Measuring Temporal and Contextual Proximity

Big Text-Data Analytics in Concept Maps

Elan Sasson
Industrial Engineering
Tel Aviv University
Tel Aviv, Israel
e-mail: sasson.elan@gmail.com

Gilad Ravid
Industrial Engineering & Management
Ben-Gurion University of the Negev
Beer-Sheva, Israel
e-mail: rgilad@bgu.ac.il

Nava Pliskin
Industrial Engineering & Management
Ben-Gurion University of the Negev
Beer-Sheva, Israel
e-mail: pliskinn@bgu.ac.il

***Abstract* - Despite being important, time and context have yet to be formally incorporated into the process of visually representing the temporal and contextual proximity between keywords in a concept map. In response to the context and time challenges, this study improves automated conventional concept mapping by measuring the temporal and contextual distance between pairs of co-occurring concepts. After generating a conventional concept map, it is temporally and contextually augmented in this work by applying an unsupervised temporal trend detection algorithm and a novel measure of contextual proximity. This proposed approach is demonstrated and validated without loss of generality for a spectrum of information technologies, showing that the resulting assessments of temporal and contextual proximity are highly correlated with subjective assessments of experts. The contribution of this work is emphasized and magnified against the current growing attention to big data analytics in general and to big text-data analytics in particular.**

Keywords- Concept map;Temporal proximity;Contextual Proximity; Big text-data analytics

I. INTRODUCTION

Managing and gaining insights from the vast amount of textual data on the web is challenging [3] and, yet, mining unstructured web data can help organizations gain insights needed for decision support [62]. The 3Vs challenges [19;40] are associated with acquiring, managing and analyzing data sets characterized by extensive Volume from gigabytes to terabytes, Variety from semi-structured to unstructured, and Velocity from batch to streaming [52]. Responding to the 3Vs challenges, *big data analytics* encompasses a new paradigm for utilizing big data sources while using advanced and sophisticated data analysis approaches and tools [44]. For example, text mining (TM) and information extraction (IE) are used for capturing, processing, analyzing and visualizing a big corpus of unstructured textual data drawn from the web [9;53;72]. Such information integration and analytical capabilities aim to reduce the mined text data to a relatively small number of keywords that can visualized as concept map.

This work harnesses big data analytics to go beyond the automatic generation of a concept map, based on massive amounts of textual web data, to quantify the temporal and contextual distance between two concepts on the map. Via a network visualization application, conventional concept mapping involves producing merely a graph of keywords based on co-word analysis of concepts extracted from a big corpus of unstructured textual data [48]. The reason for going in this study beyond conventional concept mapping is that the generated visual structure that represents keywords in a concept map leaves decision makers wondering how closely concept pairs are *temporally* and *contextually* related.

To better support decision making, this work proposes to augment concept maps by quantifying the distance between two concepts temporally and contextually, combining pair-wise temporal analysis with measurement of the extent to which concept pairs on the map are contextually related (proposed previously [61]). Given the fast pace of the global business arena, where innovations are occurring at increasing speed and life cycles are considerably shorter, both temporally and contextually augmenting concept maps is essential in support of decision-making processes such as technology assessment.

The research presented in the current paper makes innovative theoretical and practical contributions beyond the literature reviewed in the next section. From the theoretical perspective, this study is the first attempt to model the addition of both temporal and contextual insights to conventional concept maps. From the practical perspective, based on the software modules developed to obtain its results, this study contributes to the development of a managerial decision-support tool for managers dealing with the absence of temporal and contextual knowledge in concept maps. The modeling approach is described in Section III and the method used for demonstration and validation is outlined in Section IV. The concluding Section VI follows the results in Section V.

II. LITERATURE REVIEW

To make well-informed decisions, Bolshakov & Gelbukh [7] acknowledge that decision makers must read an enormous quantity of web text when faced with the challenge of identifying emerging technologies with the greatest potential just in time [14;23;56;58]. Lee, Baker, Song & Wetherbe [36] assert that manual analysis of unstructured textual data is never practical while Dixon [17] refers to the impossibility of processing massive quantities of textual information.

The objective of text mining (TM) is to exploit information contained in textual documents in various ways, including associations among text objects like concepts on the other [21]. Linguistics-based TM is guided by natural language processing (NLP) and involves information extraction (IE). IE involves extracting named entities and

978-1-5090-5753-5/16 $31.00 © 2016 IEEE
DOI 10.1109/APWC.on.CSE.2016.31

factual assertions from a textual corpus [74], transforming the unstructured document space to the structured concept space toward applying co-word analysis.

The following definitions reflect the literature on co-word analysis [10;15;55]. A concept is defined as a logical semantically cohesive unit of text. A co-word/co-occurrence relation is defined as a one-to-many mapping that associates each concept with a list of related concepts ranked quantitatively by relatedness similarity. A concept map is defined as a dynamic graphical map that visually presents concepts and relevant relationship clusters as an undirected graph G= (V, E) consisting of a set of vertices V and edges E. Within an extensive body of literature on co-occurrence analysis (e.g., [11;15;38]), co-word analysis is praised as a proven quantitative tool for knowledge discovery [18;26]. Yet, concept maps that dedicated TM applications generate from text documents automatically [35] do not reflect the temporal and contextual distance between concept pairs.

Time stamps, as publication dates, like the underlying temporal and evolutionary structure are frequently ignored by TM software tools [41]. On the other hand, Temporal Text Mining (TTM), concerned with discovering temporal patterns [49] in textual data [8;12; 39;43;54], introduces the time dimension into web mining to reflect current trends and help predict future ones regarding emerging and hot topics [16]. Within TTM, emerging trend detection (ETD) applications are mentioned as automating the discovery of emerging and hot trends [51;65]. However, current ETD systems have three limitations that the current work aims to overcome. First, most ETD systems require a user to subjectively finalize concept classification [5;6;12;25;47;50;57]. Second, previous ETD studies [6;16] focus on unary-based analysis for detecting trends at the risk of missing important synchronized temporal information. Finally, most ETD systems use a unitary static monolithic text corpus from human-maintained indexed databases [37;65;75;34;41;77;63;13], as INSPEC or TDT, thus risking limited diversity, variety and richness and imposing major drawbacks as data coverage and indexer effect [1;4;28]. For temporal augmentation, these limitations are overcome here by 1) using objective temporal operators, 2) detecting temporal attributes of co-occurring concepts at a two-item level, and 3) building an open dynamic corpus of relevant documents without the need to subjectively evaluate the cardinality or the authority of the feed sources.

For contextual augmentation, this work harnesses bibliometrics methods which use information such as word counts, date, word co-occurrence and citation to track activity in a subject area applying analytics and statistics [29;42]. The number of publications is often used as a bibliometric indicator to measure and interpret scientific advances [45;69;70;71]. Also used in bibliometric analysis are co-occurrences, potentially providing information on emerging technologies [50]. The explosion of web has led to the term 'webometrics' has been coined [2], for applying bibliometrics to web data. For example, web impact assessment is defined as counting how often ideas are mentioned online, using the reported hit count estimate (HCE) which appears near the top of the results page in a commercial search engine, harnessing the total number of search results available as impact evidence [67]. In many webometric studies [27;66;68;76], the HCE is the most commonly used indicator. Moreover, bibliometrics can be applied in combination with network analysis such as in a social setting [64]. Thus, Guston and Sarewitz [22] argue that successful concept mapping can be accomplished through TM and bibliometric approaches, as already discussed in a series of papers [30;31;32;33]. Following arguments in favor of exploiting bibliometrics and TM to support decision-making, the assumption underlying the research model presented next is that the HCE is a reliable measure of contextual proximity.

III. Research Model

At the core of the proposed research is augmentation of concept mapping via two expansions: A) pair-wise temporal analysis for the links between every concept pair on the map, using automatic trend detection (PTA) algorithm and B) measuring relatedness proximity between two concepts by means of webometrics method, using extended co-word analysis. Modeling both expansions follows two preparatory methods. The first method involves gradually building a corpus of unstructured textual data from diverse web-based sources about a target topic by using the Google Alerts (GA) content change-detection and notification service that automatically notifies subscribers when new web content matches a set of search terms associated with the target topic. The second method involves generating a concept map via co-word analysis after TM and IE extract the concepts from each document (HTML page) in the corpus to (i.e., keywords).

A. Pair-wise Temporal Analysis

To determine concept categorization based on the time dimension, whether hot or not, two quantitative pair-wise temporal operators are defined. The *Age* of relevant documents where concepts co-occur, represents the level of "hotness" since a concept pair is considered hot if its concepts are semantically richer at a later time than at an earlier time [20;51]. The *Frequency* rate (i.e., activity ratio) of publishing relevant documents where concepts co-occur in a given time interval, represents the level of "activeness" for tracking the recurrence of known events, which is one of the five types of tasks described in the Topic Detection and Tracking (TDT) project [12;73].

For expressing *Age* as old or young, the current study enhances the scalability of the Vector Space Model (VSM) by conducting pair-wise temporal analysis that exploits the cosine similarity measure [59;60]. 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 chronologically ordered accordingly. The cosine similarity measure is then applied to the 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 (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, and close to 0 if $\vec{x}$ and $\vec{y}$ have little in common, indicating an old temporal relationship between the co-occurring concepts. 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 temporal operator of the pair-wise temporal analysis ('b' in Figure 1), is accordingly, defined as (1):

$$Age(\vec{x},\vec{y}) = \frac{\sum_{k=1}^{n} x_k\, y_k}{\sqrt{(\sum_{k=1}^{n} {x_k}^2)*(\sum_{k=1}^{n} {y_k}^2)}} \qquad (1)$$

Where $\vec{y} = (y_1, \ldots, y_n);\ \vec{x} = (x_1, \ldots, x_n);\ x_{1\ldots n} = 1$

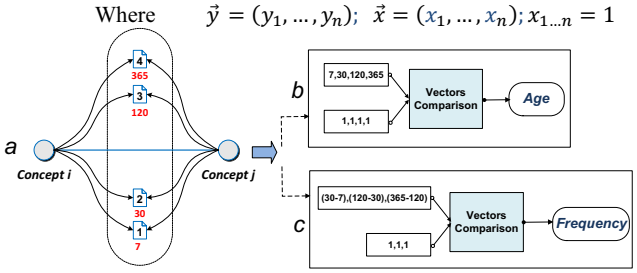

Figure 1 Concept co-occurrences demonstrated for 4 different documents

Similarly, *Frequency* ('c' in Figure 1), is 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 $\vec{y}$, yielding minimal values since high publication ratio indicates an imaginary notion of documents published on a daily basis. The cosine similarity measure is then applied to a reference Vector $\vec{x}$ with $n-1$ dimensions, where all coordinates are 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). $Frequency(\vec{x},\vec{y})$ is defined as (2):

$$Frequency(\vec{x},\vec{y}) = \frac{\sum_{k=1}^{n}(x_k - x_{k-1})\,(y_k - y_{k-1})}{\sqrt{(\sum_{k=1}^{n}(x_k - x_{k-1})^2)*(\sum_{k=1}^{n}(y_k - y_{k-1})^2)}} \qquad (2)$$

Where $(x_k - x_{k-1}) \stackrel{\text{def}}{=} 1$

Since a time-tagged corpus is under investigation here, both operators should be re-evaluated constantly to reflect progress of time. Daily recalculation is advised since 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 detect co-occurring hot concepts, in addition to the already-calculated two quantitative pair-wise temporal operators, a third operator - *CCDR* (concept co-occurring documents ratio) is calculated, reflecting the capacity of the number of documents in which hot concepts potentially co-occur, rescaled to the range 0 to 1 using observed *min max* values. The three operators are then weighted using $\omega_1, \omega_2, \omega_3$ in the following linear equation (3):

$$\underbrace{PTA}_{hot}{}_{ij} = \left(\omega_1 * Age_{ij} + \omega_2 * Frequency_{ij} + \omega_3 * CCDR_{ij}\right) \qquad (3)$$

Finally, to complete detection of co-occurring hot concepts, classification according to the PTA value is implemented in the research model by determining that Concepts j and i are co-occurring hot concepts if $PTA_{ij} \geq \tau$. The threshold value τ may be set between 0 and 1 either manually or empirically [29;65]. In either case, it is quite common in practice to allow manual calibration of the threshold values that is acceptable to the decision maker, with a τ value close to 1 reflecting the most rigorous classification requirement.

B. Relatedness Proximity Measurement via Co-Word Anlysis

To meet the challenge of measuring the contextual distance between co-occurring concept pairs, webometric co-word analysis is used to measure co-occurrence relatedness proximity via the Similarity Link Value (SLV), also known as Equivalence Index (E), defined by Callon et al. [10]. In this definition, C_i and C_j respectively count number of documents in which each term appears and C_{ij} is the number of documents in which both terms i and j co-occur:

$$SLV_{ij} = \frac{{C_{ij}}^2}{C_i * C_j},\ 0 \leq SLV_{ij} \leq 1,\ C_{ij} = C_{ji} \geq 0 \qquad (4)$$

The contextual part of the research model includes three steps (Figure 2). First, for each relation between Concept i and Concept j, an a-priori $aSLV_{ij}$ is calculated by using NLP-based TM to complete the IE task, extracting significant semantic concepts (*named entities* such as person, company, location, product) from the time-tagged textual corpus (e.g., TXT files, PDF, HTML files etc.) . Second, a bibliometric $bSLV_{ij}$ based on web counts (HCEs) is calculated for each concept pair via webometric queries to a web search engine about Concept i, Concept j and their conjunctive Concept i+ Concept j, using the AND Boolean operator. Third, both a-priori $aSLV_{ij}$ and bibliometric $bSLV_{ij}$ are synthesized into a combined $cSLV_{ij}$ for each concept pair, measuring relatedness proximity for the Concept-pair i, j based on the following additive formula:

$$cSLV_{ij} = f\left(aSLV_{ij}, bSLV_{ij}\right) \approx \left(\frac{(aC_{ij} + bC_{ij})^2}{(aC_i + bC_i) * (aC_j + bC_j)}\right) \qquad (5)$$

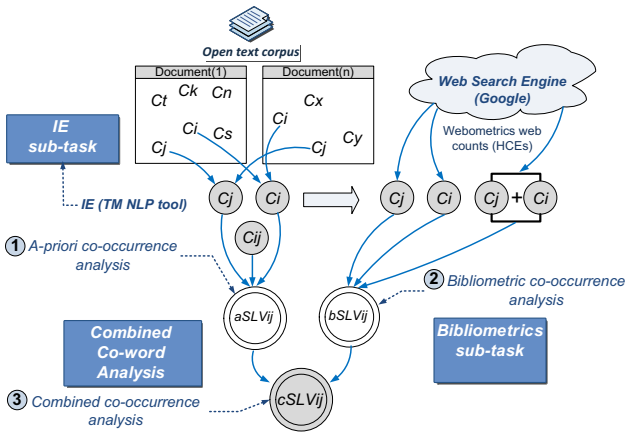

Figure 2 A conceptual workflow describing the co-word analysis process

The basic premise underlying the measuring of relatedness proximity based on webometric web counts is that the outcome of the combined co-occurrence analysis yields a more accurate, compact and valuable concept map for identification and selective extraction of significant concept co-occurrences. Upon using the combined co-word analysis proposed here, weak or strong signals which are normally detected in the conventional a-priori co-word analysis are improved via weighted synthesis with the co-occurrence values obtained by applying bibliometrics. Thus, conventional co-word analysis is enhanced by assimilating the knowledge background available on the web in the form of webometric web counts.

IV. Methodology

A web-based research instrument was developed to demonstrate and validate the research model developed in the current study and presented in Section III. The instrument is divisible into the following six main stages (Figure 3):

a) *Temporal GAs collection* tasks involve collecting a repository of Google Alert (GA) email updates, 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 in Figure 3 depict this stage.

b) *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 in order 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 in Figure 3 depict this stage.

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

d) *Post-processing* analysis tasks include all procedures and methods required for conducting the relatedness proximity measurement and the pair-wise temporal analysis toward augmented mapping. Steps 8 to 15 in Figure 3 depict this stage.

e) *Presentation* tasks and browsing functionality include easy-to-use, point-and-click, and browser-based user interface and listing capabilities. The presentation layer components of the research instrument display the knowledge map with references to co-occurrence weights calculated at each step, as well as the detected co-occurring hot concept. Step 16 in Figure 3 depicts this stage.

f) Evaluation tasks carried out by the decision maker while interpreting the acquired results are not depicted in Figure 3.

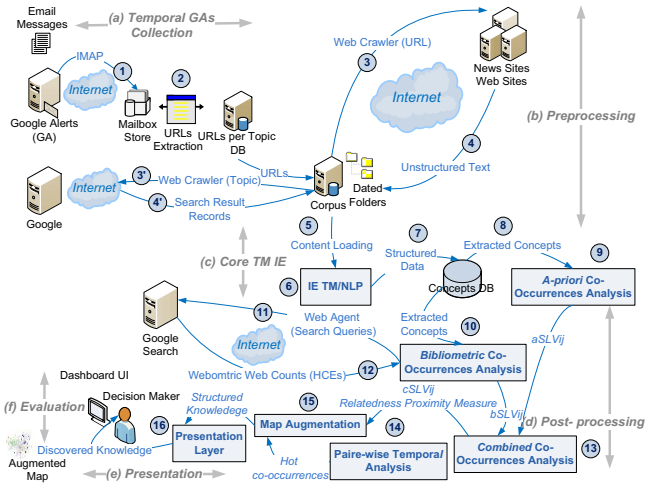

Figure 3 The stages and tasks of the research instrument

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 is gradually built by the research instrument, collecting textual data from diverse web-based sources to fully leverage the potential of hotness discovery. A corpus of unstructured textual data is gradually created about a target topic, using the Google Alerts (GA) service. 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 (converted to text files toward concept mapping). Since the GA service determines source validity, this method of corpus building allows collecting relevant documents without the need to subjectively evaluate the cardinality or the authority of the feed sources. To accomplish concept mapping in this study via the IE process, NLP-based TM analysis was applied to the text files in the time-tagged corpus, using IBM's SPPS/PASW Text Analytics Version13 (former SPSS Text Mining Modeler) and AlchemyAPI. Although each of these tools provides all the functions necessary for the IE process, both

were used for rigor and robustness. This IE process employs 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.

Five IT topics were chosen to demonstrate and validate the derivation of temporal and contextual proximity according to the research model. The demonstration was accomplished for each IT by conducting the PTA and by measuring the extended relatedness proximity. In the final step, technology savvy decision makers explored the generated augmented concept map via the graphical user interface component of the research instrument. This visualization tool for representing relationships among concepts is based on the dynamic network analysis software application Meta Network ORA Network Visualizer.

Validation of the proposed model began with validating the relatedness proximity measurement and continued with validating the PTA using the match-to-expert scoring approach for each of the five target ITs. Respondents were recruited to respond for each IT to a simple survey questionnaire distributed over the web internationally. Using two major sources of domain experts for recruiting, LinkedIn and a leading global IT consulting firm, yielded a total of n = 136 respondents. The survey questionnaire was comprised of two major parts along with respondent demographic data. The first part, for relatedness proximity validation, included questions about 20 pairs of co-occurring concepts. Each respondent was asked to determine for each pair weighted relationship scores, using visual analogue scale (VAS) as a continuous evaluation device. The second part, for temporal proximity validation, included questions about randomly-selected10 pairs of co-occurring concepts, asking each respondent to rate on a binary scale whether a pair of co-occurring concepts is hot or not.

V. MODEL DEMONSTRATION AND VALIDATION

To demonstrate and validate the model, a temporal corpus was built via collection of GA messages throughout 190 days in 2011, allowing reflection four years later. At the time of data collection, the Cloud Computing technology was much-hyped, the Grid Computing technology was expected to be substituted by cloud, the Business Process Management (BPM) technology attracted a lot of attention, the new Semantic Web technology was regarded as particularly promising, and the Service Oriented Architecture (SOA) technology was already considered a de-facto standard on the web. The corpus included 12,535 documents about Cloud Computing, 6,470 about Grid Computing, 8,908 about BPM, 6,030 about Semantic Web, and 5,781 about SOA. Since each collected GA was an aggregation of URLs linking 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. Generation of a temporally and contextually augmented concept map was carried out disjointedly for each assessed technology. For yielding high-quality concept maps in the assessment process for each technology, the dictionary used as 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.

Table I presents a partial list of automatically-identified co-occurring hot concepts for each IT topic, with normalized obtained PTA. The three quantitative pair-wise temporal operators, *Age, Frequency* and *CCDR*, were respectively weighted as follows: $\omega_1 = 0.2$, $\omega_2 = 0.4$ $\omega_3 = 0.4$ established based on an exploratory survey of 38 technology experts (were not among the 136 raters of the validation survey).

TABLE I. AUTOMATICALLY-IDENTIFIED CO-OCCURRING HOT CONCEPTS (PARTIAL LIST)

Topic	Co-occurring hot concepts	
Cloud Computing	PaaS[2]	IaaS[1]
	Salesforce	SaaS[3]
	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[4]
Service Oriented Architecture	Web Services	ESB
	Cloud	SaaS[3]
	Amazon	Cloud Computing
Business Process Management	SharePoint	Microsoft
	IBM	BPO[5]
	ERP[6]	Business Intelligence

[1]IaaS (Infrastructure as a Service)
[2]PaaS (Platform as a Service)
[3]SaaS (Software as a Service)
[4]RDF (resource description framework)
[5]ESB (enterprise service bus)
[6]BPO (business process optimization)
[7]ERP (enterprise resource planning)

Table II presents for each IT topic the top 3 co-occurring concepts under Concept 1 and Concept 2, with high $cSLV_{ij}$ values. The concept map yielded by the research instrument developed in this study allows a decision-maker to zoom-in to a specific concept which becomes the governing identification to which all other concepts would be related. This approach follows Novak & Gowin [46] and Harnisch et al. [24], who argue that concept maps are best constructed if a single focal root concept guides the selection of concepts and their organization in clusters on the map.

TABLE II. CO-OCCURRING CONCEPTS WITH HIGH $cSLV_{ij}$ VALUES

Topic	Concept 1	Concept 2
Cloud Computing	Government	Web Services
	Oracle	Amazon EC2
	IBM	IaaS
Grid Computing	IBM	Finance
	Parallel Processing	Utility Computing
	Google	Parallel Processing
Semantic Web	Microformats	Google
	RDF	Social Web
	SEO[1]	RDF
Service Oriented Architecture	Software AG	Information Technology
	Cloud	PaaS
	WSDL[2]	UDDI[3]
Business Process Management	BPO	Simulation
	Aris	Java
	Microsoft	BPMN[4]

[1]SEO (search engine optimization)
[2]WSDL (web service description language)
[3]UDDI (universal description discovery and integration)
[4]BPMN (business process model and notation)

To test the validity of the mechanism for detecting co-occurring hot concepts, predicative validity and inter-rater reliability (IRR) were processed in a typical case-by-variable statistical data structure, with the cases being the respondents and the variables being their subjective ratings (Table III).

TABLE III. PREDICATIVE VALIDITY AND FLEISS KAPPA COEFFICIENT

Topic	Percentage agreement	Fleiss Kappa Coefficient*
Buisienss Process Managemnt	85.52%	0.725
Cloud Computing	87.83%	0.745
Grid Computing	87.80%	0.765
Semamtic Web	86.07%	0.728
Service Oriented Architecure	81.85%	0.698
Total	**85.69%**	

* Values in a range of **0.61-0.8** indicate **substantial agreement**

Responses from respondents (n = 136) showed high values of Fleiss Kappa reliability-of-agreement measures for all five assessed technologies, indicating general substantial agreement and average predictive validity higher than 85%. In addition, to compare $cSLV_{ij}$ results with the human rankings for validating the improved similarity measures of relatedness proximity, inter-rater reliability measures and correlation coefficient measures were statistically analyzed. The average reliability for all the expert respondents is a measure of internal consistency, providing an index of homogeneity of responses based on the Intraclass Correlation Coefficient (ICC). As seen in Table IV, presenting obtained ICC values for each IT, homogeneity and similarity of responses indicate a high degree of inner resemblance of expert rankings for all five topics.

TABLE IV. ICC VALUES

Topic	ICC value
Buisienss Process Managemnt	0.943
Cloud Computing	0.983
Grid Computing	0.920
Semamtic Web	0.972
Service Oriented Architecure	0.978

Moreover, A Pearson's correlation coefficient was used to compare the average ranking produced by human subjects (i.e., respondents) with the model-generated value $cSLV_{ij}$. All measures performed well for all five ITs, with high correlations for $cSLV_{ij}$ values in the range of 0.879 – 0.951. When the same procedure was applied to $aSLV_{ij}$, low correlations were obtained for four ITs (0.250 – 0.541) with the exception of Grid Computing (0.763). This finding is due to the fact that $aSLV_{ij}$ values are attributed to conventional concept mapping without added knowledge. Low correlation values between human rankings and $aSLV_{ij}$ are expected for non-mature IT topics (as opposed to the potentially obsolete Grid Computing), since conventional co-word analysis based on a time-tagged corpus frequently lacks the contextual background available on the web. Thus, it can be concluded that by discovering and assimilating contextual knowledge in the form of webometric web counts, the research model thus elevates the conventional concept map to an augmented concept map, as long as the technology is not as outdated or mature. According to Google Trends, a service which indicates the frequency of topic searches over time, the interest in Grid Computing by the worldwide IT community was diminishing at the time of data collection, as indicated by an ongoing decrease in 2004-2011 of the search value index from 3 to 0.4.

Figure 4 presents, for example and for demonstration purposes, a concept map created based on the research model for Cloud Computing visualized in a concept-centric view. Its focal point (yellow circle) is ‘Business Intelligence’. Note also that hot co-occurring concepts constructed by the pair-wise temporal analysis for the same focal point are visualized by red lines. One line connects ‘Business Intelligence’ to ‘Business Objectives’ (dashed yellow circle) and another line connects ‘Business Intelligence’ to ‘Dashboards’ (dashed yellow 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.

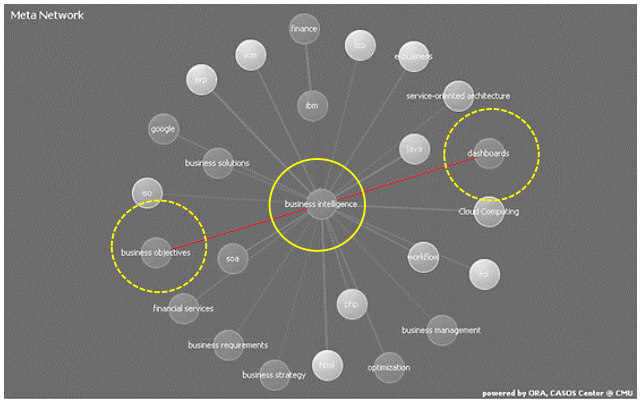

Figure 4 Concept map for Cloud Computing (concept-centric view)

VI. CONCLUSIONS

The PTA construct and the relatedness proximity measure developed in the current research to reflect the temporal and contextual distance between concepts were found to improve conventional concept mapping. The proposed research model is a framework for discovery of temporal and contextual knowledge in unstructured textual data synthesized from web sources. The developed textual data-driven research model and instrument provide a novel big data analytics framework which copes with the 3V attributes – volume, variety and velocity – of big data architecture. Since textual information accounts for the vast majority of traffic flowing over the web, this study's approach to textual data-driven decision making, harnessing research areas such as information extraction, text mining, web mining, temporal trend detection, concept mapping, and visualization, holds the potential promise to improve decision making processes toward acquiring and maintaining competitive advantage.

The innovation of this research is manifested in theoretical and practical contributions. The theoretical contribution is the harnessing of an open web-based time-tagged textual corpus and the development of unsupervised temporal trend detection construct and newly computed relatedness proximity measurement. The practical contribution of the current research is the design, development and implementation of an innovative research instrument that can serve as a prototype toward evolution of an automated tool for technology assessment.

The contributions of this work are consistent with the needs of decision makers charged with identifying future technological trends upon conducting an evaluation process of alternatives. Moreover, applying Google Alerts as a temporal and most updated source of web data is a major key element in the creation of a textual open and dynamic corpus in the current research. Finally, the managerial contribution of this work is in the ability of the research model and instrument to timely extract from textual data an augmented concept map that serves as a basis for deriving temporal and contextual insights, improving the visibility of information required to support top executives in their decision making processes.

While generalizing the current approach beyond the five topics used for demonstration and validation is left for future research, it is safe to conclude that the current work can most probably be generalized. It can also be concluded that the novel automated instrument developed to demonstrate and validate the research model have the potential to morph into a decision support system in the realm of managerial decision processes.

REFERENCES

[1] Alexa, M, (1997). Computer-assisted text analysis methodology in the social sciences. ZuMA-Arbeitsbericht, 97

[2] Almind, T.C. and Ingwersen, P., (1997). Informetric analyses on the World Wide Web: Methodological approaches to webometrics. Journal of documentation, 3, 404-426.

[3] Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. Journal of Paralle and Distributed Computing, 79, 3-15.

[4] Banko, M. & Brill, E. (2001). Scaling to Very Large Corpora for Natural Language Disambiguation. In Proceedings of ACL-01.

[5] Blank, G.D., Pottenger, W.M., Kessler, G.D., Herr, M., Jaffe, H., Roy, S., Gevry, D. & Wang, Q. (2001). Cimel: Constructive, Collaborative Inquiry-Based Multimedia E-learning. SIGCSE Bulletin, 33(3), 179.

[6] Blank, G.D., Pottenger, W.M., Kessler, G.D., Roy, S., Gevry, D.R., Heigl, J.J., Sahasrabudhe, S.A. & Wang, Q. (2002). Design and Evaluation of Multimedia to Teach Java and Object Oriented Software Engineering. In Proceedings of the 2002 American Society for Engineering Education Annual Conference & Exposition.

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

[8] Boykin, S. & Merlino, A., (2000). Machine Learning of Event Segmentation for News on Demand. Communications of the ACM, 43(2), 35-41.

[9] Budanitsky, A. & Hirst, G. (2006). Evaluating Wordnet-Based Constructs of Lexical Semantic Relatedness. Computational Linguistics, 32(1), 13-47J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[10] Callon, M., and 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 chemistry. *Scientometrics,* 11, 155 -205.

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

[12] 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, 57(3), 359 – 377.

[13] Chen, W. & Chundi, P. (2011). Extracting Hot Spots of Basic and Complex Topics from Time Stamped Documents. Data and Knowledge Engineering, 70(7), 642- 660.

[14] Courseault, C.R. (2004). A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions, PhD Dissertation, Georgia Institute of Technology.

[15] Courtial, J. P. (1994). A Coword Analysis of Scientometrics. Scientometrics, 3, 251-260.

[16] Desikan, P. & Srivastava, J. (2004). Mining Temporally Evolving Graphs. In the Proceedings of the Sixth WEBKDD Workshop in Conjunction with the 10th ACM SIGKDD conference, 22.

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

[18] Feldman, R., Klbsgen, W., Ben-Yehuda, Y.,Kedar, G. & Reznikov, V. (1997). Pattern Based Browsing in Document Collections. Principles of Data Mining and Knowledge Discovery: First European Symposium, PKDD'97.

[19] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

[20] Goorha, S. & Ungar, I. (2010). Discovery of Significant Emerging Trends. The Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 57-64.

[21] Grobelnik, M., Mladenic, D. & Milic-Frayling, N. (2000). Text Mining as Integration of Several Related Research Areas: Report on KDD'2000 Workshop on Text Mining. SIGKDD Explorations, 2(2), 99-102.

[22] Guston, D.H., and Sarewitz, D.,(2002), Real-Time Technology Assessment, Technology in Society Elsevier.

[23] Halsius, F. & Lochen, C. (2001). Assessing Technological Opportunities and Threats: An Introduction to Technology Forecasting. Division of Industrial Marketing, Lulea University of Technology.

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

[25] Havre, S., Hetzler, E., Whitney, P. & Nowell, L. (2002). Themeriver: Visualizing Thematic Changes in Large Document Collections. IEEE Transactions on Visualization and Computer Graphics, 9-20.

[26] He, Q. (1999). Knowledge Discovery through Co-Word Analysis. Library Trends, 48, 133-159.

[27] Ingwersen, P. (1998). The calculation of Web Impact Factors. Journal of Documentation 54(2), 236-243.

[28] Keller, F. & Lapata, M. (2003). Using the Web to Obtain Frequencies for Unseen Bigrams. Computational linguistics, 29(3), 459.

[29] 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: Berry, M. (ed.), Survey of Text Mining: Clustering, Classification, and Retrieval. Springer.

[30] Kostoff, R.N., and Eberhart, H.J., and Toothman, D.R., (1998), Database tomography for technical intelligence: A roadmap of the near-earth space science and technology literature. Information Processing and Management, 34, 69-85.

[31] Kostoff, R. N., and Tshiteya, R., and Pfeil, K. M., an Humenik, J. A., and Karypis, G., (2005). Power source roadmaps using bibliometrics and database tomography. Energy, 30(5), 709–730.

[32] Kostoff, R.N., and Johnson, D., and Bowles, C.A., and Bhattacharya, S., and Icenhour, A.S., and Nikodym, K,. and Barth, R.B., and Dodbele, S.,(2007). Technological forecasting and social change. Technological forecasting and social change, 74 (9): 1574-1608.

[33] Kostoff, R.N., (2008). Literature-Related Discovery (LRD): Introduction and background. The Journal of Technological Forecasting and Social Change, 75, 165 – 185.

[34] Kumaran, G. & Allan, J. (2004). Text Classification and Named Entities for New Event Detection. In the Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 297-304.

[35] Leake, D., Maguitman, A. & Canas, A. (2001). Assessing Conceptual Similarity to Support Concept Mapping. In the Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, 172-186.

[36] Lee, S., Baker, J., Song, J. & Wetherbe, J.C. (2010). An Empirical Comparison of Four Text Mining Methods. In the Proceeding of the 43rd Hawaii International Conference of System Sciences (HICSS), 1-10.

[37] Lent, B., Agrawal, R. & Srikant, R. (1997). Discovering Trends in Text Databases. In the Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD), 227 – 230.

[38] Leydesdorff, L. & Hellsten, I. (2006). Measuring the Meaning of Words in Contexts: An Automated Analysis of Controversies about 'Monarch Butterflies,' 'Franken Foods ' and 'Stem Cells'. Scientometrics. 67(2), 231–258.

[39] Ma, J. & Perkins, S. (2003). Online Novelty Detection on Temporal Sequences. In the Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 613-618.

[40] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. The management revolution. Harvard Bus Rev, 90(10), 61-67.

[41] Mei, Q. & Zhai, C.X. (2005). Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining. In the Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 198-207.

[42] Moed, H.F, and Glänzel, W., and Schmoch, U., (2004). Handbook of Quantitative Science and Technology Research: the use of publication and patent statistics in studies of S&T systems. Springer.

[43] Morinaga, S. & Yamanishi, K. (2004). Tracking Dynamics of Topic Trends Using a Finite Mixture Model. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining. 811-816.

[44] Müller, O., Junglas, I., vom Brocke, J., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems.

[45] Narin, F. and Olivastro, D. and Stevens, K. A. (1994). Bibliometrics/Theory Practice and Problems, Evaluation Review, 18(1), pp. 65-76.

[46] Novak, J.D. & Gowin, D. (1984). Learning How to Learn. Cambridge University Press:NY.

[47] Nowell, L.T., France, R.K. & Hix, D. (1997). Exploring Search Results with Envision. In Proceeding of Computer Human Interaction-CHI'97, 14-15.

[48] Plotnick, E. (1997). Concept Mapping: A Graphical System for Understanding the Relationship between Concepts. An ERIC digest. Clearinghouse on Information and Technology.

[49] Porter, A.L. & Cunningham, S. W. (2005). Tech Mining – Exploiting New Technologies for Competitive Advantage. Hoboken, NJ:John Wiley & Sons Publisher.

[50] Porter, A.L. and Detampel, M.J. (1995). Technology opportunities analysis, Technological Forecasting and Social Change, (49:3), pp. 237-255.

[51] Pottenger, W.M. & Yang, T. (2001). Detecting Emerging Concepts in Textual Data Mining. Computational Information Retrieval, 1-17.

[52] Power, D. J. (2016). Data science: supporting decision-making. Journal of Decision Systems, 1-12.

[53] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 3.

[54] Rajaraman, K. & Tan, A.H. (2001). Topic Detection, Tracking, and Trend Analysis Using Self-Organizing Neural Networks. In Advances in Knowledge Discovery and Data Mining. 102-107.

[55] Rapp, R. (2002). The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In the Proceedings of the 19th International Conference on Computational linguistics, 1-7.

[56] Rousseau, D.M. (1979). Assessment of Technology in Organizations: Closed Versus Open Systems Approaches. The Academy of Management Review, 4(4), 531-542.

[57] Roy, S., Gevry, D. & Pottenger, W.M. (2002). Methodologies for Trend Detection in Textual Data Mining. In the Proceedings of the Textmine'02 Workshop, Second SIAM International Conference on Data Mining, 58.

[58] Russell, A.W., Vanclay, F.M. & 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, 28(2), 109-116.

[59] Salton, G. (1988). Automatic Text Processing. Addison-Wesley Publishing Company.

[60] Salton, G., Wong, A. & Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM, 18(11), 613–620.

[61] Sasson, E., Ravid, G., & Pliskin, N. (2015). Improving similarity measures of relatedness proximity: Toward augmented concept maps. Journal of Informetrics, 9(3), 618-628.

[62] Schomm, Fabian, Florian Stahl, and Gottfried Vossen. "Marketplaces for data: an initial survey." ACM SIGMOD Record 42.1 (2013): 15-26.

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

[64] Su, H.N., and Lee, P.C.,(2010). Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. Scientometrics, 85(1), 65-79.

[65] Swan, R. & Jensen, D. (2000). Timemines: Constructing Timelines with Statistical Models of Word Usage. KDD-2000 Workshop on Text Mining.

[66] Thelwall, M. (2008). Quantitative comparisons of search engine results. Journal of the American Society for Information Science and Technology, 59 , 1702-1710.

[67] Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences . Morgan & Claypool Publishers 2009.

[68] Thomas, O. and Willett, P., (2000). Webometric analysis of departments of librarianship and information science. Journal of Information Science, 26(6), 421-428.

[69] Twiss, B.C., (1992). Forecasting For Technologists And Engineers – A practical guide for better decisions, Peter Peregrinus, Stevenage.

[70] Van Raan, A.F.J., and Van Leeuwen, Th. N., (2004). Assessment of the Scientific Basis of Interdisciplinary, Applied Research, Center for Science and Technology Studies (CWTS), University of Leiden, The Netherlands.

[71] 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.

[72] Wang, Y., Kung, L., Wang, C., Yu, W., & Cegielski, C. (2014). Developing a Big Data-Enabled Transformation Model in Healthcare: A Practice Based View.

[73] Wayne, C.L. (1997). Topic Detection and Tracking (TDT). On Workshop held at the University of Maryland, 27, 28-30.

[74] Wilks, Y. (1997). Information Extraction as a Core Language Technology, Lecture Notes in Computer Science, 1-9.

[75] Wong, P.C., Cowley, W., Foote, H., Jurrus, E. & Thomas, J. (2000). Visualizing Sequential Patterns for Text Mining. Information Visualization, 105-111.

[76] Wormell, I., (2001). Informetrics and webometrics for measuring impact, visibility, and connectivity in science, politics, and business. Competitive Intelligence Review, 12(1), 12-2.

[77] Zhang, K., Zi, J. & Wu, L.G. (2007). New Event Detection Based on Indexing-Tree and Named Entity. In the Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 215-222.