emerging each year) did not attract any additional attention until they became more familiar and popular within the corresponding research community. As new concepts were introduced and became better acquainted in subsequent years, they tended to gain more attention from the cross-disciplinary scientific research community. Three datasets showed that with the temporal evolution of co-word networks, the number of new keywords declined, but the use of old keywords from the previous year(s) was widely extended (see Table 3). In the OBS dataset, the number of new keyword arrivals declined to zero in the final year (i.e., 2013). Simultaneously, in the same year, the number of co-occurrences among old keywords almost doubled compared to the year 2009. This phenomenon shows the evolving nature of scientific research.

The arrival of new keywords is indeterminate and unpredictable, but a reasonable amount of new keywords found to have formed links with the popular keywords from the set of old keywords. As stated above, co-word networks conform to the properties of complex networks (i.e., preferential attachment) where rich get richer. They also maintain a power law degree distribution. Further, it also demonstrates that co-word networks are highly assortative or dis-assortative. The assortativity co-efficient measures the level of homophily (tendency to bond with similar counterpart) of the network based on node attributes. Generally it is measured in regards to the number of direct neighbours (degree) of nodes in the network. A network is highly assortative when high degree nodes, on average, connect to other high degree nodes and similarly, low degree nodes, on average, show tendency to connect with low degree nodes. On the contrary, in a disassortative

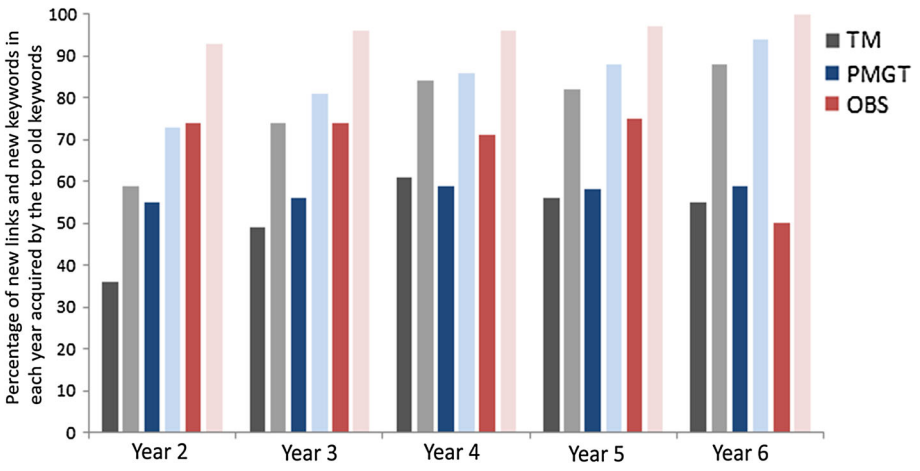

Fig. 1 The relative percentage of new links from E_{common} and new keywords in K_{new}, arriving in each year, acquired by the top old keywords from K_{old} arrived in the preceding intervals. The top old keywords here denote the top 30 % popular keywords, high in degree centrality, from the preceding year(s) before each interval. Dark colored bar represents the percentage of links and light colored bar represents the percentage of new keywords acquired by the popular old keywords in each year. The year-wise measurement exclude the first year of each dataset

network (negative assortativity), nodes with high degree tend to form association with the low degree nodes (Noldus and Van Mieghem 2015). In this study, we observed disassortative co-word networks. For example, OBS has assortativity rate −0.122 and −0.118 recorded in 2008 and 2013 respectively. TM has assortativity rate −0.132 and −0.229 recorded in 2010 and 2015 respectively. Finally, in PMGT, the disassortative index was recorded −0.086 in 2010 and −0.095 in 2015. We observed that the disassortativity index was decreasing in OBS. Later, we will find in this study that the number of new keyword appearance decreases temporally prompting the existing keywords to form associations among them. However, in PMGT and TM disassortativity index was increasing that signifies the degree heterogeneity in forming association among keywords.

This study attempted to analyse the trend and pattern of forming associations among new keywords (arriving each year) in relation to the top 30 % of old keywords from the previous year(s). These top keywords were selected based on their degree centrality with respect to their corresponding co-word network in preceding year(s) of a particular year. Next, this study sought to determine the effect of these top old keywords in persuading new keywords in each year to form links. Figure 1 shows the relative percentage of acquired links (acquired by these top keywords) of the total links in E_{common} in each year excluding the first. This dark bars in the figure for each dataset represent the percentage of new links. The light coloured bars in the figure represent the percentage of new keywords from K_{new} in each year that participated in these acquired links. It is apparent from Fig. 1 that as the co-word networks evolved temporally, the rich, old keywords from the preceding year(s) remained rich, as the majority of the new keywords tended to form associations with these rich, old keywords. Similar results were found across all three datasets; however, it is evident that not all links in E_{common} emerged from these top degree nodes. Some

Table 4 Number of new and unique links (E_{unique}), appearing in each year, from pairs of old keywords in K_{old} (see Table 3) with different length of geodesic distance

	Year 2	Year 3	Year 4	Year 5	Year 6
TM					
Unique edges among old keywords, E_{unique}	93	154	283	310	321
Node pairs within GD = 2	47	83	175	210	218
Node pairs within GD > 2	33	62	96	85	99
Node pairs with no path	13	9	12	15	4
PMGT					
Unique edges among old keywords, E_{unique}	1611	2135	2475	2733	2895
Node pairs within GD = 2	807	1291	1641	1890	2154
Node pairs within GD > 2	706	780	813	816	740
Node pairs with no path	98	64	21	27	1
OBS					
Unique edges among old keywords, E_{unique}	14,130	16,260	19,889	20,633	18,543
Node pairs within GD = 2	12,239	14,807	18,775	19,977	18,241
Node pairs within GD > 2	1857	1440	1114	656	302
Node pairs with no path	34	13	0	0	0

GD represents the length of geodesic distance between keyword pairs in the co-word network of the previous years(s)

new links also developed between the new keywords and unpopular keywords from the previous year(s); for example, in the TM dataset, the co-word network in 2012 (i.e., Year 3) had 244 edges (marked as E_{common} in Table 3). In 120 of these 244 edges (i.e., 49 %), one end contained a new keyword that arrived in 2012 and the other end contained a keyword from the top 30 % of high degree keywords in 2010–2011. Further, 76 of the 103 new keywords (approximately 74 %), arrived in 2012 participated in these 120 edges. Figure 1 also shows that the majority of new keywords (i.e., on average 60–90 %) tended to associate with the top nodes from the preceding year(s) across the three datasets; however, not all links with old and new keywords emerged from these top nodes.

Although new keywords arrived each year; however, most of the new links developed among the prevailing old keywords. This aspect of co-word network evolution suggests that existing and commonly familiarised concepts are widely employed to validate new hypotheses in scientific research. In some instances, the relative percentage of new associations among old keywords exceeded 90 % (E_{old} in Table 3). Thus, this study next sought to explore the effect of distance between pairs of these old keywords before the formation of links between themselves. Specifically, this study attempts to analyse whether the old keywords tended to form associations with remote counterparts or other old keywords from close neighbourhoods. If two keywords are more than two 'hops' away in their shortest path, the neighbourhood and the local topological metrics are inconsiderable. Conversely, if two keywords are within two hops of each other than the neighbourhood effect is considered higher in forming links. In such circumstances, local topological similarity metrics (rather than global topological similarity metrics) provide a better prediction of new links.

Table 4 sets out the number of links between keywords within different geodesic distances for each year for three datasets. For this purpose, only the links between old keywords that were unique to a particular year were considered and repeating links were ignored; for example, in the PMGT dataset of 2012, 2800 links (E_{old} in Table 3) emerged from 802 old keywords (K_{old}) from the preceding duration 2010–2011. Of these links, 2135 edges were new and unique links that had been absent in 2010–2011. The remaining 665 links represented the duplicate or repeating links of the previous year(s). Of these 2135 links, 1291 links had keywords those were two steps away from each other in the 2010–2011 network, 780 links had keyword pairs those were three hops away in their geodesic distance and 64 links emerged among keywords for which no path existed (see Table 4). Evidently, we observed that at least 50 % of the new links among the old keywords occurred between two nodes those were previously two steps away from each other. This percentage increased with the temporal evolution of co-word networks; however, there were some node pairs with no geodesic path existed in the network. The non-existence of geodesic path between keyword pairs may have occurred because all the keywords within a particular domain were not considered; however, this would have been practically infeasible in the context of this study and would be a resource intensive task.

This study also attempted to explore the aforementioned two mechanisms of link formations with the help of co-word networks from datasets (see Figs. 2, 3). This excludes the mechanism of link formation between new keyword pairs. Figures 2a, b present two snapshots of OBS subgraphs in 2008 and 2010 using three target keywords; that is, 'Vascular Disease', 'Dementia' and 'Alzheimer's Disease'. Figures 2c, d present two snapshots of TM subgraphs in two intervals; 2010–2013 and 2014–2015, using four target keywords; that is, 'Big Data', 'Topic Detection', 'Topic Analysis' and 'Topic Mining'. These keywords are marked blue in four figures. The green coloured keywords from Fig. 2b, d represent the corresponding keywords coloured red in Fig. 2a, c. The red

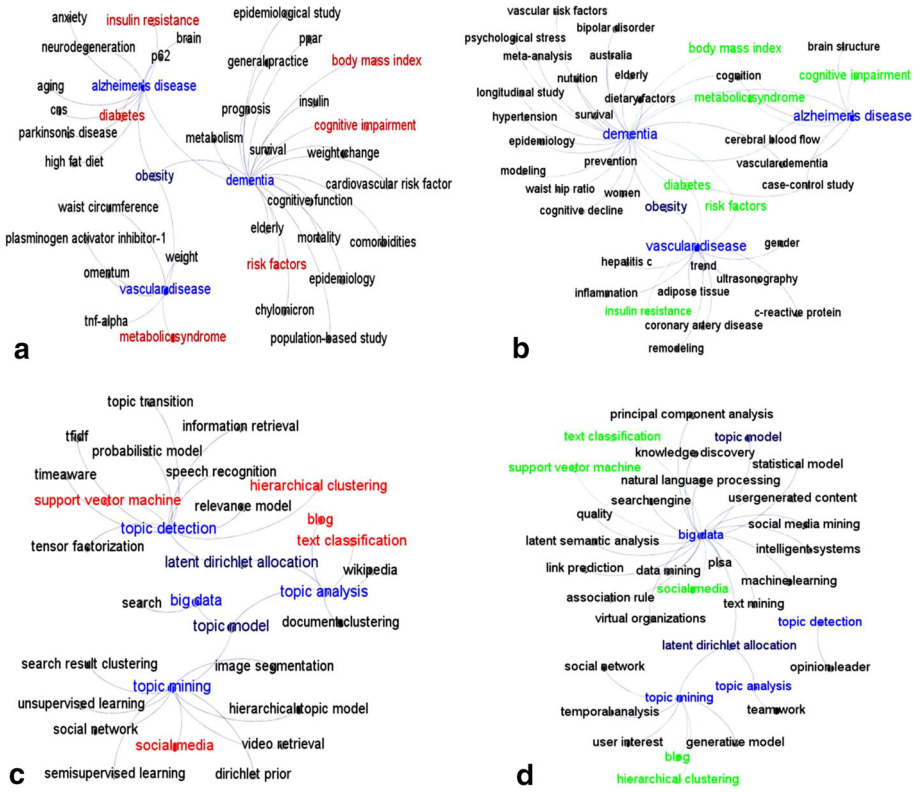

Fig. 2 Snapshots of co-word networks from OBS dataset in the year **a** 2008 and **b** 2010 and TM dataset in the duration **c** 2010–2013 and **d** 2014–2015. Each network in **a** and **b** represents a subnetwork snapshot for three keywords namely 'Vascular Disease', 'Dementia' and 'Alzheimer's Disease' in the respective year. Similarly, each network in **c** and **d** represent a subnetwork snapshot for three keywords namely 'Topic Mining', 'Topic Analysis', 'Topic Detection' and 'Big data' in the respective duration. The *red coloured* keywords from the networks on the *left* are the corresponding *green coloured* keywords in the networks on the *right*. The *navy coloured* keywords are the top old keywords in both dataset with high degree centrality

coloured keywords formed new links with neighbours of neighbours. Although, new links evolved at different path length, however, these red keywords acquire new connections with target keywords that are linked by the important mediator keywords (coloured navy) like 'Obesity' in the OBS domain and 'Latent Dirichlet Allocation' or 'Topic Model' in the TM domain; for example, 'Cognitive Impairment' previously associated with 'Dementia' in Fig. 2a formed a new link with 'Alzheimer's Disease' in Fig. 2b, a direct neighbour of 'Dementia' in Fig. 2a. Most interestingly (not represented in the figures), we found that majority of the keywords that formed new links in Fig. 2b, d with the three target keywords, were formerly associated with the navy coloured keywords, common neighbour to the three target keywords, or one of the top old keyword with high centrality measures.

The well-known concept of 'Preferential Attachment' is observable in co-word network evolution in Fig. 3a, b. These figures provide two snapshots of the co-word networks for the PMGT dataset in 2011 and 2014. These two figures represent the top nodes in the two subgraphs in relation to their degree centrality. The colour code represents the sequential order of degrees; blue represents nodes with the highest number of connections, black, the

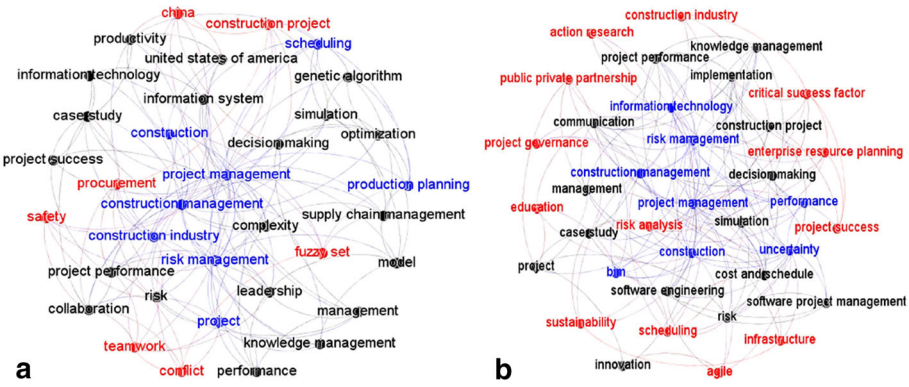

Fig. 3 Two subnetworks with top ($n \cong 35$) keywords with high degree centrality in PMGT dataset in the year **a** 2011 and **b** 2014. The *colour code* (i.e., *blue*, *black*, *red*) represents the sequential order of degree centrality of keywords in the respective co-word network with blue coloured keywords having highest number of connections in each year

second highest and red being the third highest. It is evident that five of the top eight keywords (coloured blue) match both intervals (i.e., 2011 and 2014). Other keywords have greater connections with temporal evolution; however, a relative percentage of the top degree nodes from Fig. 3b can be found among other top nodes in Fig. 3a (e.g., for the keywords 'Performance' and 'Construction Industry'). These figures establish the second mechanism of link formation (described above); that is, most new keywords form associations with the top nodes of the preceding intervals.

Time series and forecasting results

In building the time series dataset for forecasting purpose, the topological similarity values for both positively and negatively labelled links were first calculated. If two keywords, incidental to an edge, co-appeared together in any subgraph (i.e., G_1^T, G_2^T,…, G_n^T) of the training phase then a value for their topological similarity (or zero otherwise) was obtained. It was observed that in PMGT and TM the majority of the keywords did not appear as pairs in all five subgraphs of the training phase. Therefore, a link, either with a positive or negative label, present in at least one subgraph of five and the topological similarity values for the nodes incidental to it was greater than zero [i.e., score $(u, v) > 0$], was considered to be a member of the time series data for forecasting purpose; for example, two keywords from the TM dataset (i.e., 'Latent Dirichlet Allocation' and 'Climate Change') were found forming a link in $T + 1$, denoting a positive instance of the classification dataset. Conversely, these two keywords did not appear together within any subgraph in T. As these two keywords did not form an edge within any subgraph of the training phase, their temporal information of topological evolution was unknown and thus not considered in the time series. However, keywords such as 'Construction Industry' and 'Turnover' in the PMGT dataset formed a link both in the 2014 (training phase) and the 2015 (test phase). As both keywords co-appeared together within one of the subgraphs in T, this link was considered as a prospective link for time series construction. This link had a topological similarity score for the year 2014 and zero for rest of the four subgraphs. Table 5 sets out the total number of links with positive and negative labels in the

forecasting datasets for OBS, PMGT and TM. Figure 4a–c show the different estimates from the three forecasting methods (i.e., ARIMA, exponential smoothing and random walk) for three different time series. These time series were built on the degree of connections in each year, acquired by the top three keywords in the study (i.e., 'Obesity', 'Project Management' and 'Latent Dirichlet Allocation'). The dotted lines in the figures represent the lower bound of the forecasted predictions with 80 % confidence interval in relation to the three methods used in this study.

Classification performance

In supervised link prediction, each future potential non-connected node pair has either a positive label or negative label. Depending of their appearances in the test phase, if a potential node pair truly appears in the test phase, it is labelled as positive class instance in the classification dataset and a negative class instance otherwise (Al Hasan et al. 2006). As stated above, there were four classification datasets (i.e., G_{ARIMA}, G_{ETS}, G_{RWF} and G_{static}) with both positive and negative classes of links. The first three datasets consider topological evolution in dynamic networks and the fourth dataset considers a static network. Pairs of keywords (representing positive and negative instances in the datasets) were chosen at random from a list of qualifying links. The feature vectors were computed considering the local, global and aggregated metrics for each keyword pair incidental to those links. Next, five well-known classifiers (see "Classification algorithms" section) were used and their performances were measured to classify the positive and negative examples of links. For all classifiers, a tenfold cross-validation and the mean scores were used to determine the accuracy of the results. Tables 6, 7 and 8 compare the performance of different classifiers on G_{ARIMA}, G_{ETS}, G_{RWF} and G_{static} using the TM, PMGT and OBS datasets.

Earlier in this study, it was observed that future associations among keywords evolve not only from similar and adjacent nodes, but also from dissimilar and distant nodes (Fig. 2). Consequently, both local and global topological similarity features can be used to model co-word network evolution. These features can be engineered to distinguish between positive and negative classes of links. The tables show that network structural metrics can be used as an important set of features in supervised link prediction in relation to co-word networks. Considering different classifiers, ensemble classifiers were observed to perform better than basic linear classifiers such as logistic regression. Instead of finding the optimum value for the parameters of ensemble classifiers, the default parameters for all datasets were run. This parameter optimisation task should be the subject of future exploration.

In the random forest analysis, the number of trees was set as 150 with a depth level of five. In the KNN analysis, the number of neighbours (K) was set to five. While in the

Table 5 Total number of positive and negative edges considered as part of the time series for forecasting purpose and build the final classification datasets

Dataset name	Positive edges	Negative edges
OBS	13,032	106,376
PMGT	737	761
TM	77	144

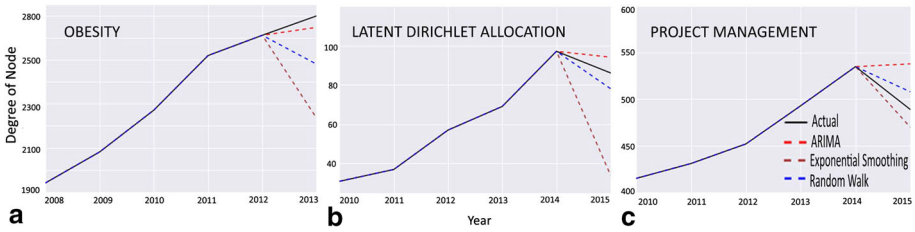

Fig. 4 Forecasted values of nodal degrees predicted by the three forecasting models for three important keywords in this study, **a** Obesity, **b** Latent Dirichlet Allocation and **c** Project Management. The forecasted values represent the lower bound of the range of possible values at 80 % confidence interval. The *dotted lines* represent the prediction by three different forecasting models i.e., ARIMA (*red*), Exponential Smoothing (*maroon*) and Random Walk (*blue*)

Table 6 Performance of different classification algorithms for TM dataset

Classifier	Accuracy	AUC	Precision	Recall	*F*-score
G_{ARIMA}					
Logistic Regression	0.74	0.716	0.625	0.277	0.384
Naive Bayes	0.77	0.692	0.700	0.389	0.500
Bagging	0.78	0.784	0.579	0.611	0.595
Random Forest	0.78	0.821	0.563	0.504	0.529
K-Nearest Neighbour	0.76	0.726	0.458	0.611	0.524
G_{ETS}					
Logistic Regression	0.76	0.747	0.503	0.421	0.457
Naive Bayes	0.77	0.789	0.733	0.579	0.647
Bagging	0.78	0.839	0.667	0.737	0.700
Random Forest	0.81	0.849	0.654	0.684	0.661
K-Nearest Neighbour	0.75	0.727	0.505	0.579	0.537
G_{RWF}					
Logistic Regression	0.72	0.771	0.667	0.202	0.308
Naive Bayes	0.77	0.778	0.688	0.550	0.611
Bagging	0.73	0.778	0.714	0.500	0.588
Random Forest	0.73	0.833	0.786	0.550	0.647
K-Nearest Neighbour	0.73	0.777	0.600	0.600	0.600
G_{Static}					
Logistic Regression	0.74	0.791	0.786	0.500	0.611
Naive Bayes	0.79	0.824	0.813	0.591	0.684
Bagging	0.81	0.826	0.882	0.682	0.769
Random Forest	0.83	0.848	0.727	0.727	0.727
K-Nearest Neighbour	0.78	0.814	0.608	0.636	0.622

bagging analysis, a decision tree was used as a base classifier. Of the three ensemble classifiers, the bagging analysis of the TM and OBS datasets and the random forest analysis of the PMGT dataset showed superior performances. In relation to the linear classifiers,

Table 7 Performance of different classification algorithms for PMGT dataset

Classifier	Accuracy	AUC	Precision	Recall	*F*-score
G_{ARIMA}					
Logistic Regression	0.67	0.715	0.724	0.549	0.624
Naive Bayes	0.64	0.720	0.821	0.348	0.489
Bagging	0.66	0.738	0.684	0.638	0.66
Random Forest	0.70	0.764	0.673	0.679	0.676
K-Nearest Neighbour	0.65	0.656	0.602	0.607	0.604
G_{ETS}					
Logistic Regression	0.67	0.727	0.686	0.553	0.612
Naive Bayes	0.64	0.713	0.767	0.364	0.494
Bagging	0.66	0.756	0.676	0.645	0.667
Random Forest	0.71	0.771	0.671	0.622	0.646
K-Nearest Neighbour	0.66	0.703	0.652	0.622	0.637
G_{RWF}					
Logistic Regression	0.67	0.765	0.733	0.561	0.635
Naive Bayes	0.65	0.754	0.807	0.400	0.535
Bagging	0.69	0.778	0.704	0.691	0.697
Random Forest	0.72	0.816	0.747	0.717	0.732
K-Nearest Neighbour	0.64	0.731	0.643	0.696	0.668
G_{Static}					
Logistic Regression	0.68	0.748	0.741	0.567	0.642
Naive Bayes	0.63	0.736	0.792	0.369	0.503
Bagging	0.62	0.677	0.638	0.576	0.605
Random Forest	0.69	0.733	0.659	0.696	0.677
K-Nearest Neighbour	0.63	0.652	0.605	0.585	0.595

logistic regression outweighed Naïve Bayes in PMGT. Despite its high performance in document classification, Naïve Bayes had poor accuracy in the OBS and PMGT datasets. The inferior performance by Naïve Bayes is subject to inter-dependencies among different features. This algorithm performs well when features are independent and if dependences are distributed evenly in classes; however, it is incompetent at handling feature interactions. The poor logistic regression performance was due to the linearly inseparable feature values. Despite having a single parameter, bagging is prone to over fitting and computationally expensive, as it considers all available features to split a node in decision trees. Conversely, random forest, a special case of bagging, randomly considers only a subset of the best features of those available. Thus, its performance was superior to that of bagging in some cases.

In a supervised classification problem, evaluation metrics can be broadly categorised into two main classes: (1) fixed threshold metrics such as accuracy, precision and recall; and (2) *k*-equivalents and threshold curves such as a receiver operating characteristics (ROC) curve, a precision-recall (*P–R*) curve and the area under the ROC curve (AUC) (Yang et al. 2015). Precision has been defined as the proportion of true positive predictions

Table 8 Performance of different classification algorithms for OBS dataset

Classifier	Accuracy	AUC	Precision	Recall	*F*-score
G_{ARIMA}					
Logistic Regression	0.91	0.772	0.716	0.171	0.276
Naive Bayes	0.89	0.794	0.461	0.383	0.419
Bagging	0.97	0.954	0.933	0.756	0.835
Random Forest	0.94	0.917	0.936	0.509	0.659
K-Nearest Neighbour	0.96	0.885	0.923	0.607	0.733
G_{ETS}					
Logistic Regression	0.89	0.783	0.560	0.086	0.149
Naive Bayes	0.82	0.811	0.324	0.643	0.431
Bagging	0.95	0.923	0.887	0.631	0.738
Random Forest	0.93	0.902	0.968	0.336	0.499
K-Nearest Neighbour	0.94	0.853	0.833	0.504	0.628
G_{RWF}					
Logistic Regression	0.89	0.779	0.577	0.103	0.175
Naive Bayes	0.88	0.824	0.459	0.437	0.448
Bagging	0.95	0.934	0.916	0.682	0.782
Random Forest	0.93	0.892	0.946	0.383	0.545
K-Nearest Neighbour	0.95	0.875	0.90	0.57	0.70
G_{Static}					
Logistic Regression	0.91	0.819	0.602	0.145	0.234
Naive Bayes	0.90	0.841	0.387	0.416	0.401
Bagging	0.96	0.916	0.958	0.756	0.845
Random Forest	0.94	0.916	0.984	0.369	0.537
K-Nearest Neighbour	0.91	0.748	0.591	0.213	0.315

of all the positive predictions. Conversely, recall has been defined as the proportion of true positive predictions of all true levels. The *F*-score is the harmonic mean of precision and recall. This score is sometimes considered a better performance measure than the accuracy metric, especially when the class populations in the training set are biased. ROC graphs are two-dimensional graphs in which the true positive rate is plotted on the *Y* axis and the false positive rate is plotted on the *X* axis to show the relative trade-offs among the two class values. This graph directly depicts the screening capability of the predictors. Conversely, *P–R* curves, often used in information retrieval, can be used as an alternative to ROC curves for models with a large skew in the class distribution. *P–R* curves can sometimes expose differences between classifiers that are not apparent in the ROC curves. The AUC, an important traditional measure, is used in imbalanced classification problems. It relates the true positive rate and true negative rate of a classifier. AUC is the second most popular metric (after accuracy) used in binary classification. Accuracy only classifies the class label right or wrong; however, AUC quantifies the uncertainty associated with classifiers by introducing a probability value. In relation to a random classifier that has an equal probability of 0.5 in successfully classifying the positive and negative class labels, the more a classifier exceeds this threshold the better it gets. In relation to the tenfold cross-validated

accuracy measure and AUC, higher values in TM and OBS datasets were observed. The worst result was found in the PMGT dataset. To determine the reason for these results, the distributions of feature values for the positive and negative samples were analysed. Both the highest performing (i.e., G_{ARIMA} in OBS) and the lowest performing dataset (i.e., G_{Static} in PMGT) in relation to accuracy and AUC were selected. Figure 5 presents the distribution of positive and negative class density for three important topological features in each dataset according to the Random Forest classifier. Feature importance is discussed further below. The features selected for the highest performing dataset were AdamicAdar, CommonNeighbours and Katz. Conversely, for the low performing dataset, the three most important features were AdamicAdar, Resource Allocation and SimRank. For the sake of the comparison, the feature distribution was normalised so that the areas under both curves were similar. Classifiers can pick patterns where there is significant difference between both class distributions. The overlapping region between the class density of the positive and negative samples in Fig. 5 were the reason for the misclassification. In the lower performing dataset, the overlapping regions were comparably higher.

In relation to the other performance measures, most classifiers were observed to have comparatively higher precision than recall indicating a higher number of false negatives than false positives across most of the dataset. There were also some exceptions where recall was higher than precision; for example, in the G_{ETS} classification dataset, Bagging and KNN in TM and Naïve Bayes in OBS had higher recall denoting the existence of higher false positives. A higher precision, but lower recall value indicates the conservatism of a classifier whereas a higher recall, but lower precision indicates liberalism. These two

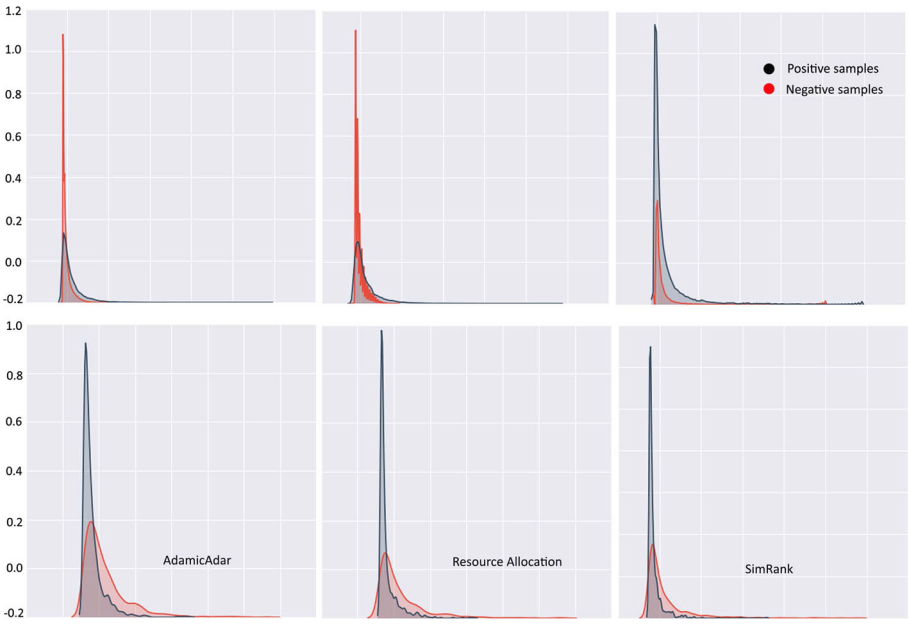

Fig. 5 Positive and negative class density of three topological features in high performing (*top*) and low performing (*bottom*) datasets. The high and low performing datasets are G_{ARIMA} in OBS and G_{Static} in PMGT respectively. Topological features are selected according their importance in Random Forest classifier

measures generally varied inversely with each other; that is, as one increases the other decreases and vice versa. This trend was observed across most of the classification datasets. It is difficult to achieve high recall and high precision simultaneously. The trade-off between precision and recall is important, as a classifier is tuned to stay between the extremes of conservatism and liberalism with the application of labels. Figure 6a–d represent the *P–R* and ROC curves and show the screening capability of the classifiers. Figure 6a, b represent the *P–R* curves of classifiers in the low and high (left to right) performing datasets selected above. Similarly, the bottom row represents the ROC curves. For the low performing dataset, the performances of the classifiers were similar to each other; however, in the high performing dataset, the difference between the linear and ensemble classifiers was precise.

Feature importance

Once it was observed that topological features that depend on co-word network structures are useful in predicting future links among keywords from scientific literature, the study sought to compare these eight features to judge their relative strength in the classification task. Table 9 sets out the quantitative comparisons of these metrics in relations to several algorithms for measuring feature importance. The values in the table denote the rank of importance for each feature demonstrated by the different algorithms. The ranks appear in decreasing order where one denotes the highest rank and eight denotes the lowest rank. To present the rank of feature importance, the best performing classification datasets from TM, PMGT and OBS were selected in relation to accuracy score and AUC (i.e., G_{Static}, G_{RWF} and G_{ARIMA}). The rows with the name of classifiers in the table denote the importance of features according to the respective classifier. The table shows that in TM, global topological metrics such as Katz and SimRank out-performed the local metrics.

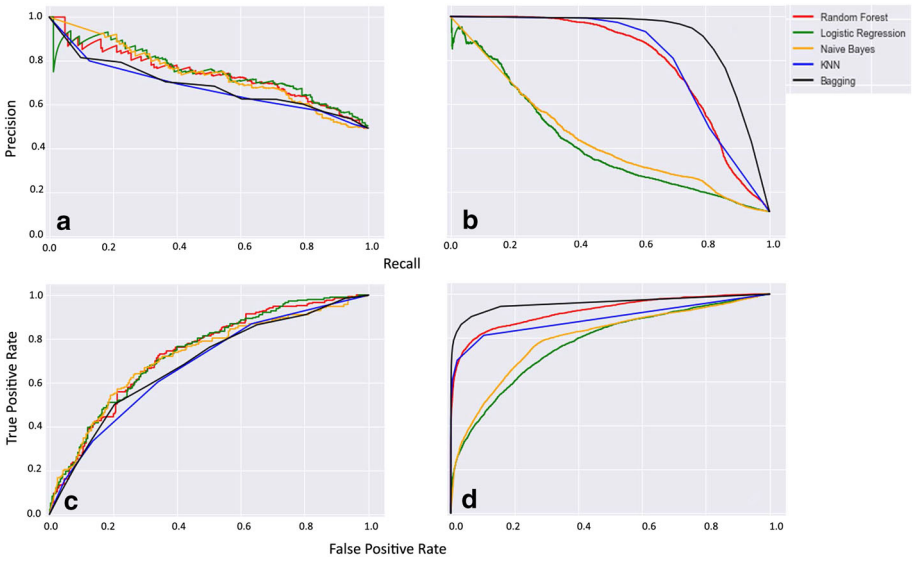

Fig. 6 *P–R* Precision–Recall curves (*top*) and ROC curves (*bottom*) of different classifiers in both low performing (*left*) and high performing (*right*) datasets

Table 9 Rank of different network topological features used in this study for different datasets

	AA	CN	RA	Katz	RPR	SR	DC	CC
G_{Static} in TM								
Chi Square Attribute Value	5	6	7	1	8	4	3	2
Gain Ratio	7	4	8	1	3	5	6	2
Information Gain	5	6	7	1	8	4	3	2
Random Forest	6	7	5	1	8	3	4	2
Bagging	4	6	5	3	7	2	8	1
G_{RWF} in PMGT								
Chi Square Attribute Value	1	6	2	4	7	3	5	8
Gain Ratio	1	6	2	3	8	5	7	4
Information Gain	1	6	2	3	7	4	5	8
Random Forest	1	6	3	4	8	2	5	7
Bagging	1	7	5	3	8	2	6	4
G_{ARIMA} in OBS								
Chi Square Attribute Value	2	1	3	5	8	6	4	7
Gain Ratio	3	4	1	5	8	6	2	7
Information Gain	4	6	1	5	8	2	3	7
Random Forest	2	3	6	1	7	5	4	8
Bagging	2	3	5	1	6	8	4	7

The ranks are ordered in decreasing order with 1 denoting the highest importance. The features are *AA* AdamicAdar, *CN* CommonNeighbours, *RA* Resource Allocation, *RPR* RootedPageRank, *SR* SimRank, *DC* Sum of Degree Centrality, *CC* Sum of Closeness Centrality

Conversely, the local similarity metrics, such as AdamicAdar and ResourceAllocation were useful features for the other datasets. Katz and AdamicAdar appeared to be the most important features for both the linear and ensemble classifiers. This signifies the fact we observed in Fig. 2 where different links were formed at different length among keywords. Although different metrics were found important in different datasets, however, RootedPageRank was found to be the least important feature. It is understandable that this metric is dependent on transition probabilities that vary in relation to network structure and nodal importance. Therefore, time series forecasting method would have little impact in providing meaningful prediction for this type of feature.

Performance improvement in dynamic networks

One of the main benefits of link prediction in dynamic networks, based on time series forecasting rather than a static snapshot of the whole network, is performance improvement in predicting future links among nodes. As observed above, there were better outcomes for link prediction task in relation to performance metrics in some classification datasets built upon dynamic networks (i.e., G_{ARIMA} and G_{RWF}). These datasets were constructed based on the historical evolution of different metrics considered in "Time series and forecasting models" section. Notably, in PMGT and OBS, G_{RWF} and G_{ARIMA} were observed to have higher performance measures than G_{Static}; however, in some respects, G_{Static}, unsurprisingly, out-performed the dynamic datasets. This fact is true when a set of topological

features is considered together and classifiers tune themselves to pick the best patterns from the best performing feature(s). Table 10 compares the three classification datasets from dynamic networks using important features from Table 9 and their respective G_{Static} in OBS and PMGT. In relation to AUC, the table shows that the time-aware information outweighed the static information in supervised link prediction tasks for some network structural metrics.

Discussion

A knowledge graph (alternatively, a network) is a specific kind of knowledge representation technique that uses a semantic network structure comprising textual content (e.g., keywords, concepts) as nodes and causal relationships among nodes as links or edges (Popping 2003). Researchers speculated that there is a reciprocal relationship between causal relationships and statistical co-occurrence patterns (McNorgan et al. 2007; Cheng et al. 2009). Non-trivial co-occurrence pattern among keywords is a complex structure that represents their semantic affinity (Montemurro and Zanette 2013) and relatedness (Schulz et al. 2014). Therefore, co-word networks are considered as knowledge networks comprised of micro knowledge entities (Ding et al. 2013) like keywords or key concepts. Science mapping—an effective tool for science strategy and evaluation that uses the network of links between scholarly documents to understand the structure of science (Börner et al. 2003), provides spatial representations of relationships among various knowledge entities and displays the structural and dynamic aspects of scientific research (Cobo et al. 2011). Further, Noyons and van Raan (1994) noted that with the help of co-word networks, science mapping facilitates the extraction of important features where time series analyses of these features can provide a dynamic view of the structural changes in scientific knowledge.

Table 10 Relative average performances of three classification datasets, built upon dynamic networks with forecasting models, with respect to static network in the training phase

	AUC (OBS)				AUC (PMGT)			
	G_{ARIMA}	G_{ETS}	G_{RWF}	G_{Static}	G_{ARIMA}	G_{ETS}	G_{RWF}	G_{Static}
Random Forest								
AdamicAdar	0.733	0.722	0.721	0.764	0.781	0.753	0.777	0.727
CommonNeighbours	0.709	0.701	0.712	0.753	0.669	0.668	0.686	0.725
Resource Allocation	0.784	0.777	0.778	0.777	0.756	0.768	0.730	0.729
Katz	0.733	0.775	0.781	0.681	0.687	0.664	0.736	0.714
Sum of Degree Centrality	0.752	0.748	0.754	0.729	0.700	0.707	0.674	0.705
Naïve Bayes								
AdamicAdar	0.650	0.644	0.658	0.683	0.726	0.699	0.702	0.712
CommonNeighbours	0.662	0.631	0.650	0.678	0.669	0.668	0.664	0.705
Resource Allocation	0.754	0.745	0.744	0.682	0.724	0.679	0.645	0.689
Katz	0.647	0.742	0.757	0.661	0.692	0.668	0.715	0.705
Sum of Degree Centrality	0.695	0.683	0.708	0.723	0.699	0.664	0.607	0.647

Scientific knowledge changes over time and most of these changes are incremental although some revolutionary and fundamental changes do occur. Since the study of sociology of science (Latour and Woolgar 2013), quantitative analysis of scientific literature at macro-level (disciplinary) shifted increasingly towards micro-analysis (keywords) with its focus on scientific communications (Leydesdorff and Milojević 2015). Formal communication among scientific communities through scholarly publications is found significant in scientific knowledge creation and innovation. Co-word analysis is a form of scholarly communication that extracts thematic concepts of scholarly contents and their linkages directly from the textual contents. Constituents of co-word network such as author selected keywords provide the conceptualization of these contents that authors intend to convey the world. Therefore, co-word network is a useful and objective approach in identifying the dynamics of conceptual structures in various disciplines and socio-cognitive structures of science. It concurrently reveals the patterns of scientific knowledge growth through the collective understanding of scholarly communities. Thus, it can be argued that co-word network is conducive of identifying emerging research topics if it is explored from the aspect of network dynamics. Further, Canals (2005) noted that knowledge diffusion occurs over network structures, as the process of knowledge diffusion involves interactions among networked agents. Therefore, in this study, we attempted to comprehend the dynamics of co-word network in relation to future nodal relationships and link formations mechanisms among keywords that can contribute to an understanding of the knowledge evolution mechanism. Thus the two main topics in this study are co-word networks and link prediction.

Earlier in this study, we mentioned that co-word networks display scale-free and small-world phenomenon. In this study, we have observed that co-word network also conforms to preferential attachment process found in real world complex networks. This demonstrates that co-word networks evolve through a self-reinforcing mechanism. Preferential Attchment also denotes that high degree nodes are very densly connected to each other and the nodes with low degree centrality rarely connects to each other. Therefore, in this study, we also observed that most recent and contemporary keywords form associations with popular keywords within the research domain as co-word network evolves temporally. It signifies the "Rich Club Phenomenon" in terms of co-word network evolution. Concurrently, it raises the degree heterogeneity (deviations from the regular network in terms of degree) of a network where the degree distribution follows a power law. Further, the dis-assortativity (negative assortativity) of co-word networks built in this study denotes the tendency of forming linkages between dissimilar keywords in regards to their degree centrality.

On the other hand, link prediction is a time-evolving model for network analysis problems that directly predicts links in the future based on previous trends to model network dynamics. When dynamics is concerned it is also imperative to include the temporal information in analysis. Thus, this study built and analysed the evolution of co-word networks in relation to both static and time-aware network structural information. It also attempted to perform link prediction task by using this information to predict future associations among keywords/key-concepts extracted from scientific literature. In this study, three scholarly datasets were obtained from Scopus on the topics of 'Obesity', 'Project Management' and 'Topic Model'. Each dataset comprised 6 years of scientific articles and associated author selected keywords. After the necessary text pre-processing and data standardisation tasks, co-word networks of the author provided keywords were constructed based on the co-appearances of keywords in the same article. For the purpose of the link prediction, the range of publication years was divided into two non-overlapping sub-ranges; that is, the training phase and test phase. This study attempted to use different

topological metrics in relation to keyword pairs from the co-word network in the training phase to predict future associations in the test phase. These topological metrics are based on network structure and defined in relation to neighbourhood, ensembles of network path, random walk and centrality measures. These features were used to describe instances of classification datasets comprising both positive and negative classes of links. These classification datasets were then fed into supervised learning models so that the two classes of links could be classified. Supervised link prediction is a binary classification task and may be selected over unsupervised link prediction for two reasons (Lichtenwalter et al. 2010). Firstly, under the supervised method, algorithms are able to capture the interacting relationships among different topological properties and, secondly, supervised approaches are adaptive whereas unsupervised methods are invariant. Additionally, if a classifier is trained using a single unsupervised method, it is capable of outperforming the ranks generated by sorting the scores of the respective method.

To build the classification datasets in this study, topological features of both dynamic and static networks were considered. To emulate the dynamic network perspective, the training network was split into five smaller units with one subnetwork for each year of training phase. Based on these splits, the topological evolutions were determined for each non-connected node pair, either with a positive or negative class label, of the classification dataset. This constructed a time series of topological metrics for each non-connected node pair. Three forecasting models were then used to predict their final value as input to the classification datasets. Thus, three classification datasets were constructed considering the time-aware network evolution information. A final classification dataset was also built that considered the aggregated static version of these splits. Of the three datasets, the PMGT dataset had a similar number of positive and negative samples, whereas the OBS and TM datasets had negative samples out-numbering the positive ones by a certain ratio. Of the many possible variations for the ARIMA and exponential smoothing model, the best variation was identified by the R forecast package. As the time series was short in duration, a random walk method of forecasting was also used. The difference among the three forecasting models used in this study was also noted. Future studies should explore the applicability of other prevailing forecasting methods, including multiple variations of the methods used in this study. These forecasting methods could also be used to predict the future values of the network specific properties that allow emerging trends to be predicted; for example, time series of nodal degrees, citation counts per keywords or trends in author-keyword relationships could be used with forecast models to predict their future values that will enable growth in scientific research across different concepts to be identified.

The result section sought to understand the co-word network evolution. Co-word networks were observed to evolve with the accumulation of new keywords (arrived at each interval) and different types of association among keywords. In this study, three types of links formed among keywords in a co-word network: (i) a link between a pair of new keywords; (ii) a link between a new and old keyword; and (iii) new associations within the old keywords. The old keywords at each interval denote the set of keywords already introduced in the preceding intervals. Some repetitive links may evolve among old keywords at every interval; however, the results showed that new links always emerge from the recombination of old keywords. The number of new links among existing keywords increased with time, especially where most keywords came from similar or closely related domains. Further, the old popular keywords also attracted new keywords and most of the new keywords formed associations with already familiar and popular old keywords. This allowed co-word networks to demonstrate ‘preferential attachment’. Similar to other complex networks, the co-word networks were found to conform to the power law

phenomenon; that is, popular keywords had more connections than the others; however, a few new keywords gained popularity as time evolved. We also observed the impact of geodesic distance between keywords in evolving networks over developing future associations among them. We found that most of the new links among existing keywords emerged between 2-hop distant keywords. This fact suggests the importance of neighbourhood around keywords in co-word network impacting on the formation of new associations. However, some new links emerged between keywords with distance more than 2-hops from each other and simultaneously many keyword pairs at distance two did not associate with each other in future signifying the 'small world' property of co-word network.

In relation to classification, five different classifiers were used. This included a simple linear classifier (logistic regression) and some ensemble classifiers. The ensemble classifiers showed superior performances across the three datasets. The performance of dataset with imbalanced class samples was greater than that of the dataset with balanced class labels. This study also attempted to identify feature importance in relation to different algorithms and individual classifiers. Different classifiers performed in different ways depending on the structural variations of the co-word network and the density of class distributions. In some cases, classifiers preferred the local similarity metrics, while some preferred global metrics. In terms of aggregated feature importance, degree centrality was preferred over closeness centrality in most cases. In majority datasets, RootedPageRank was found as the least important features. This is evident as it relies on random walk in network and produces different results in relation to transition probabilities in different types of networks. Moreover, transition probabilities, depending to nodal importance, vary with network structure. Time series forecasting of this metric could contribute a little towards meaningful predictions. Further, historical information of topological evolution helped to improve link prediction performances. The performance measures of dynamic networks sometimes superseded the measures of the static network; however, in some cases, the static network topological information was also found useful in predicting future associations among keywords.

This study sought to contribute to understandings of co-word network evolution, identify important network structural and topological features along with their temporal evolutionary information to describe instances in classification datasets and use supervised learning models to accurately predict future links among keywords from scientific literature. The methods and results of this study will facilitate to predict the future links between literary concepts. This will also benefit the identification of emerging trends, information retrieval and building associative concept spaces. The future research should seek to identify more generic features other than features used in this study such as, statistical significance of keywords, citation count, the number of authors attracted to keywords, to build predictive science mapping.

References

Abbasi, A., Hossain, L., Uddin, S., & Rasmussen, K. J. R. (2011). Evolutionary dynamics of scientific collaboration networks: Multi-levels and cross-time analysis. *Scientometrics, 89*(2), 687–710.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks, 3*(25), 211–230.

Al Hasan, M., Chaoji, V., Salem, S. & Zaki, M. (2006). Link prediction using supervised learning. In *6th SDM' workshop on link analysis, counter-terrorism and security, Bethesda, Maryland, Society for Industrial and Applied Mathematics*.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (revised ed.). San Francisco, CA: Holden-Day.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Canals, A. (2005). Knowledge diffusion and complex networks: A model of high-tech geographical industrial clusters. In *Proceedings of the 6th European conference on organizational knowledge, learning, and capabilities* (pp. 1–25). Boston, MA.

Cheng, X., Miao, D., & Wang, L. (2009). A statistics-based semantic relation analysis approach for document clustering. In P. Witold, M. Duoqian, S. Dominik, P. Georg, H. Qinghua, & W. Ruizhi (Eds.), *Rough sets and knowledge technology* (pp. 332–342). Shanghai: Springer International Publishing.

Choi, J., Yi, S., & Lee, K. C. (2011). Analysis of keyword networks in MIS research and implications for predicting knowledge evolution. *Information and Management, 48*(8), 371–381.

Chung, F., & Zhao, W. (2010). Pagerank and random walks on graphs. In G. O. H. Katona, A. Schrijver, T. Szonyi, & G. Sagi (Eds.), *Fete of combinatorics and computer science* (pp. 43–62). Berlin: Springer.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402.

Davis, D., Lichtenwalter, R. & Chawla, N. V. (2011). Multi-relational link prediction in heterogeneous information networks. In *Proceedings of the 2011 international conference on advances in social networks analysis and mining*, IEEE Computer Society.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting, 22*(3), 443–473.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PLoS ONE, 8*(8), e71416.

Elsevier. (1880). *Scopus*. Amsterdam: Elsevier B. V.

Güneş, İ., Gündüz-Öğüdücü, Ş., & Çataltepe, Z. (2016). Link prediction using time series of neighborhood-based node similarity scores. *Data Mining and Knowledge Discovery, 30*(1), 147–180.

Guns, R. (2014). Link prediction. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 35–56). Cham: Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133.

Huang, Z., & Lin, D. K. (2009). The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing, 21*(2), 286–303.

Huang, Z., & Zeng, D. D. (2006). A link prediction approach to anomalous email detection. In *IEEE international conference on systems, man and cybernetics*.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3), 1–22.

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: The state space approach*. Berlin: Springer.

Jeh, G., & Widom, J. (2002). SimRank: A measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton, Alberta, Canada, Association for Computing Machinery.

Kastrin, A., Rindflesch, T. C., & Hristovski, D. (2014). Link prediction on the semantic MEDLINE network. In S. Džeroski, P. Panov, D. Kocev, & L. Todorovski (Eds.), *Discovery science* (Vol. 8777, pp. 135–143). Bled: Springer International Publishing.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.

Khan, A., Choudhury, N., Uddin, S., Hossain, L., & Baur, L. (2016). Longitudinal trends in global obesity research and collaboration: A review using bibliometric metadata. *Obesity Reviews, 17*(4), 377–384.

Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S., & Phelps, D. J. (2004). A survey of emerging trend detection in textual data mining. In M. W. Berry (Ed.), *Survey of text mining: Clustering, classification, and retrieval* (Vol. 1, pp. 185–224). New York: Springer.

Latour, B., & Woolgar, S. (2013). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6), 481–497.

Leydesdorff, L. (1996). Scientometrics and science studies: From Words and co-words to information and probabilistic entropy. *Journal of the International Society for Scientometrics and Informetrics, 2*, 33–39.

Leydesdorff, L. (2002). Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports. *Scientometrics, 53*(1), 131–159.

Leydesdorff, L., & Milojević, S. (2015). Scientometrics. In D. J. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (pp. 322–327). Oxford: Elsevier.

Li, X., Du, N., Li, H., Li, K., Gao, J. & Zhang, A. (2014). A deep learning approach to link prediction in dynamic networks. In *SIAM international conference on data mining*, Philadelphia, USA, Society of Industrial & Applied Mathematics.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031.

Lichtenwalter, R. N., Lussier, J. T. & Chawla, N. V. (2010). New perspectives and methods in link prediction. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM*.

McNorgan, C., Kotack, R. A., Meehan, D. C., & McRae, K. (2007). Feature-feature causal relations and statistical co-occurrences in object concepts. *Memory and Cognition, 35*(3), 418–431.

Montemurro, M. A., & Zanette, D. H. (2013). Keywords and co-occurrence patterns in the Voynich Manuscript: An information-theoretic analysis. *PLoS ONE, 8*(6), e66344.

Newman, M. (2001). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 25102.

Noldus, R., & Van Mieghem, P. (2015). Assortativity in complex networks. *Journal of Complex Networks, 3*(4), 507–542.

Noyons, E. C., & van Raan, A. F. (1994). Bibliometric cartography of scientific and technological developments of an R & D field. *Scientometrics, 30*(1), 157–173.

Pan, R. K., Sinha, S., Kaski, K., & Saramäki, J. (2012). The evolution of interdisciplinarity in physics research. *Scientific Reports, 2*, 551.

Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information, 42*(1), 91–106.

Rip, A., & Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics, 6*(6), 381–400.

Ronda-Pupo, G. A., & Guerras-Martin, L. Á. (2012). Dynamics of the evolution of the strategy concept 1962–2008: A co-word analysis. *Strategic Management Journal, 33*(2), 162–188.

Rousseau, R. (2014). Library science: Forgotten founder of bibliometrics. *Nature, 510*(7504), 218.

Schulz, S., Costa, C. M., Kreuzthaler, M., Miñarro-Giménez, J. A., Andersen, U., Jensen, A. B. & Maegaard, B. (2014). Semantic relation discovery by using co-occurrence information. In: *9th Language resources and evaluation conference*. Reykjavik: European Language Resources Association.

Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. *Journal of the American Society for Information Science and Technology, 63*(1), 78–85.

Smalheiser, N. R., & Swanson, D. R. (1998). Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine, 57*(3), 149–153.

Soares, P. R. d. S., & Prudêncio, R. B. C. (2012). Time series based link prediction. In *The 2012 international joint conference on neural networks (IJCNN), IEEE*.

Su, H., & Lee, P. (2010). Network perspective of science and technology policy research community in Taiwan. *Technology management for global economic growth (PICMET), 2010 Proceedings of PICMET'10: IEEE*.

Sun, X., Kaur, J., Milojević, S., Flammini, A., & Menczer, F. (2012). Social dynamics of science. *Scientific Reports, 3*, 1069.

Tylenda, T., Angelova, R., & Bedathur, S. (2009). Towards time-aware link prediction in evolving social networks. In: *Proceedings of the 3rd workshop on social network mining and analysis*, Paris, France, Associations of Computing Machinery.

Uddin, S., Hossain, L., Abbasi, A., & Rasmussen, K. (2012). Trend and efficiency analysis of co-authorship network. *Scientometrics, 90*(2), 687–699.

Uddin, S., Hossain, L., & Rasmussen, K. (2013). Network effects on scientific collaborations. *PLoS One, 8*(2), e57546.

Uddin, S., Khan, A., & Baur, L. A. (2015). A framework to explore the knowledge structure of multidisciplinary research fields. *PLoS One, 10*(4), e0123537.

van der Eijk, C. C., van Mulligen, E. M., Kors, J. A., Mons, B., & van den Berg, J. (2004). Constructing an associative concept space for literature-based discovery. *Journal of the American Society for Information Science and Technology, 55*(5), 436–444.

Van Raan, A. (1997). Scientometrics: State-of-the-art. *Scientometrics, 38*(1), 205–218.

Van Raan, A. (2003). The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. *Technology Assessment—Theory and Practice, 1*(12), 20–29.

Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629–635.

Wang, X., Jiang, T., & Li, X. (2010). Structures and dynamics of scientific knowledge networks: An empirical analysis based on a co-word network. *Chinese Journal of Library and Information Science, 3*(3), 19–36.

Wang, C., Satuluri, V. & Parthasarathy, S. (2007). Local probabilistic models for link prediction. In *7th IEEE international conference on data mining, ICDM 2007*, Omaha, NE, IEEE.

Wang, X., & Sukthankar, G. (2014). Link prediction in heterogeneous collaboration networks. In R. S. Missaoui & I. Sarr (Eds.), *Social network analysis-community detection and evolution* (pp. 165–192). Cham: Springer.

Wang, P., Xu, B., Wu, Y., & Zhou, X. (2015). Link prediction in social networks: The state-of-the-art. *Science China Information Sciences, 58*(1), 1–38.

Wu, C., & Leu, H. (2014). Examining the trends of technological development in hydrogen energy using patent co-word map analysis. *International Journal of Hydrogen Energy, 39*(33), 19262–19269.

Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics, 8*(2), 295–309.

Yang, Y., Lichtenwalter, R. N., & Chawla, N. V. (2015). Evaluating link prediction methods. *Knowledge and Information Systems, 45*(3), 751–782.

Yu, Q., Long, C., Lv, Y., Shao, H., He, P., & Duan, Z. (2014). Predicting co-author relationship in medical co-authorship networks. *PLoS One, 9*(7), e101214.

Zelinka, I., Davendra, D. D., Chadli, M., Senkerik, R., Dao, T. T., & Skanderova, L. (2012). Evolutionary dynamics as the structure of complex networks. *Handbook of Optimization: From Classical to Modern Approach, 38*, 215.

Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B, 71*(4), 623–630.