Creation of knowledge-added concept maps: time augmention via pairwise temporal analysis

Elan Sasson, Gilad Ravid and Nava Pliskin

Elan Sasson is based at Tel Aviv University, Tel Aviv, Israel. Gilad Ravid and Nava Pliskin are both based at Ben-Gurion University of the Negev, Beer-Sheva, Israel.

Abstract

Purpose – *Although acknowledged as a principal dimension in the context of text mining, time has yet to be formally incorporated into the process of visually representing the relationships between keywords in a knowledge domain. This paper aims to develop and validate the feasibility of adding temporal knowledge to a concept map via pair-wise temporal analysis (PTA).*

Design/methodology/approach – *The paper presents a temporal trend detection algorithm – vector space model – designed to use objective quantitative pair-wise temporal operators to automatically detect co-occurring hot concepts. This PTA approach is demonstrated and validated without loss of generality for a spectrum of information technologies.*

Findings – *The rigorous validation study shows that the resulting temporal assessments are highly correlated with subjective assessments of experts (*n = *136), exhibiting substantial reliability-of-agreement measures and average predictive validity above 85 per cent.*

Practical implications – *Using massive amounts of textual documents available on the Web to first generate a concept map and then add temporal knowledge, the contribution of this work is emphasized and magnified against the current growing attention to big data analytics.*

Originality/value – *This paper proposes a novel knowledge discovery method to improve a text-based concept map (i.e. semantic graph) via detection and representation of temporal relationships. The originality and value of the proposed method is highlighted in comparison to other knowledge discovery methods.*

Keywords *Pair-wise temporal analysis (PTA), Technology assessment, Temporal trend detection, Time-augmented concept map, Vector space model (VSM)*

Paper type *Research paper*

1. Introduction

Managing and gaining insights from the vast amount of textual data produced on the Web is both a challenge and a key to obtaining and maintaining a competitive advantage (Assunção *et al.*, 2015). Mining unstructured content – either privately warehoused or publically available on the Web – can help organizations gain insights from large amounts of data (Schomm *et al.*, 2013). This data deluge, often referred to as *big data* (McAfee *et al.*, 2012; Gandomi and Haider, 2015), encompasses the challenges associated with acquiring, managing and analyzing data sets characterized by extensive volume, variety and velocity (3Vs). Data volume, from gigabytes to terabytes, refers to the increasing number of stored bytes. Data variety, from semi-structured to unstructured, refers to the increasing number of data formats. Data velocity, from batch to streaming, refers to the increasing speed of data generation, processing and use (Power, 2016). Marko and Mladeni (2005) assert that knowledge management (KM) has been enhanced by maturation of internet.

Received 12 July 2016
Revised 18 October 2016
Accepted 19 October 2016

DOI 10.1108/JKM-07-2016-0279

A closely related term – *big data analytics* – encompasses a new paradigm for using big data sources while using advanced and sophisticated data analysis approaches and tools (Müller *et al.*, 2016). It goes without saying that the potential value of big data analytics is unlocked and materialized only when leveraged to drive decision-making (Gandomi and Haider, 2015). Big data analytics is applied to massive amounts of textual documents available on the Web in a specific knowledge area via text mining (TM) applications, which harness co-word analysis to automatically generate concept maps (Porter and Detampel, 1995; Plotnick, 1997; Budanitsky and Hirst, 2006; Waltman *et al.*, 2010). Methods for extracting meaningful keywords and concepts facilitate content analysis by applying various technologies for capturing, processing, analyzing and visualizing an immense volume, variety and velocity of unstructured data drawn from multiple Web sources. One of the main ideas behind content analysis is that large bodies of text are reduced to a relatively small number of concepts, so that a large corpus can be easily managed and understood via concept mapping (Wang *et al.*, 2014). Information integration and analytical capabilities thus provide proactive business insights and trends essential for decision-making (Raghupathi and Raghupathi, 2014; Zhuge, 2015).

In the KM literature published over the past two decades, mapping of knowledge has been identified as a key activity designed to help assess and explore knowledge about a topic (Coleman and Li, 1999; Gloet and Terziovski, 2004). Alavi and Leidner (1999) assert that free text and concepts are the foundation of *information-based* KM systems, and that TM is a major component of *technology-based* KM systems. In fact, the structurally mapped knowledge may be seen as visually describing the relationships between ideas (Jonassen and Grabowski, 1993), thus allowing decision makers to quickly grasp significant visual patterns essential for dealing with the fast pace of global economy (Porter and Detampel, 1995). Zhuge (2015) asserts further that knowledge representation of semantic links in a knowledge map can improve decision-making by helping decision makers extract insights from massive textual sources.

There are four key definitions used in this paper (Callon *et al.*, 1986; Courtial, 1994; Rapp, 2002). A *concept* is defined as a logical, semantically cohesive unit of text. A *co-occurrence relation* is defined as a one-to-many mapping, which associates each concept with a list of related concepts ranked quantitatively by relatedness similarity. A *concept map* is defined as a dynamic graphical map of a knowledge domain that visually presents concepts and relevant relationship clusters, which can be portrayed as an undirected graph $G = (V, E)$ consisting of a set of vertices V (which represent the concepts) and a set of edges E (which represent the co-occurrence). A *knowledge map*, or knowledge cartography, is defined as methods and tools for visually analyzing knowledge areas to enable decision makers to discover business-relevant features or meaning in a comprehensive form (Speel *et al.*, 1999). Moreover, Marko and Mladeni (2005) have argued that information systems related to KM are moving from a data-processing mode toward a concept-processing mode, meaning that the basic processing unit is becoming increasingly less an atomic piece of data and progressively more a semantic concept that carries an interpretation and exists together with other concepts.

Sedighi and Jalalimanesh (2014) assert that KM publications generally focus on knowledge in organizations and on knowledge creation, highlighting bibliometric methods (and more specifically, co-word occurrences) as a major component to the analysis of networks of documents and keywords. Moreover, they argue that knowledge visualization and knowledge (concept) maps are commonly used in KM processes. Ding *et al.* (2000) used bibliometric techniques to break down an area of knowledge into its main elements and represent the areas and sub-areas graphically. Borner *et al.* (2003) use science mapping for analyzing the networks of links between information entities as keywords to understand the structure of science, thereby making it usable as a tool for science strategy and evaluation. Visualization tools as concept maps presently play a major role in creating maps that are more informative and easier to understand (Besselaar and Heimeriks, 2006). According to Lee and Chen (2012), such tools are used to derive new insights by identifying trends, or clusters, in the large amounts of data associated with a field of study.

However, the terms *concept map* and *knowledge map* are not synonymous. Produced via analytical applications that perform TM to extract concepts from a large textual corpus, followed by co-word analysis (Porter and Detampel, 1995; Plotnick, 1997; Budanitsky and Hirst, 2006; Waltman *et al.*, 2010), concept maps do not cope with concepts proximity. Co-word analysis reduces the data into a specific visual concept map representation. This representation is based on the nature of keywords which are important carriers of scientific ideas and knowledge (Raan and Tijssen, 1993), revealing patterns and trends in a specific discipline and responding to the challenge of extracting useful relationships among concepts (Ding *et al.* 2001). While providing visual representations of knowledge structures and knowledge organization, automatically generated concept maps leave users wondering how closely concept pairs on the map are contextually and temporally related.

We have already coped with the first challenge of contextual proximity (Sasson *et al.*, 2015) by developing, demonstrating and validating a research model for augmenting concept maps contextually. This novel method of co-word analysis helps to determine the *contextual* relatedness of concepts on the map by using webometrics Web counts to improve similarity measures of relatedness proximity. To measure relatedness proximity, we adopted the similarity link value (SLV) co-occurrence measure, also known as Equivalence Index (E), which is defined by Callon *et al.* (1991) as the ratio of the squared number of documents in which both concepts co-occur and the multiplication of number of documents in which each concept appears solely. A corpus-based SLV was calculated first, followed by calculation of a bibliometric SLV based on webometric hit count estimates (HCEs), derived from the Google search engine (Web counts). Then, the two SLV values were combined to an extended SLV value. The 2015 research paper, while focusing on contextual knowledge, acknowledged the need for further research to discover and assimilate the temporal knowledge available on the Web. The potential transformative process of combining contextual and temporal proximity measurements is beyond the scope of the present study. It is left for future research to explore whether a concept map augmented by the combination of both may yield further improvements and hopefully, yield a more understandable and accurate map than when augmented by each separately.

The aim of the current work is thus to continue where the previous paper left off and cope with the challenge of temporal proximity, acknowledging that *time* is an important factor in acquiring the knowledge needed to detect whether concepts are obsolete, emerging or hot. Hot concepts, for example, are crucial in decision-making processes concerning recent and progressive new developments in a specific scientific field under strategic investigation and assessment (Porter and Cunningham, 2005). To identify hot concepts on time, there is a pressing need for analytical exploitation of temporal textual information massively collected from a myriad of sources. The abundance of information and exponential growth of time-based textual content are well-established phenomena on the Web (Varian, 2006; Prado and Ferneda, 2007; Hauber *et al.*, 2012). Yet, one faces inherent limitations when relying on search engines for identifying hot concepts, as the vast scope of data makes it difficult to extract quickly and efficiently valuable knowledge, especially insights. This is consistent with Courtney *et al.*'s (1997) assertion that in determining what knowledge to include in KM systems, omitting the unimportant is as important as considering the important.

In the research model at the focus of this paper, objective quantitative temporal operators are used to draw upon past work on emerging trend detection (ETD) and discover how closely related are the concepts on the map *temporally*. The motivation for taking this research direction of closing the temporal gap between concept and knowledge maps is rooted in the rapidly increasing pace of events in the present harsh global business arena, where innovations occur at ever-increasing speeds and life cycles are considerably shorter.

The approach presented in this work temporally augments concept maps via pair-wise temporal analysis (PTA) by using a temporal trend detection algorithm based on the vector space model (VSM). The PTA approach focuses on expressing the temporality value of the relationship between two co-occurring concepts, and particularly identifying pairs of hot concepts. Implementing this novel, unsupervised and automated algorithm, which uses objective quantitative pair-wise temporal operators to automatically detect co-occurring hot concepts, facilitates the temporal augmentation of a *concept map* to a *knowledge map*. A decision maker using the time-augmented concept map is able to derive hype-free insights and view a comprehensive picture – not only of the targeted general landscape evaluated but also of important temporal trends, providing actionable knowledge that might not be cost-effectively achievable otherwise.

The temporal augmentation approach demonstrated and validated in this work leverages, in addition to the PTA approach, a synergy of two methods to uncover hidden patterns in textual data. The first synergetic method to precede demonstration and validation of the PTA approach involves gradually building a corpus of unstructured textual data from diverse Web-based sources about a target topic. Google Alerts (GA) content change-detection and notification service, which automatically notifies subscribers when a new Web content matches a set of search terms associated with the target topic, was used to initiate corpus building. As GA determines source validity, this corpus building method allows collecting relevant documents without the need to subjectively evaluate the cardinality or the authority of the feed sources. While other approaches use controlled and limited content in closed databases (such as digital libraries of articles), possibly missing useful and relevant knowledge, the harnessing of temporal data referenced in GA messages to build a dynamic and open corpus is novel. The second synergetic method to precede demonstration and validation of the PTA approach involves uncovering hidden patterns in the corpus and generating a concept map in the form of a co-occurrence network of textual entities (i.e. keywords). This was achieved by applying information extraction (IE) to each document (HTML page) collected earlier to the corpus, using TM guided by natural language processing (NLP) to discover and extract the concepts, followed by a co-word analysis.

The research presented in the current paper provides innovative theoretical and practical contributions to KM in general decision-making processes, especially those related to technological assessment. From the theoretical perspective, this study presents the first attempt to model the addition of temporal knowledge to a concept map. The innovation is attributed to the simplicity of implementing temporal measures based on objective time properties derived via two quantitative temporal pair-wise operators from the time-tagged textual corpus. From the practical perspective, the contribution lies in the development of an automated research instrument capable of supporting decision makers in knowledge discovery tasks. In technology assessment (TAS), for instance, the developed instrument helps decision makers gain clear knowledge about a specific technology when they evaluate technology alternatives, and then identify future technological trends. Specifically, based on the software modules developed to obtain its results, this study contributes a managerial decision-support tool for managers which deals with the absence of temporal knowledge from traditional concept maps. For TAS, the research instrument assists technology-savvy decision makers in quickly getting actionable information in the form of a comprehensive and parsimonious picture of the general landscape of an assessed technology in the form of a concept map, identifying important indications of trends over time.

To the best of our knowledge, this study represents the first attempt to measure temporal proximity, thus upgrading traditional concept maps into innovative *knowledge maps*. It is worth noting that big data analytics applications generally consider one, two or three of the 3V ingredients (volume, variety and velocity). At the heart the current work, the two components of volume and variety are considered specifically, as high volume of textual data derived from the Web (clearly beyond the human capability to cope with in a timely manner) and high variety of unstructured data (such as presented in a gradually constructed Web-based text corpus).

The theoretical foundation for these contributions is presented next (Section 2), followed by the research model developed (Section 3) and the method used in this work (Section 4). The last two sections are primarily devoted to a review of the results, focusing on demonstration of the PTA analysis and on its validation respectively, with concluding remarks devoted to discussing the study's limitations and future research.

2. Literature review

Decision makers need to foresee advances and assess new innovations for insights about new developments and for strategic business planning (Halsius and Lochen, 2001). Many studies (Rousseau, 1979; Russell *et al.*, 2010) describe such insight assessments as essential in managerial decision-making processes, especially for managers faced with the challenge of quickly identifying emerging and hot technologies with the greatest potential (Courseault, 2004). Bolshakov and Gelbukh (2004) state that well-informed decisions require decision makers to read an enormous quantity of Web text. As the data collection space is rapidly expanding with many new sources and types, managers face both the challenges and the opportunities associated with timely analysis of big data for strategic KM-driven decisions (Zhuge, 2015). In fact, Dzone software's blog (2013) puts keeping data accurate, making sure information is relevant, and interpreting data efficiently on the list of top ten KM challenges.

A potential application of the current work lends itself to TAS. To survive in the hyper-competitive business environment, organizations regularly engage in TAS processes prior to decision-making regarding investments in existing, emerging and hot technologies. Assessment of an information technology (IT) is especially challenging due to increasing rates of innovation and shorter life cycles. Swanson and Ramiller (2004) argue that the ultimate goal of the TAS process is to provide guidance to managers on the question of whether, when and how to innovate. With regard to technologies, it is imperative that organizations have a clear understanding of new strategies, evolving architectures, hot trends, emerging products and new standards with emphasis on deploying KM technologies (Ashrafi *et al.*, 2006). Looking for relevant information via Web search engines, decision makers involved in TAS processes face an abundance of information that limits their ability to assess technologies within a reasonable time frame. Moreover, Web searches do not necessarily lead to discovery of relevant knowledge, being mainly based on retrieval of textual documents without ability to screen noise, find hidden knowledge patterns or sift through "the wheat from the chaff".

Unable to deal with the rich knowledge scattered all over the Web, organizations turn to leading consulting firms and analyst groups. These vendors compile expensive reports whose objectivity is sometimes in question. An additional problem with acquiring knowledge from domain experts is the knowledge acquisition bottleneck resulting from limits on the amount of available time (Li and Zhong, 2004). Against this background emerges the following research question:

RQ1. How is it possible to extract valuable knowledge from a diverse corpus of textual data on the web?

Power (2016) asserts that supporting decision-making using such new data sources presents an important challenge and calls for more research, discussion and analysis in response for this challenge. The remainder of the literature reviews sets the stage for the current paper to address this challenge by developing a knowledge mapping research model and instrument that handle massive amounts of unstructured textual in support of decision-making processes with a temporal focus.

Almost two decades have passed since Dixon (1997) alluded to the impracticality and impossibility of processing or understanding massive quantities of textual information. Unsurprisingly, Lee *et al.* (2010) assert that manual analysis of unstructured textual data is becoming increasingly unfeasible. In the present complex business environment, where emergence is characteristic of a decision-making situation (Tredinnick, 2009), knowledge needs to be continuously reviewed to assess its relevancy and validity (Intezari and Pauleen, 2013). Obviously, this knowledge review must rely on some automation.

Unique data storage, management, analysis and visualization technologies allow big data analytics techniques to process automatically a data corpus featuring immense volume, variety and velocity (Chen *et al.*, 2012; Zikopoulos and Eaton, 2012; Das and Kumar, 2013). Wang *et al.* (2014) assert that the relative analytical advantage of big data architecture over more traditional architectures stems from the fact that the former has a unique ability to analyze unstructured textual data, to process large data volumes in parallel and to parse data in real time or near real time. For example, Méndez-Torreblanca *et al.* (2000) note that the rapid expansion of text documents on the Web, and the increased difficulty of extracting potentially useful knowledge from massive amounts of textual data, make TM methods of great importance on the Web (Matsuo *et al.*, 2006; Das and Kumar, 2013). Debortoli *et al.* (2016) claim that manual analysis of textual Web data is virtually impossible and suggest applying automated TM techniques.

The objective of TM is to exploit information contained in textual documents in various ways, including automated discovery of patterns or trends as well as associations among text objects like concepts (Grobelnik *et al.*, 2000). Linguistics-based TM is guided by NLP and involves IE. IE is the task of extracting named entities and factual assertions from a textual corpus (Wilks, 1997; Gandomi and Haider, 2015), allowing transformation from the unstructured document space to the structured concept space and, when followed by co-word analysis, paving the way to analysis of concept interactions. Confronting the problem of knowledge discovery from textual data sources, TM and NLP processing are two big data analytics techniques that Das and Kumar (2013) highlight as core components of unstructured data analysis on the Web.

There is a fairly extensive body of literature on co-word/co-occurrence analysis (Callon *et al.*, 1986; Courtial, 1994). Feldman *et al.* (1997) published an early seminal work on concept maps featuring co-occurrence relationships in a corpus of documents. Their paper paved the way for co-word analysis to be viewed as a powerful and proven quantitative tool for knowledge discovery in a certain research field (He, 1999). According to Rapp (2002), concepts that co-occur tend to be related. Demonstrating relatedness association, co-occurring concepts have therefore been considered as carriers of meaning across different domains, and as general indicators of activity in textual document sets (Leydesdorff and Hellsten, 2006). Moreover, the recognition of the implicit links of co-occurring concepts helps to establish insights and to support analysis of knowledge in various domains under investigation (Zhuge, 2015).

TM applications generate graphic representations of concepts and relationships from text documents in a knowledge domain (Leake *et al.*, 2001). Time stamps, like publication dates, and the underlying temporal and evolutionary structure, are unfortunately ignored in the process (Mei and Zhai, 2005). This drawback of TM is highlighted by Desikan and Srivastava (2004), who emphasize the significance of accounting for the time dimension, given the need to reflect current trends and help predict future ones regarding emerging and hot topics. Pottenger and Yang (2001) similarly symbolize the process of detecting emerging conceptual content in regions of semantic locality in concept maps, analogous to the operation of a radar system which aims to effectively assist in the differentiation of mobile (i.e. dynamic) and stationary (i.e. static) objects. In the big data analytics space, Zhuge (2015) specifically emphasizes the time dimension as a key component, which determines the way to organize and operate multi-dimensional data in applications (such as life-cycle analysis of an observed technology). Temporal text mining (TTM), aims to discover temporal patterns in textual data collected over time and, thus, responds to the TM challenge of detecting emerging and hot concepts. Relying on concept mapping from the digital domain to the semantic domain in a plethora of domains, TTM facilitates concept map augmentation in a temporally sensitive environment as (Chen, 2006), summarizing events in news feeds for example (Boykin and Merlino, 2000; Ma and Perkins, 2003; Morinaga and Yamanishi, 2004; Rajaraman and Tan, 2001).

A growing TTM literature (Swan and Jensen, 2000; Pottenger and Yang, 2001), devoted to ETD applications that help address the temporal challenge of concept mapping, suggests that much progress has been made toward automating discovery of emerging and hot trends. Kontostathis *et al.* (2004), in an in-depth survey of the literature, indicate that ETD systems can be classified as either fully-automatic (Swan and Jensen, 2000; Pottenger and Yang, 2001) or semi-automatic (Porter and Detampel, 1995; Blank *et al.*, 2001; Roy *et al.*, 2002). In contrast to fully-automatic systems, semi-automatic ones first require user input about a certain topic and then usually provide user-friendly reports and screens that summarize the evidence available on the topic, indicating whether it is truly an emerging topic. ETD systems, however, have three limitations that the approach demonstrated in the current work aims to overcome.

First, most ETD systems impose limitations, such comparability and completeness, by requiring a human reviewer to subjectively finalize concept classification (Porter and Detampel, 1995; Nowell *et al.*, 1997; Blank *et al.*, 2001; Roy *et al.*, 2002; Havre *et al.*, 2002; Blank *et al.*, 2002; Chen, 2006). As indicated above, semi-automatic ETD applications (Nowell *et al.*, 1997; Blank *et al.*, 2001; Roy *et al.*, 2002) require user input prior to providing trend evidence. The constructive collaborative inquiry-based multimedia e-learning system, using document count-based search of the major indexing INSPEC database of scientific and technical literature, also requires user input (Blank *et al.*, 2001). The ThemeRiver system, which visualizes thematic variations over time within a large collection of documents, similarly relies on human expertise (Havre *et al.*, 2002). Technology opportunities' analysis couples bibliometric methods with expert opinion to provide trend insights regarding specific technologies (Porter and Detampel, 1995). The current work aims to automatically add temporal knowledge to a concept map without explicit user input. This is accomplished by implementing the process of deriving objective time properties in an unsupervised mode by using objective temporal operators based on the VSM and the cosine similarity measure in PTA.

Second, unlike the approach demonstrated in the current work, previous ETD studies focused on detection of a single stand-alone concept at a one-item level. Such inclusion of unary-based analysis for detecting trends carries a significant risk of missing important temporal indications regarding co-occurring concept pairs. For example, past ETD research projects (Desikan and Srivastava, 2004; Blank *et al.*, 2002) performed single node analysis of time-stamped documents to detect changes by identifying periods with a burst of activities related to a stand-alone topic (concept). In these works, if the frequency of documents referencing the concept increases over time, the potential emerging concept is confirmed as a bona fide trend with respect to the main topic under assessment. To be considered a *hot concept* in such unary-based analysis of a large corpus, for example, only the number of occurrences of a particular concept should exhibit an accelerating occurrence. To accomplish the goal of improving the temporal dimension of concept mapping in the current study, detection of temporal attributes of co-occurring concepts is conducted at a two-item level. This PTA first adds knowledge by presenting the decision maker with an augmented concept map, one that goes beyond tracing changes over time for frequency of one single concept with reference to other concepts as well. Then, if the study of one concept at a certain period is found to influence or stimulate the study of another concept at the same period, the discovery of such an evolutionary relationship between concepts can reveal not only the hidden concepts as a single semantic entity but also the latent inter-linkage of timely synchronized hot co-occurring concepts.

Finally, a third limitation of most ETD systems relates to the use of a unitary static monolithic text corpus from human-maintained indexed databases, such as INSPEC, topic detection and tracking (TDT) or COMPENDEX (Nowell *et al.*, 1997; Lent *et al.*, 1997; Swan and Jensen, 2000; Wong *et al.*, 2000; Kumaran and Allan, 2004; Mei and Zhai, 2005; Zhang *et al.*, 2007; Subasic and Berendt, 2010; Chen and Chundi, 2011). A closed static textual data corpus suffers from limited diversity, variety and richness and must be periodically refreshed, imposing major drawbacks, such as data coverage and indexer effect (Alexa, 1997; Zweigenbaum *et al.*, 2001; Banko and Brill, 2001; Keller and Lapata, 2003). These drawbacks possibly cause controlled and limited content, as well as missed useful and relevant content, as indexing might reflect prejudices of professional indexers (King, 1987; Law and Whittaker, 1992). In response to Debortoli *et al.*'s (2016) recommendation, the current work aims to use open data repositories, such as the Web that contains large collections of textual data sets, instead of controlled and limited data repositories. This goal is accomplished by harnessing temporal data referenced in GA messages to build a dynamic and open corpus.

The current paper adds value to past work, as evident if compared to Marko and Mladeni's (2005) review of selected knowledge discovery methods. Their emphasis is on TM examples for document categorization, document clustering, ontology learning and concept mapping as a form of sematic graphs. Pattuelli and Miller (2015) present a novel approach to the development and semantic enhancement of a social network to support the analysis and interpretation of digital oral history data from jazz archives and special collections. Their approach is similar to the current work in providing an automated IE method to create a social network. Cunningham *et al.* (2005) provide a high-level introduction to IE and descriptions of application scenarios for KM tools that exploit IE. Dasgupta *et al.* (2015) have developed a new digital library architecture that supports a polyhierarchic ontology structure where a child concept representing an interdisciplinary subject area can have multiple parent concepts. However, they use a closed static corpus of a digital library, in contrast to the open dynamic corpus used in the current work. Van den Berg and Popescu (2005) apply knowledge map tools for knowledge diffusion by professionals in organizations and communities of practice in The Netherlands. None of these works address the temporal challenge addressed here.

Using a closed static corpus and focusing on detection of a single stand-alone keyword, as opposed to the method proposed here, Ribiere and Walter (2013) provided keyword frequency and content analysis based on TM of all 235 journal articles published in 2003-2012 by the journal *Knowledge Management Research and Practice*. Sedighi and Jalalimanesh (2014) identified the research trend in the field of KM by presenting a systematic and analytical scientometrics approach based on the Web of Science (WoS) database. Their co-word occurrence analysis for mapping KM research topics showed that the structure of fundamental subject areas within the KM field has changed and expanded dynamically in 2004-2010. However, they use a single source to construct the corpus, as opposed to various source types (news, Web, blogs and discussion group sites) used in the current research. Khasseh and Mokhtarpour (2016) applied the "Referenced Publication Years Spectroscopy" (RPYS) method, based on frequency analysis of the references cited in the scientific publications of a particular area based on their publication years, to show formation of knowledge using citation analysis. The temporal approach presented in their study is based on a closed corpus extracted from the WoS database, with the number of references constituting the temporal variable. The current research uses two pair-wise temporal operators and uses a more diverse corpus.

Having set the background for adding temporal knowledge to a concept map using a web-based textual corpus that grows over time, the current research model aims to go beyond conventional ETD systems to discover co-occurring hot concepts by implementing temporal pair-wise concept analysis.

3. Research model: pair-wise temporal analysis

At the core of the research model underlying this work is the addition of temporal knowledge to a concept map. This is accomplished by means of PTA based on objective time properties of co-occurring hot concepts derived from time-tagged textual corpus to determine concept categorization based on the time dimension via two quantitative pair-wise temporal operators defined as follows:

- *age* of relevant documents where concepts co-occur; and
- *frequency* rate of publishing relevant documents where concepts co-occur in a given time interval.

Following Pottenger and Yang (2001) and Goorha and Ungar (2010), a pair of concepts is classified as *hot* if the concepts are semantically richer at a later time and co-occur more frequently as an increasing number of documents reference them. Thus, the level of "hotness" (i.e. Age) and the level of "activeness" (i.e. Frequency) can, respectively, be described as *young* (hot) and *active* (progressive). Detection of hot topics is a sub-process of topic detection and tracking (TDT) (Wayne, 1997; Chen, 2006), primarily aimed at detecting changes in topics (e.g. disruptive events) and exhibiting discontinuities in semantics in localized data sources, such as newscasts (Carbonell *et al*., 1999). TDT is generally focused on five types of tasks (Wayne, 1997; Chen, 2005):

1. story segmentation (e.g. finding topically homogeneous regions);
2. topic detection (e.g. detecting the occurrence of new events);
3. topic tracking (e.g. tracking the recurrence of known events);
4. first-story detection (e.g. detecting the first time occurrence of new stories); and
5. story-link detection (e.g. detecting linkages to stories).

In the context of the current study, TDT is relevant to PTA.

To express the temporality value of a co-occurrence relationship in a concept map and describe it as *old* or *young*, thus adding temporal values to concept maps, the current study enhances the scalability of the VSM by conducting PTA that exploits the cosine similarity measure (Salton, 1988; Salton *et al*., 1975). Given that Concept i and Concept j co-occur in n documents, a Vector $\vec{y}$ with n dimensions (where each coordinate reflects the number of days since creation of each document) is assembled, and documents are accordingly chronologically ordered. The cosine similarity measure is then applied to Vector $\vec{x}$ (with n dimensions as $\vec{y}$), where all its coordinate reference values are 1's to express fresh temporal notions and chronological proximity to the present time. The cosine similarity measure – i.e. the angle between Vector $\vec{x}$ and Vector $\vec{y}$ is $0 \leq \cos(\vec{x},\vec{y}) \leq 1$ – would be close to 1 if $\vec{x}$ and $\vec{y}$ are nearly identical, indicating a young temporal relationship between the co-occurring concepts. Contrastingly, the cosine similarity measure would be close to 0 if $\vec{x}$ and $\vec{y}$ have little in common, indicating an old temporal relationship between the co-occurring concepts. Although other measures, such as an average, could be used to capture the temporal notion of age, it was found that the VSM model is more resilient to outlier values and more robust for normalization of vectors of different lengths.

Figure 1 demonstrates this approach in the case of Concept *i* and Concept *j*, which co-occur in four ($n = 4$) different documents published one year ago, four months ago, one month ago and one week ago ("a" in Figure 1). The two corresponding variable-size date-ordered vectors ($x_i y_i \in R^n$) are: $\vec{x} = (1, 1, 1, 1)$ and $\vec{y} = (7, 30, 120, 365)$. *Age*($\vec{x}$,$\vec{y}$), the first quantitative pair-wise temporal operator of the PTA ("b" in Figure 1), is accordingly defined as:

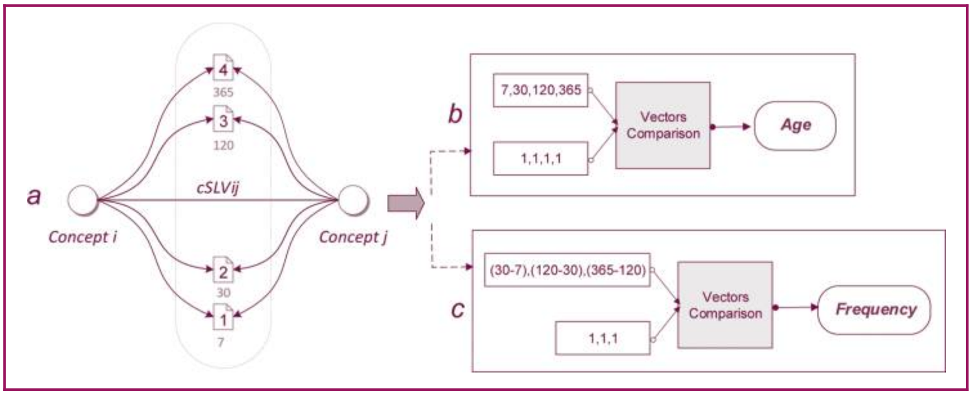
$$Age(\vec{x}, \vec{y}) = \frac{\sum_{k=1}^{n} x_k \, y_k}{\sqrt{\left(\sum_{k=1}^{n} x_k^2\right) \times \left(\sum_{k=1}^{n} y_k^2\right)}} \quad where\, \vec{y} = (y_1, \ldots, y_n); \vec{x} = (x_1, \ldots, x_n); x_{1 \ldots n} = 1$$

It is worth noting that in rare case where there are two vectors with an identical cosine similarly value yet different magnitude (i.e. Euclidean norm) values, a compensation factor is applied. The factor is derived from the division ratio of the two corresponding vectors' length, thus preserving the difference in magnitude while computing the cosine similarity values. Although this mechanism is included in the research instrument, such cases were not observed in the corpus collected for the current study. *Frequency*($\vec{x}$,$\vec{y}$), the second quantitative pair-wise temporal operator of the PTA ("c" in Figure 1), is also a vector whose sequentially ordered coordinates are calculated by subtracting the values of two subsequent coordinates $|y_k - y_{k-1}|$ of the corresponding Vector $\vec{y}$. This subtraction should yield minimal values, as a high publication ratio indicates an imaginary notion of documents published on a daily basis. Thus, the cosine similarity measure is then applied to a reference Vector $\vec{x}$ with $n - 1$ dimensions, where all coordinates assume the value 1 to reflect an *active* temporal value and where 0 is not applicable. The two corresponding variable-size date ordered vectors ($x_i y_{i,} \in R^n$) are $\vec{x} = (1,1,1)$ and $\vec{y} = (23,90,245)$. *Frequency*($\vec{x}$,$\vec{y}$), the second quantitative pair-wise temporal operator of the PTA ("c" in Figure 1) is accordingly defined as:

$$Frequency(\vec{x}, \vec{y}) = \frac{\sum_{k=1}^{n-1} (x_k - x_{k-1})(y_k - y_{k-1})}{\sqrt{\left(\sum_{k=1}^{n-1} (x_k - x_{k-1})^2\right) \times \left(\sum_{k=1}^{n-1} (y_k - y_{k-1})^2\right)}} \quad where\, (x_k - x_{k-1}) \overset{def}{=} 1$$

In the current study, the quantitative pair-wise temporal operators should be re-evaluated constantly to reflect progress of the time dimension, as a time-tagged corpus is being investigated. In fact, assuming that a document's discovery date is identical to its publication date, daily recalculation is advised, as the value of the current date (seed date) is the basis for computing the values of both *Age*($\vec{x}$,$\vec{y}$) and *Frequency*($\vec{x}$,$\vec{y}$). To automatically detect co-occurring hot concepts in the current research model, the following three sub-tasks are implemented automatically.

Figure 1 Concept co-occurrences demonstrated for four different documents



In the first implemented sub-task, in addition to the already-calculated two quantitative pair-wise temporal operators *Age* and *Frequency*, the present research model calculates *CCDR* (concept co-occurring documents ratio) as a quantitative normalized operator reflecting the capacity of the number c_{ij} of documents in which hot concepts potentially co-occur:

$$CCDR_{ij} = \frac{c_{ij} - min(l)}{max(l) - min(l)}$$

Where $l = (c_{11}, \ldots, c_{nm})$, is a list of all pairs of concepts defined in the co-word analysis.

The three operators – *Age*, *Frequency* and *CCDR* – are combined in the second implemented sub-task of the present research model into the PTA formative construct, using the weights ω_1, ω_2, ω_3 in the following weighted linear equation:

$$\underbrace{PTA_{ij}}_{hot} = f\{(\omega_1 \times Age_{ij}(\vec{x}, \vec{y})), (\omega_2 \times Frequency_{ij}(\vec{x}, \vec{y})), (\omega_3 \times CCDR_{ij})\}$$

$$= (\omega_1 \times Age_{ij}(\vec{x}, \vec{y}) + \omega_2 \times Frequency_{ij}(\vec{x}, \vec{y}) + \omega_3 \times CCDR_{ij})$$

where:

$$i, j = 1, \ldots, \#concepts\ \omega_1 + \omega_2 + \omega_3 = 10 < PTA_{ij} \leq 1$$

To automatically detect co-occurring hot concepts, classification according to the PTA value is implemented in the third and final implemented sub-task of the research model, as follows:

Concepts *j* and *i* are co-occurring hot concepts if $\underbrace{PTA_{ij}}_{hot} \geq \tau$ and otherwise if $\underbrace{PTA_{ij}}_{hot} < \tau$.

The threshold value τ may be set between 0 and 1 either manually or empirically (Swan and Jensen, 2000; Kontostathis *et al.*, 2004). In either case, it is quite common in practice to allow manual calibration of the threshold values, acceptable to the decision maker, with a τ value close to 1 reflecting the most rigorous classification requirement.

Previous studies refer to hot topics as temporally synonymous to emerging topics (Bun and Ishizuka, 2006; Chen *et al.*, 2007; Goorha and Ungar, 2010; Kasiviswanathan *et al.*, 2011). However, the joint appearance of the former is characterized by a rather late start in the corpus and then steady growth over time (Blank *et al.*, 2001), while the joint appearance of the latter is characterized during a certain period by stable yet frequent appearances in the corpus (Bun and Ishizuka, 2003).

The current study therefore uses the *CCDR* operator to differentiate between emerging and hot co-occurring concepts as carriers of different temporal features. To automatically detect emerging co-occurring concepts, as opposed to hot ones, concepts that have only recently started to appear in the corpus, and thus have low SLV, are identified as emerging in this research. This means that the *CCDR* operator is omitted, and the PTA construct for emerging co-occurring concepts includes only the two quantitative pair-wise temporal operators *Age* and *Frequency* rather than all three operators.

While the temporal pair-wise research model has been developed in the current study with this differentiation in mind, its validation in Section 6 focuses on the weighted sum of all three operators in the case of hot concepts. Omission of a third component from the research model in the case of emerging concepts is left to future research and is therefore not presented nor discussed further in the context of the current paper. It is also worth noting that in this study, unlike previously discussed trend detection projects, a human reviewer was not required to subjectively finalize concept classification. As elaborated upon in the following methodology section, expert input was required for model validation purposes only.

4. Methodology

Instead of using controlled and limited content in closed databases as digital libraries of articles, possibly missing useful and relevant knowledge, a dynamic temporal corpus was gradually built by collecting textual data from diverse Web-based sources, thus fully leveraging the potential of hotness discovery. A corpus of unstructured textual data was progressively created about a target topic using the GA change-detection and notification service, which automatically notifies subscribers when new textual content appears, matching a set of search terms associated with that target topic. Each GA message includes one (or more) URL link(s) to Web documents (e.g. HTML, XML) about the specific topic published on the Web. The setting for the delivery rate of the GA messages was defined on an "as-it-happens" basis. As the corpus is continuously created, it is safe to assume that there is no time lag between the publishing date of a document and the respective GA creation date. Moreover, as the GA service determines source validity, this method of corpus building allowed for collecting relevant documents without the need to subjectively evaluate the cardinality or the authority of the feed sources.

In addition to the unique and novel approach to building a temporal open corpus, the temporal augmentation approach demonstrated and validated in this work has leveraged a synergy of several well-established research fields to uncover hidden patterns in the corpus. Generating a concept map, and forming a co-occurrence network of textual entities (i.e. keywords), was achieved by TM guided by NLP. To accomplish a rigorous and robust TM/NLP-based IE, as well as discover the concepts within each document (HTML page) in the time-tagged corpus, IBM's SPPS/PASW Text Analytics Version13 (formerly SPSS TM Modeler) and AlchemyAPI were used in parallel with a domain-specialized related dictionary add-on. This IE process uses a named-entity processor which allows identification of multi-gram concepts (i.e. NLP phrases), such as person names, location names and names of organizations. Moreover, to improve the accuracy of text extraction, domain-dependent linguistic resource file is applied during the extraction phase to refine the underlying rules and dictionaries. Both IE processes are regarded as black-box facilitators that are used mainly for pre-processing tasks, but alone cannot automatically detect hot co-occurring concepts.

Given a concept map derived from a time-tagged textual corpus, the algorithm facilitates addition of temporal knowledge by means of PTA, based on objective time properties of co-occurring hot concepts. For each pair of extracted concepts, three quantitative pair-wise temporal operators are calculated and stored with relevant meta-data – *Age, Frequency* and *CCDR*- and are then combined to a weighted sum using the $\omega_1,\omega_2,\omega_3$ weights to obtain the formative PTA construct according to which concepts are categorized as hot or not.

Without loss of generality, demonstration and validation of the research model has been accomplished in the present study to assess five information technologies, and to conduct an exploratory survey of 38 technology experts to establish the values of ω_1, ω_2, ω_3. The demographic distribution of data collected from these respondents shows that 65 per cent of them were practitioners. Their years of experience were distributed as follows: 10 per cent with 1-3 years of experience, 13 per cent with 4-8 years and 76 per cent with over nine years of experience. Hence, the majority of respondents are experienced practitioners. Using constant sum questions, which permit collection of ratio data on a scale of 100 points, each respondent was asked to distribute the 100 points and give the more important ω_i a greater number of points, so that the total equals exactly 100 points. Data analysis reveals the following statistics:

$$\omega_1 \rightarrow (Age\,(\vec{x},\vec{y})) = 22.4\ (\text{Median: } 20,\ \text{SD: } 9.52)$$

$$\omega_2 \rightarrow (Frequncy\,(\vec{x},\vec{y})) = 37.1\ (\text{Median: } 40,\ \text{SD: } 9.71)$$

$$\omega_3 \rightarrow\ (CCDR) = 40.5\ (\text{Median: } 40,\ \text{SD: } 12.77)$$

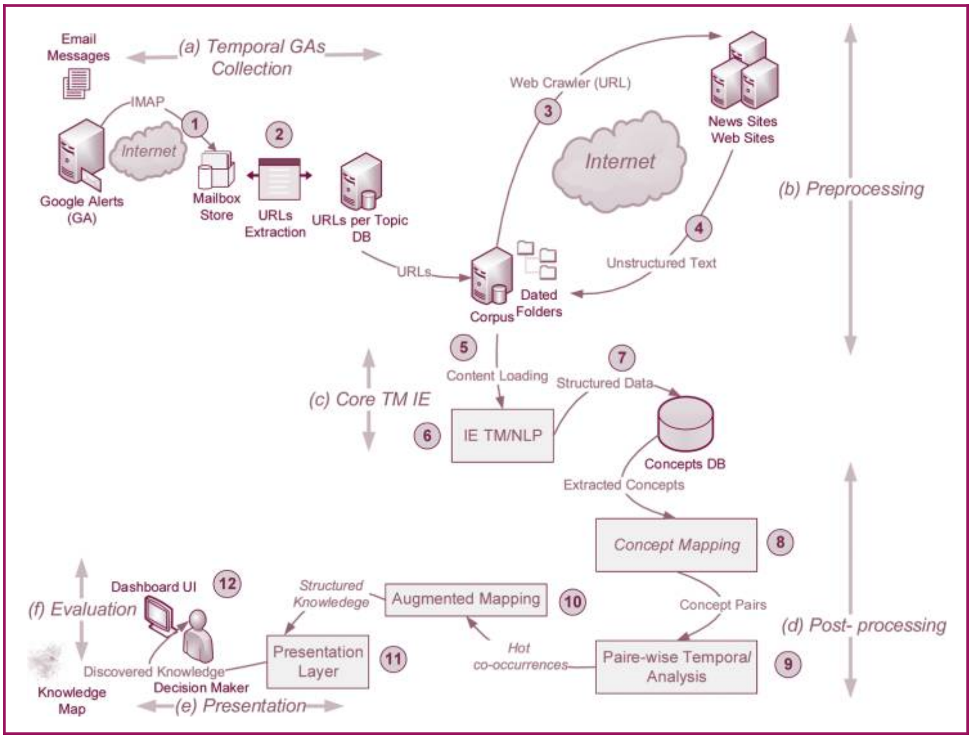
Moreover, as 11 respondents (33 per cent) specifically indicated the following values: $\omega_1 = 20\%$, $\omega_2 = 40\%$ and $\omega_3 = 40\%$, it is reasonable to adopt these exact proportions $\omega_1 = 0.2$, $\omega_2 = 0.4$ and $\omega_3 = 0.4$ in the PTA analysis as weights for the *Age*, *Frequency* and *CCDR* operators, respectively.

As indicated above, co-occurring concepts are classified as *hot* if their PTA is above the threshold τ value. The value of τ is set either manually or empirically between 0 and 1 (1 being the most rigorous classification requirement), often allowing the decision maker to fine tune τ manually. Fine tuning the τ value for the TAS demonstration and validation in the present study has revealed that $\tau = 0.6$ includes *hot* and *near-hot* co-occurring concept pairs and excludes *not-hot* ones. It is beyond the scope of the current study to devise a scheme for adjusting the threshold τ for an investigated domain, given the ability to demonstrate and validate the research model with a fixed threshold. In future research, however, it would be worthwhile to find out whether the threshold is domain-dependent.

To allow technology-savvy decision makers to automatically generate and explore a temporally augmented concept map via a graphical user interface (UI), differentiating between co-occurring hot/near-hot and not-hot concepts by means of a meaningful spring-force network layout algorithms, the research model must be converted to a useful application which is beyond the scope of this paper. To demonstrate and validate the research model developed in the current study, a Web-based research instrument was developed and, as depicted in Figure 2, may be divided into the following six main stages and 12 tasks:

1. *Temporal GA collection* tasks involve collecting a repository of GA email update messages, each including one or more URL links to domain-specific (i.e. IT topic) Web documents (e.g. HTML, XML) in diverse web sites. Steps 1 and 2 depict this stage in Figure 2.

Figure 2 The stages and tasks of the research instrument



2. *Preprocessing* tasks include all routines, processes and methods required for using crawling techniques to fetch the actual HTML files. A crawler Web agent is applied to automate the execution of the actual textual data gathering, starting from a list of URLs stored in the repository created in Stage (a), including all the links embedded in the GA email messages received over time. The crawler follows all links to actually collect the required Web pages, and locally stores and indexes the collected textual data in a repository on a dedicated corpus server for further use and analysis. Steps 3 to 4 depict this stage in Figure 2.

3. *Core TM and IE* NLP-based tasks are routines and processes for concept discovery in the document corpus yielded by Stage (b), extracting and storing for further analysis categorized, keyword-labelled and time-stamped concepts and their relevant metadata (e.g. time stamp, total number of appearances, average concept distribution, etc.). Steps 5 to 7 depict this stage in Figure 2.

4. *Post-processing* analysis tasks include all procedures and methods required for conducting the PTA toward knowledge mapping. Steps 8 to 10 depict this stage in Figure 2.

5. *Presentation* tasks and browsing functionality include easy-to-use, point-and-click and browser-based UI and listing capabilities. *Presentation layer* components display the knowledge map with references to co-occurrence weights calculated at each step, as well as the detected co-occurring hot concept. We chose to use the ORA Network Visualizer, developed by researchers at the Centre for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. Step 11 depicts this stage in Figure 2.

6. *Evaluation* tasks are carried out by the decision maker who considers and interprets the acquired results, and are therefore not depicted in Figure 2. Generalization, pruning or requiring collection of additional textual data to enrich the corpus may be implemented in this stage by the user.

To validate the PTA, the match-to-expert scoring approach was applied to more than 23 respondents for each assessed technology, all of which were recruited to respond to a simple Web-based questionnaire. Survey questionnaires were distributed internationally over the Web for each of the five assessed technologies to a database of domain experts obtained from two major sources: LinkedIn and a leading global IT consulting firm, yielding a total of n = 136 respondents.

The survey questionnaire comprised two major parts:

1. questions about respondents' demographic characteristics (nationality, position and years of experience relevant to an assessed IT topic); and

2. for PTA validation, questions about 10 pairs of co-occurring hot concepts for each assessed technology asking respondents to rate whether a pair of co-occurring concepts is hot (binary scale).

Five pairs were randomly selected from a list of co-occurring hot concepts that the PTA has previously detected as co-occurring hot concepts, and five pairs were randomly selected from a list of co-occurring concepts which the PTA did not previously detect as co-occurring hot concepts.

The PTA was conducted disjointedly for each of the five assessed information technologies to demonstrate and validate the research model. To ensure that these five technologies could serve as use cases with relevance to temporally augmented concept mapping regarding other topics as well, the spectrum of lifecycle maturation stages of these technologies was sufficiently diverse upon starting to build the temporal corpus. Model demonstration follows in the next section for the five ITs.

5. Results: model demonstration

GA messages were collected onto a temporal corpus in the course of 190 days in 2011. In the IT domain – one of the most over-hyped industries characterized by increasing rates of innovation and rather short life cycles – a six months period is sufficient for hot topics to appear in Web publications. A corpus collected over a shorter period may fail to facilitate temporal categorization of co-occurring concepts as hot due to insufficient representation of the domain under assessment. Moreover, because in the validation process (described in the next section) respondents were asked to evaluate whether co-occurring concepts are hot, six months is a reasonable period for them to recognize and be fully be aware of the evaluated concepts.

With each collected GA being an aggregation of URLs to the latest news articles about one of the five technologies, 39,724 URLs of various source types (news, Web, blogs and discussion group sites) were used altogether (after converting all HTML files to text files) to demonstrate and validate the PTA research model. More specifically, the collected corpus included 12,535 documents about the Cloud Computing technology (much-hyped in 2011), 6,470 documents about Grid computing technology (expected in 2011 to be substituted by cloud), 8,908 documents about business process management technology (which attracted in 2011 a lot of new attention), 6,030 documents about Semantic Web technology (regarded in 2011 as new and particularly promising) and 5,781 documents about service-oriented architecture (SOA) technology (already considered in 2011 a *de facto* standard on the Web).

To yield high-quality concept maps, the used dictionary (i.e. domain-specific resource file) was IT-sensitive, allowing extraction of multi-words and acronyms of IT-specific concepts (such as *operating system*, *Amazon Web services* and "*HTML 5*" to name a few extracted concepts). Generation of a temporally augmented concept map was carried out disjointedly for each of the five information technologies. Table I presents a partial list of automatically identified co-occurring hot concepts (i.e. NLP phrases discovered via IE) for each technology with normalized obtained PTA. Figure 3 displays, for example, a partial list of co-occurring hot concepts for SOA as a browser-based table.

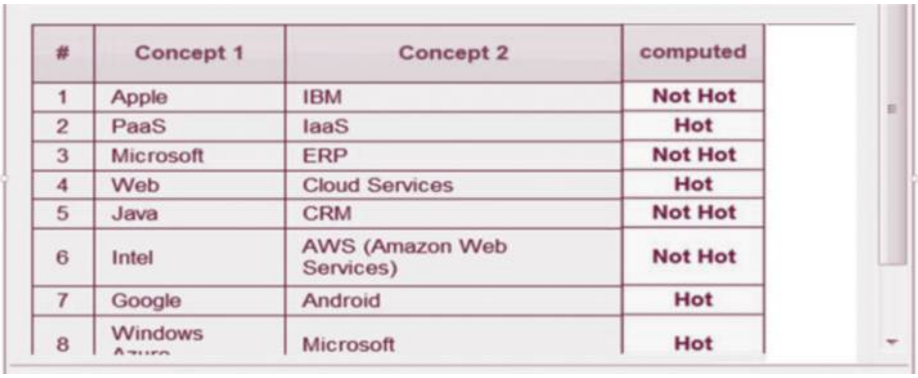
The concept map yielded by the presentation software module of the research instrument developed in this study allows a decision-maker to zoom into a specific concept, which then becomes the governing identification to which all other concepts would be related. This approach follows Novak and Gowin (1984) and Harnisch *et al.* (1994), who argue that

Table I Automatically identified co-occurring hot concepts (partial list)

Topic	*Co-occurring hot concepts*	
Cloud computing	PaaS[b]	IaaS[a]
	Salesforce	SaaS[c]
	Microsoft	Windows Azure
Grid computing	Parallel processing	Cloud services
	VMware	Parallel processing
	Cloud services	Virtualization
Semantic Web	Latent semantic	Search engines
	HTML 5	Flash
	Search engines	RDF[d]
Service-oriented architecture	Web services	ESB[e]
	Cloud	SaaS
	Amazon	Cloud computing
Business process management	SharePoint	Microsoft
	IBM	BPO[f]
	ERP[g]	Business intelligence

Notes: [a]IaaS (Infrastructure as a Service); [b]PaaS (Platform as a Service); [c]SaaS (Software as a Service); [d]RDF (Resource Description Framework); [e]ESB (Enterprise Service Bus); [f]BPO (Business Process Optimization); [g]ERP (Enterprise Resource Planning)

Figure 3 PTA results for SOA



#	Concept 1	Concept 2	computed
1	Apple	IBM	Not Hot
2	PaaS	IaaS	Hot
3	Microsoft	ERP	Not Hot
4	Web	Cloud Services	Hot
5	Java	CRM	Not Hot
6	Intel	AWS (Amazon Web Services)	Not Hot
7	Google	Android	Hot
8	Windows Azure	Microsoft	Hot

concept maps are best constructed if a single focal root concept guides the selection of concepts and their organization in clusters on the map. The UI was thus designed in the current study to limit the ability of users to link independent concepts, avoiding a hyper-complex spaghetti-like concept map and minimizing the risk of "seeing the trees but not the forest". The shift to a concept-centric analysis away from a map-centric analysis is a common exploratory method which is characteristic of many other related analytical tools. Also common in big data analytics applications is design of the UI aimed toward dynamic and highly focused knowledge discovery, providing a simple way to present the main and significant concepts of a highly dimensional space, as done in the present study. This feature is remarkably important, as suggested by Feldman *et al.* (1997), given that users do not necessarily know in advance the concepts and the associations that may interest them. Notably, this leads to simple "who, what, where" types of investigations in masses of textual data, which must be examined by a decision maker (Hutchins and Benham-Hutchins, 2010). Auxiliary visuals, such as graphs, motion charts and histograms, have been embedded in the UI component by implementing the Google Visualization API, using technologies such as HTML5/SVG with cross-browser compatibility. In addition to the UI component, based on a Web browser as a loosely coupled front-end, the research tool was designed to engage in rudimentary data exploration via middle-tier application of business logic and an intermediary connection interface to a database server.

Figure 4 demonstrates, for example, a concept map based on the research model and generated by the research instrument for Cloud Computing. Its focal point (see solid circle), visualized in a concept-centric view, is "Business Intelligence". All relevant concepts, such as "Dashboards", "Business Objectives", "Cloud Computing" and "IBM" (see arrows), are linked to "Business Intelligence". Note also that hot co-occurring concepts constructed by the PTA for the same focal point are visualized by red lines. One line connects "Business Intelligence" to "Business Objectives" (see dashed circle), and another line connects "Business Intelligence" to "Dashboards" (see dashed circle). Both red lines serve as a graphical notation to highlight hot co-occurring concepts in the augmented concept map, providing a simple way to present the main and significant concepts of a high-dimensional space.

6. Model validation and conclusion

In the validation phase, experts rated pairs of concepts in response to the Web-based survey, as partially depicted in Figure 5 (Part 2 of the questionnaire). Once complete questionnaires were received, the responses were combined to form complete evaluation matrices. To test the validity of the mechanism for detecting co-occurring hot concepts, predicative validity and inter-rater reliability were analyzed in a typical case-by-variable

Figure 4 Concept map for cloud computing based on a concept-centric view

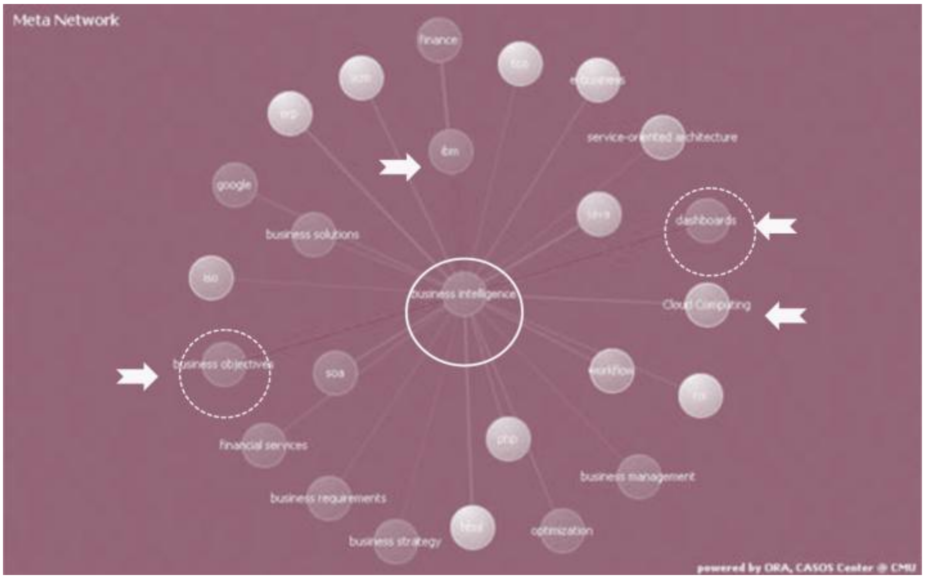
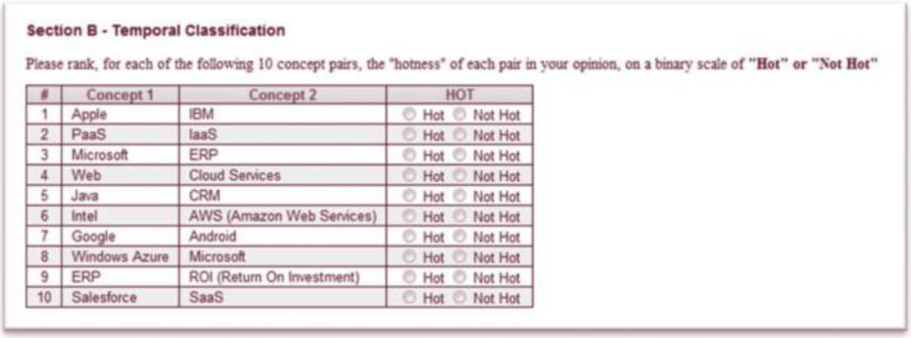


Figure 5 The Web-based questionnaire – PTA validation



Section B - Temporal Classification

Please rank, for each of the following 10 concept pairs, the "hotness" of each pair in your opinion, on a binary scale of **"Hot"** or **"Not Hot"**

#	Concept 1	Concept 2	HOT
1	Apple	IBM	○ Hot ○ Not Hot
2	PaaS	IaaS	○ Hot ○ Not Hot
3	Microsoft	ERP	○ Hot ○ Not Hot
4	Web	Cloud Services	○ Hot ○ Not Hot
5	Java	CRM	○ Hot ○ Not Hot
6	Intel	AWS (Amazon Web Services)	○ Hot ○ Not Hot
7	Google	Android	○ Hot ○ Not Hot
8	Windows Azure	Microsoft	○ Hot ○ Not Hot
9	ERP	ROI (Return On Investment)	○ Hot ○ Not Hot
10	Salesforce	SaaS	○ Hot ○ Not Hot

statistical data structure, with the cases being the raters (respondents) and the variables being their subjective ratings (responses). Responses from respondents (n = 136) showed high values of Fleiss Kappa reliability-of-agreement measures for all five assessed technologies (Table II), indicating general substantial agreement and average predictive validity higher than 85 per cent. The predictive validity is based on a percentage-

Table II Predicative validity and Fleiss Kappa coefficient

Topic	*Percentage agreement*	*Fleiss Kappa coefficient**
Business process management	85.52	0.725
Cloud computing	87.83	0.745
Grid computing	87.83	0.765
Semantic Web	86.07	0.728
Service-oriented architecture	81.85	0.698
Total	85.69	

Note: *Values in a range of 0.61-0.8 indicate substantial agreement (Landis and Koch, 1977) agreement measure, which seems to be the most prevalent method for calculating the consensus estimate $A = O/P$, where the agreement rate A is the division of the observed agreement O by the possible agreement P (Grayson and Rust, 2001). Missing values (e.g. *don't know* responses), which were at a significantly low rate (less than 0.01 per cent), were not included in the analysis.

The validation results show that the PTA construct developed in the current research is valuable and accomplishes this study's goals. It reflects the temporal distance between concepts by automatically detecting co-occurring hot and near-hot concepts. Moreover, the dynamic open textual data corpus at the basis of this detection is essential for improving the concept mapping process via temporal augmentation.

The innovation of this research is evident in terms of theory and practice, especially to management. The theoretical innovation lies in the development of an unsupervised temporal trend detection model, implementing a temporal construct for concept pairs via quantitative pair-wise temporal operators based on objective time properties derived from the time-tagged textual corpus. The practical innovation of the current research lies in the design, development and implementation of an innovative research model and automated instrument for knowledge mapping. These major innovations are evident in particular in the TAS arena and consistent with the needs of decision makers charged with identifying future technological trends when conducting evaluation of investment alternatives. Moreover, applying GA as a temporal and supremely updated source of Web data is indeed a major key element in the creation of a textual open and dynamic corpus. Finally, the managerial innovation of this work lies in the ability of the research model to extract from textual data an augmented concept map in a timely fashion, serving as a basis for deriving temporal insights and improving the visibility of knowledge in support of top executives' decision-making.

The contributions of the study, underscored by the growing attention to the big data phenomenon, are threefold:

1. generation of a time-augmented concept map;
2. development of a novel PTA approach to co-word analysis based on age, frequency and CCDR; and
3. identification of a way to distinguish between hot and established co-occurring concept pairs.

Toward accomplishing these contributions, we have proposed and developed a textual data-driven method which relies on established research areas, such as IE, TM, Web mining, concept mapping, visualization and KM.

The research model developed in this work may potentially improve concept mapping by adding temporal knowledge on the basis of a novel pair-wise analysis aimed at reflecting the temporal distance between co-occurring concepts. Observations of temporal indicators within concept maps play a major role in the dynamic nature of knowledge structures and knowledge representation. The developed textual data-driven research model provides a novel analytics framework for coping with the 3V attributes of big data architecture, which embraces and synthesizes the streaming of large bodies of unstructured temporal textual data from Web sources. The demonstrated and validated addition of temporal knowledge to concept mapping was proven valuable in detecting hot and near-hot technological trends in the IT arena. Evidently, the implemented computed PTA construct was found to be highly correlated with subjective ratings of experts (n = 136), exhibiting substantial reliability-of-agreement measures and average predictive validity above 85 per cent.

The current paper which focuses on temporal proximity measurement, and the previous one (Sasson *et al.*, 2015) which has focused on contextual proximity measurement, supplement one another. Further research is required to examine the transformative process of combining the relatedness proximity measurement and the PTA to yield augmented concept maps with these two components joined together. Future research should also validate the model extension for the automatic detection of co-occurring emerging concepts. While the current work can most probably be generalized to other domains, the challenge of generalizing the current approach beyond the five technological topics used for demonstration and validation must still be investigated. Nevertheless, two observations of the current research are worth noting and further researched:

1. some dynamically created Web pages are difficult to find or access due to lack of indexing by commercial search engines (i.e. "Invisible Web"); and
2. commercial search engines display language biases in site coverage, with English as a site language.

Moreover, future work should be devoted to dealing with near-real time analysis of the data collected over time in the corpus. Finally, the novel software modules developed to demonstrate and validate the research model have the potential to morph in the future into a decision support system.

References

Alavi, M. and Leidner, D.E. (1999), "Knowledge management systems: issues, challenges, and benefits", *Communications of the AIS*, Vol. 1 No. 2es, p. 1.

Alexa, M. (1997), "Computer-assisted text analysis methodology in the social sciences", *ZuMA-Arbeitsbericht*, Vol. 97.

Ashrafi, N., Xu, P., Kuilboer, J. and Koehler, W. (2006), "Boosting enterprise agility via IT knowledge management capabilities", *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Vol. 2, pp. 46a-46a.

Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A. and Buyya, R. (2015), "Big Data computing and clouds: trends and future directions", *Journal of Parallel and Distributed Computing*, Vol. 79, pp. 3-15.

Banko, M. and Brill, E. (2001), "Scaling to very large corpora for natural language disambiguation", *Proceedings of ACL-01*, Toulouse.

Besselaar, P.V.D. and Heimeriks, G. (2006), "Mapping research topics using wordreference co-occurrences: a method and an exploratory case study", *Scientometric*, Vol. 68 No. 3, pp. 377-393.

Blank, G.D., Pottenger, W.M., Kessler, G.D., Herr, M., Jaffe, H., Roy, S., Gevry, D. and Wang, Q. (2001), "Cimel: constructive, collaborative inquiry-based multimedia e-learning", *SIGCSE Bulletin*, Vol. 33 No. 3, p. 179.

Blank, G.D., Pottenger, W.M., Kessler, G.D., Roy, S., Gevry, D.R., Heigl, J.J., Sahasrabudhe, S.A. and Wang, Q. (2002), "Design and evaluation of multimedia to teach java and object oriented software engineering", *Proceedings of the 2002 American Society for Engineering Education Annual Conference & Exposition*, *Montreal*.

Bolshakov, I.A. and Gelbukh, A. (2004), *Computational Linguistics: Models, Resources, Applications*, Center for Computing Research (CIC) of the National Polytechnic Institute, The Economic Culture Fund Press.

Börner, K., Chen, C. and Boyack, K.W. (2003), "Visualizing knowledge domains", *Annual Review of Information Science and Technology*, Vol. 37 No. 1, pp. 179-255.

Boykin, S. and Merlino, A. (2000), "Machine learning of event segmentation for news on demand", *Communications of the ACM*, Vol. 43 No. 2, pp. 35-41.

Budanitsky, A. and Hirst, G. (2006), "Evaluating wordnet-based constructs of lexical semantic relatedness", *Computational Linguistics*, Vol. 32 No. 1, pp. 13-47.

Bun, K.K. and Ishizuka, M. (2003), "Topic extraction from news archive using TF*PDF algorithm", *Proceedings of the Third International Conference on Web Information Systems Engineering*, *Singapore*, pp. 73-82.

Bun, K.K. and Ishizuka, M. (2006), "Emerging topic tracking system in WWW", *Knowledge-Based Systems*, Vol. 19 No. 3, pp. 164-171.

Callon, M., Courtial, J.P. and Laville, F. (1991), "Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemsitry", *Scientometrics*, Vol. 22 No. 1, pp. 155-205.

Callon, M., Law, J. and Rip, A. (1986), *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*, Macmillan Press.

Carbonell, J., Yang, Y., Lafferty, J., Brown, R.D., Pierce, T. and Liu, X. (1999), "CMU report on TDT-2: segmentation, detection and tracking", *Proceedings of the DARPA Broadcast News Workshop*, *Herndon, Virginia*, pp. 117-120.

Chen, C. (2006), "CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature", *The Journal of the American Society for Information Science and Technology*, Vol. 57 No. 3, pp. 359-377.

Chen, H., Chiang, R.H. and Storey, V.C. (2012), "Business intelligence and analytics: from big data to big impact", *MIS Quarterly*, Vol. 36 No. 4, pp. 1165-1188.

Chen, K.Y., Luesukprasert, L. and Chou, S. (2007), "Hot topic extraction based on timeline analysis and multidimensional sentence modelling", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19 No. 8, pp. 1016-1025.

Chen, W. and Chundi, P. (2011), "Extracting hot spots of basic and complex topics from time stamped documents", *Data and Knowledge Engineering*, Vol. 70 No. 7, pp. 642-660.

Coleman, D.J. and Li, S. (1999), "Developing a groupware-based prototype to support geomatics production management", *Computers, Environment and Urban Systems*, Vol. 23 No. 4, pp. 315-331.

Courseault, C.R. (2004), "A text mining framework linking technical intelligence from publication databases to strategic technology decisions", *PhD Dissertation*, GA Institute of Technology.

Courtial, J.P. (1994), "A coword analysis of scientometrics", *Scientometrics*, Vol. 3, pp. 251-260.

Courtney, J., Croasdell, D. and Paradice, D. (1997), "Lockean inquiring organizations: guiding principles and design guidelines for learning organizations", *Proceedings of the 1997 America's Conference on Information Systems*, *Indianapolis*.

Cunningham, H., Bontcheva, K. and Li, Y. (2005), "Knowledge management and human language: crossing the chasm", *Journal of Knowledge Management*, Vol. 9 No. 5, pp. 108-131.

Das, T.K. and Kumar, P.M. (2013), "Big data analytics: a framework for unstructured data analysis", *International Journal of Engineering Science & Technology*, Vol. 5 No. 1, p. 153.

Dasgupta, S., Pal, P., Mazumdar, C. and Bagchi, A. (2015), "Resolving authorization conflicts by ontology views for controlled access to a digital library", *Journal of Knowledge Management*, Vol. 19 No. 1, pp. 45-59.

Debortoli, S., Müller, O., Junglas, I. and vom Brocke, J. (2016), "Text mining for information systems researchers: an annotated topic modeling tutorial", *Communications of the Association for Information Systems (CAIS*), Vol. 39 No. 1.

Desikan, P. and Srivastava, J. (2004), "Mining temporally evolving graphs", *The Proceedings of the Sixth WEBKDD Workshop in Conjunction with the 10th ACM SIGKDD Conference*, *Vol*. 22.

Ding, Y., Chowdhury, G.G. and Foo, S. (2001), "Bibliometric cartography of information retrieval research by using co-word analysis", *Information Processing and Management*, Vol. 37, pp. 817-842.

Ding, Y., Chowdhury, G.G., Foo, S. and Qian, W. (2000), "Bibliometric information retrieval systems (BIRS): a web search interface utilizing bibliometric research results", *Journal of the American Society for Information Science (JASIS*), Vol. 51 No. 13, pp. 1190-1204.

Dixon, M. (1997), *An Overview of Document Mining Technology*. *Computer Based Learning Unit*, University of Leeds.

Dzone Software (2013), "10 knowledge management challenges managers face today", available at: www.dzonesoftware.com/blog/10-challenges-knowledge-managers-face-today (accessed 14 October 2016).

Feldman, R., Klbsgen, W., Ben-Yehuda, Y., Kedar, G. and Reznikov, V. (1997), "Pattern based browsing in document collections", Principles of Data Mining and Knowledge Discovery: First European Symposium, PKDD'97.

Gandomi, A. and Haider, M. (2015), "Beyond the hype: big data concepts, methods, and analytics", *International Journal of Information Management*, Vol. 35 No. 2, pp. 137-144.

Gloet, M. and Terziovski, M. (2004), "Exploring the relationship between knowledge management practices and innovation performance", *Journal of Manufacturing Technology Management*, Vol. 15 No. 5, pp. 402-409.

Goorha, S. and Ungar, I. (2010), "Discovery of significant emerging trends", *The Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Washington, DC*, pp. 57-64.

Grayson, K. and Rust, R. (2001), "Interrater reliability", *Journal of Consumer Psychology*, Vol. 10 No. 1, pp. 71-73.

Grobelnik, M., Mladenic, D. and Milic-Frayling, N. (2000), "Text mining as integration of several related research areas: report on KDD'2000 workshop on text mining", *SIGKDD Explorations*, Vol. 2 No. 2, pp. 99-102.

Halsius, F. and Lochen, C. (2001), "Assessing technological opportunities and threats: an introduction to technology forecasting", *Division of Industrial Marketing*, Lulea University of Technology.

Harnisch, D.L., Sato, T., Zheng, P., Yamagi, S. and Connell, M. (1994), *Concept Mapping Approach and its Applications in Instruction and Assessment*, The American Educational Research Association.

Hauber, R.P., Vesmarovich, S. and Dufour, L. (2012), "The use of computers and the internet as a source of health information for people with disabilities", *Rehabilitation Nursing*, Vol. 27 No. 4, pp. 142-145.

Havre, S., Hetzler, E., Whitney, P. and Nowell, L. (2002), "Themeriver: visualizing thematic changes in large document collections", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8 No. 1, pp. 9-20.

He, Q. (1999), "Knowledge discovery through co-word analysis", *Library Trends*, Vol. 48, pp. 133-159.

Hutchins, C.E. and Benham-Hutchins, M. (2010), "Hiding in plain sight: criminal network analysis", *Computational and Mathematical Organization Theory*, Vol. 16 No. 1, pp. 89-111.

Intezari, A. and Pauleen, D. (2013), "Looking beyond knowledge: can wisdom be nurtured in management programs?", *Academy of Management*, Vol. 2013 No. 1, p. 13468.

Jonassen, D.H. and Grabowski, B.L. (1993), *Handbook of Individual Differences: Learning & Instruction*, Lawrence Earlbaum Associates, Hillsdale, NJ.

Kasiviswanathan, S.P., Melville, P., Banerjee, A. and Sindhwani, V. (2011), "Emerging topic detection using dictionary learning", *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, *Glasgow*, pp. 745-754.

Keller, F. and Lapata, M. (2003), "Using the web to obtain frequencies for unseen bigrams", *Computational linguistics*, Vol. 29 No. 3, pp. 459-484.

Khasseh, A.A. and Mokhtarpour, R. (2016), "Tracing the historical origins of knowledge management issues through Referenced Publication Years Spectroscopy (RPYS)", *Journal of Knowledge Management*, Vol. 20 No. 6.

King, J. (1987), "A review of bibliometric and other science indicators and their role in research evaluation", *Journal of Information Science*, Vol. 13, pp. 261-276.

Kontostathis, A., Galitsky, L.M., Pottenger, W.M., Roy, S. and Phelps, D.J. (2004), "A survey of emerging trend detection in textual data mining", in Berry, M. (Ed.), *Survey of Text Mining: Clustering, Classification, and Retrieval*, Springer.

Kumaran, G. and Allan, J. (2004), "Text classification and named entities for new event detection", *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *Sheffield*, pp. 297-304.

Law, J. and Whittaker, J. (1992), “Mapping acidification research: a test of the co-word method”, *Scientometrics*, Vol. 23, pp. 417-461.

Leake, D., Maguitman, A. and Canas, A. (2001), “Assessing conceptual similarity to support concept mapping”, *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, *Pensacola Beach, Florida*, pp. 172-186.

Lee, M.R. and Chen, T.T. (2012), “Revealing research themes and trends in knowledge management: from 1995 to 2010”, *Knowledge-Based Systems*, Vol. 28, pp. 47-58.

Lee, S., Baker, J., Song, J. and Wetherbe, J.C. (2010), “An empirical comparison of four text mining methods”, *Proceeding of the 43rd Hawaii International Conference of System Sciences (HICSS)*, *Honolulu, HI*, pp. 1-10.

Lent, B., Agrawal, R. and Srikant, R. (1997), “Discovering trends in text databases”, *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD)*, *Newport Beach, California*, pp. 227-230.

Leydesdorff, L. and Hellsten, I. (2006), “Measuring the meaning of words in contexts: an automated analysis of controversies about ‘monarch butterflies’, ‘franken foods’ and ‘stem cells’”, *Scientometrics*, Vol. 67 No. 2, pp. 231-258.

Li, Y. and Zhong, N. (2004), “Web mining model and its applications for information gathering”, *Knowledge-Based Systems*, Vol. 17 No. 5, pp. 207-217.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. and Barton, D. (2012), “Big data: the management revolution”, *Harvard Business Review*, Vol. 90 No. 10, pp. 61-67.

Ma, J. and Perkins, S. (2003), “Online novelty detection on temporal sequences”, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Washington, DC*, pp. 613-618.

Marko, G. and Mladeni, D. (2005), “Automated knowledge discovery in advanced knowledge management”, *Journal of Knowledge Management*, Vol. 9 No. 5, pp. 132-149.

Matsuo, Y., Sakaki, T., Uchiyama, K. and Ishizuka, M. (2006), “Graph-based word clustering using a web search engine”, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, *Association for Computational Linguistics*, pp. 542-550.

Mei, Q. and Zhai, C.X. (2005), “Discovering evolutionary theme patterns from text: an exploration of temporal text mining”, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, *Chicago, IL*, pp. 198-207.

Mendez-Torreblanca, A., Montes-y-Gomez, M. and Lopez-Lopez, A. (2002), “A trend discovery system for dynamic web content mining”, *Representations*, Vol. 7 No. 8.

Morinaga, S. and Yamanishi, K. (2004), “Tracking dynamics of topic trends using a finite mixture model”, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *New York, NY*, pp. 811-816.

Müller, O., Junglas, I., vom Brocke, J. and Debortoli, S. (2016), “Utilizing big data analytics for information systems research: challenges, promises and guidelines”, *European Journal of Information Systems*, Vol. 25 No. 4.

Novak, J.D. and Gowin, D. (1984), *Learning How to Learn*, Cambridge University Press, New York, NY.

Nowell, L.T., France, R.K. and Hix, D. (1997), “Exploring search results with envision”, *Proceeding of Computer Human Interaction-CHI’*, Vol. 97, pp. 14-15.

Pattuelli, M.C. and Miller, M. (2015), “Semantic network edges: a human-machine approach to represent typed relations in social networks”, *Journal of Knowledge Management*, Vol. 19 No. 1, pp. 71-81.

Plotnick, E. (1997), “Concept mapping: a graphical system for understanding the relationship between concepts”, *An ERIC Digest*, Clearinghouse on Information and Technology.

Porter, A.L. and Cunningham, S.W. (2005), *Tech Mining – Exploiting New Technologies for Competitive Advantage*, John Wiley & Sons Publishers, Hoboken, NJ.

Porter, A.L. and Detampel, M.J. (1995), “Technology opportunities analysis”, *Technological Forecasting and Social Change*, Vol. 49 No. 3, pp. 237-255.

Pottenger, W.M. and Yang, T. (2001), “Detecting emerging concepts in textual data mining”, in Berry, M.W. (Ed.), *Computational Information Retrieval*, SIAM, Vol. 106, pp. 1-17.

Power, D.J. (2016), “Data science: supporting decision-making”, *Journal of Decision Systems*, Vol. 25 No. 4, pp. 1-12.

Prado, H.A. and Ferneda, E. (2007), *Emerging Technologies of Text Mining: Techniques and Applications. Information Science Reference*, Hershey, New York, NY.

Raan, A. and Tijssen, R. (1993), “The neural net of neural network research: an exercise in bibliometric mapping”, *Scientometrics*, Vol. 26 No. 1, pp. 169-192.

Raghupathi, W. and Raghupathi, V. (2014), “Big data analytics in healthcare: promise and potential”, *Health Information Science and Systems*, Vol. 2 No. 1, p. 3.

Rajaraman, K. and Tan, A.H. (2001), “Topic detection, tracking, and trend analysis using self-organizing neural networks”, Advances in Knowledge Discovery and Data Mining, London, pp. 102-107.

Rapp, R. (2002), “The computation of word associations: comparing syntagmatic and paradigmatic approaches”, *Proceedings of the 19th International Conference on Computational Linguistics*, *Taipei*, pp. 1-7.

Ribiere, V. and Walter, C. (2013), “10 years of KM theory and practices”, *Knowledge Management Research & Practice*, Vol. 11 No. 1, pp. 4-9.

Rousseau, D.M. (1979), “Assessment of technology in organizations: closed versus open systems approaches”, *The Academy of Management Review*, Vol. 4 No. 4, pp. 531-542.

Roy, S., Gevry, D. and Pottenger, W.M. (2002), “Methodologies for trend detection in textual data mining”, *Proceedings of the Textmine’02 Workshop, Second SIAM International Conference on Data Mining*, *Arlington*, p. 58.

Russell, A.W., Vanclay, F.M. and Aslin, H.J. (2010), “Technology assessment in social context: the case for a new framework for assessing and shaping technological developments”, *Impact Assessment and Project Appraisal*, Vol. 28 No. 2, pp. 109-116.

Salton, G. (1988), *Automatic Text Processing*, Addison-Wesley Publishing Company.

Salton, G., Wong, A. and Yang, C.S. (1975), “A vector space model for automatic indexing”, *Communications of the ACM*, Vol. 18 No. 11, pp. 613-620.

Sasson, E., Ravid, G. and Pliskin, N. (2015), “Improving similarity constructs of relatedness proximity: toward augmented concept maps”, *Journal of Informetrics*, Vol. 9, pp. 618-628.

Schomm, F., Stahl, F. and Vossen, G. (2013), “Marketplaces for data: an initial survey”, *ACM SIGMOD Record*, Vol. 42 No. 1, pp. 15-26.

Sedighi, M. and Jalalimanesh, A. (2014), “Mapping research trends in the field of knowledge management”, *Malaysian Journal of Library & Information Science*, Vol. 19 No. 1, pp. 71-85.

Speel, P., Shadbolt, N.R., Vries, W.D., Dam, P.V. and O’Hara, K. (1999), “Knowledge mapping for industrial purposes”, *Twelfth Workshop on Knowledge Acquisition, Modelling Management (KAW’99*), Vol. 2 No. 7.

Subasic, I. and Berendt, B. (2010), *From Bursty Patterns to Bursty Facts: The Effectiveness of Temporal Text Mining for News*, Citeseer.

Swan, R. and Jensen, D. (2000), “Timemines: constructing timelines with statistical models of word usage”, KDD-2000 Workshop on Text Mining, Boston, MA.

Swanson, E.B. and Ramiller, N.C. (2004), “Innovating mindfully with information technology”, *MIS Quarterly*, Vol. 28 No. 4, pp. 553-583.

Tredinnick, L. (2009), “Complexity theory and the web”, *Journal of Documentation*, Vol. 65 No. 5, pp. 797-816.

Van den Berg, C. and Popescu, I. (2005), “An experience in knowledge mapping”, *Journal of Knowledge Management*, Vol. 9 No. 2, pp. 123-128.

Varian, H.R. (2006), *The Economics of Internet Search*, University of California at Berkeley.

Waltman, L., van Eck, N.J. and Noyons, E.C. (2010), “A unified approach to mapping and clustering of bibliometric networks”, *Journal of Informetrics*, Vol. 4 No. 4, pp. 629-635.

Wang, Y., Kung, L., Wang, C., Yu, W. and Cegielski, C. (2014), “Developing a big data-enabled transformation model in healthcare: a practice based view”, *Proceedings of Thirty Fifth International Conference on Information Systems*, *Auckland*.

Wayne, C.L. (1997), “Topic detection and tracking (TDT)”, *On Workshop held at the University of Maryland*, Vol. 27, pp. 28-30.

Wilks, Y. (1997), *Information Extraction as a Core Language Technology, Lecture Notes in Computer Science*, Vol. 1299, pp. 1-9.

Wong, P.C., Cowley, W., Foote, H., Jurrus, E. and Thomas, J. (2000), “Visualizing sequential patterns for text mining”, *Information Visualization*, InfoVis 2000, IEEE Symposium, IEEE, pp. 105-111.

Zhang, K., Zi, J. and Wu, L.G. (2007), “New event detection based on indexing-tree and named entity”, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *New York, NY*, pp. 215-222.

Zhuge, H. (2015), “Mapping big data into knowledge space with cognitive cyber-infrastructure”, *arXiv preprint arXiv:1507.06500*.

Zikopoulos, P. and Eaton, C. (2012), *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media.

Zweigenbaum, P., Jacquemart, P., Grabar, N. and Habert, B. (2001), *Building a Text Corpus for Representing the Variety of Medical Language. Studies in Health Technology and Informatics*, IOS Press, pp. 290-294.

Further reading

Landis, J.R. and Koch, G.G. (1977), “The constructment of observer agreement for categorical data”, *Biometrics*, pp. 159-174.

Corresponding author

Gilad Ravid can be contacted at: rgilad@bgu.ac.il