Journal of Informetrics 9 (2015) 618–628

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Improving similarity measures of relatedness proximity: Toward augmented concept maps

Elan Sasson [a], Gilad Ravid [b,*], Nava Pliskin [b]

[a] Faculty of Engineering, Tel Aviv University, Tel Aviv, Israel
[b] Faculty of Engineering Sciences, Ben-Gurion University of the Negev, Beer Sheva, Israel

ARTICLE INFO

Article history:
Received 18 December 2014
Received in revised form 16 June 2015
Accepted 17 June 2015
Available online 9 July 2015

Keywords:
Augmented concept map
Relatedness proximity
Co-word analysis
Webometrics
Technology assessment

ABSTRACT

Decision makers relying on web search engines in concept mapping for decision support are confronted with limitations inherent in similarity measures of relatedness proximity between concept pairs. To cope with this challenge, this paper presents research model for augmenting concept maps on the basis of a novel method of co-word analysis that utilizes webometrics web counts for improving similarity measures. Technology assessment serves as a use case to demonstrate and validate our approach for a spectrum of information technologies. Results show that the yielded technology assessments are highly correlated with subjective expert assessments ($n = 136$; $r > 0.879$), suggesting that it is safe to generalize the research model to other applications. The contribution of this work is emphasized by the current growing attention to big data.

1. Introduction

In today's world, under an economic climate characterized by global competition in highly transparent markets, reduction of uncertainty and risk levels in decision making becomes crucial for survival (Henselewski, Smolnik, & Riempp, 2006). For example, technological insights based on analysis of information from a myriad of sources are prime factors in support of technology-savvy decision makers charged with decision making about enterprise-level investment decisions in existing, emerging, and hot technologies and the technology assessment it requires.

One obvious treasure trove of information is the Internet, where the exponential growth of textual content is an established phenomenon (Varian, 2006; Prado & Ferneda, 2007; Hauber, Vesmarovich, & Dufour, 2012). However, this excellent source has inherent limitations as its vast scope of data makes it difficult to extract valuable insights. Naisbitt's (1996) statement – that we are drowning in information, especially since nearly 80% of online data is unstructured (White, 2005) – seems particularly apt in this context. When relying on web search engines for relevant information, decision makers face an abundance of information that restricts their ability to absorb and analyze the information quickly and efficiently, raising three primary concerns (Feldman, 2004). First, they neither have information skills nor a roadmap of what to look for or when it is reasonable to stop looking. Second, search result pages returned by search engines for a specific query include copious information to sort, read, and integrate, inundating decision makers with too much information while they remain unequipped with adequate tools to help handle the "flood". Third, there is really no metric to compare the value

* Corresponding author. Tel.: +972 864 72772/+972 544 905 391.
E-mail address: gilad@ravid.org (G. Ravid).

http://dx.doi.org/10.1016/j.joi.2015.06.003
1751-1577/

of a 'good' and a 'bad' search, despite the fact that relevance measurement is crucial to web searching and to information retrieval (IR).

These three concerns are especially challenging in the context of technology assessment, considered in this paper as a use case relevant to other decision making situations as well. Assessment of Information Technology (IT), in particular, entails an even greater challenge since new IT innovations occur at increasing speeds and with shorter life cycles (Ashrafi, Xu, Kuilboer, & Koehler, 2006). Drowning in textual information scattered all over the web, organizations turn for answers to leading IT consulting firms and analyst groups. These vendors compile reports about various IT disciplines which are usually rather expensive and whose objectivity is sometimes in question, especially when the vendor firm is associated with a specific IT provider. Domain experts as well are overwhelmed by abundant textual data and are constrained by the amount of time spent in the retrieval and analysis process (Li & Zhong, 2004).

To deal with this complexity, decision makers and analysts attempt to exploit the massive amounts of available textual documents on the web via applications that harness text mining and co-occurrence analysis with an aim to automatically generate concept maps (Porter & Detampel, 1995; Plotnick, 1997; Budanitsky & Hirst, 2006; Waltman, van Eck & Noyons, 2010). Generally speaking, concept maps (also termed co-occurrence networks) capture concepts and their relationships in a two-dimensional visually-based graphic representation (Leake, Maguitman, & Canas, 2001), dealing with questions as: what are the most relevant concepts and what are the underlying relevant relationships among concepts in a specific domain?

Automatically-generated concept maps do respond to the challenge of extracting useful information for decision making purposes. However, one of their limitations is that they often leave decision makers wondering how closely concept pairs are contextually related on the map. The goal of this paper is to overcome this limitation by augmenting concept maps with improved similarity measure of the relatedness proximity between concept pairs. This goal is accomplished by quantifying the contextual distance between two concepts via the expansion of co-word analysis with webometrics, i.e., quantitative bibliometric counts on the web (also known as hit count estimate—HCE). Toward meeting its research goal, this work leverages a unique synergy of several well-established research fields.

To begin with, a corpus of time-tagged unstructured textual data about a specific domain is collected from diverse web-based sources. It is noteworthy that the time-tagged textual corpus was gradually built using Google Alerts (GA)—a content change-detection and notification service that automatically notifies subscribers when new Internet content matches a set of search terms (e.g., topic). This innovative corpus building method allows for the collection of relevant documents without the need to subjectively evaluate the cardinality or the authority of the feed sources, as the service supplier determines source validity. Harnessing temporal data referenced in GA messages to build a dynamic and open corpus is novel in the sense that most other approaches use controlled and limited content in closed databases, such as digital libraries of articles, possibly missing useful and relevant knowledge.

Once the corpus was available, augmented concept mapping commenced. To uncover hidden patterns in the corpus and to generate a conventional concept map, information extraction (IE) is applied to the corpus, using a text mining (TM) technique based on natural language processing (NLP) followed by co-word analysis. Then, to improve the initial concept map and create an augmented concept map, its relatedness proximity measures are processed further by numerically and visually depicting the extent to which concept pairs on the map are contextually related.

On the basis of the thorough literature review (presented in the next section), this research makes innovative theoretical and practical contributions. From the theoretical perspective, to the best our knowledge this study presents (in the third and fourth sections of the paper) the first attempt to improve measures of relatedness proximity by combining conventional and traditional corpus-level co-word analysis with webometric-based co-occurrence analysis. The improved measures enable the upgrading of concept maps based on traditional co-word analysis algorithms to augmented concept maps. From the practical perspective, as evident from the results (presented in the fifth and sixth sections of the paper), this study contributes to the development of a decision-support research model and research instrument for mangers engaged in decision-making processes. Once implemented with an automated research instrument that collects a corpus of texts on the web and presents a solid and precise picture of a specific knowledge domain in terms of an augmented concept map, this new model allows for manual derivation of insights by a decision maker, whether or not the domain is technological. The managerial contribution of this work is thus in the ability of the research instrument to timely extract an augmented concept map from textual data and help with the visualization of information required to support top executive decision making processes. By leveraging a unique synergy of several well-established research fields, this research potentially contributes to the quality of decision making processes and practices.

2. Literature review

To engage in managerial decision making as strategic business planning, decision makers apply insights that depend on their ability to anticipate future developments, understand market position vis-à-vis competitors, and identify upcoming innovations (Halsius & Lochen, 2001). Many studies (e.g., Rousseau, 1979; Russell, Vanclay, & Aslin, 2010), describe technology assessment as a valid use case of managerial decision making processes, since technology managers are constantly faced with the challenge of identifying emerging technologies with the greatest technological potential (Courseault, 2004). Technology assessment includes such decision making endeavors as technological forecasting or foresight (Anderson, 1997), technology monitoring (Porter, 1994), technology intelligence (Brockhoff, 1991), technology road mapping (Garcia & Bray, 1997), technology opportunity analysis (Porter & Detampel, 1995) and technology future analysis (Porter et al., 2004).

Considering the increasing complexity of the decision making process, Filho, Dos Santos, Coelho, and Santos (2005) stress the need for information support, whether the information is from formal and informal sources or from sources internal and external to the organization, arguing that decision support is becoming a critical success factor in the competitive global market.

Bolshakov and Gelbukh (2004) acknowledge that decision makers must read and understand an enormous quantity of internet text to make well-informed decisions. Clearly, it is beyond the ability of any person or group to comprehend such large quantities of textual data without use of quantitative indicators (Narin, Olivastro, & Stevens, 1994). Bibliometrics and scientometrics are methods which utilize quantitative indicators analysis and statistics, depicting publication patterns within a given field or body of literature (Zhu, Porter, Cunningham, Carlisie, & Nayak, 2004). Quantitative bibliometric indicators use information, such as word counts, date information, word co-occurrence information, corpus-level co-word analysis and citation information, to track activity in a subject area (Kontostathis, Galitsky, Pottenger, Roy, & Phelps, 2004). Porter and Detampel (1995) assert that a key tenet of bibliometrics are co-occurrences, presented as a linkage of concepts that can be detected in a specific domain, and considered important in bibliometric analysis.

While the amount of textual data available to us is constantly increasing, the human ability to understand and process this information remains constant and limited. Given the volume and complexity of the information involved, Lee, Baker, Song, and Wetherbe (2010) thus assert that manual analysis of unstructured textual data is ever more impractical. Conversely, automatic TM has the potential to give companies the competitive edge they need to survive by identifying patterns hidden inside vast collections of text data. The objective of TM is to exploit information contained in textual documents in various ways, including discovery of patterns and trends in textual data and associations among text objects (e.g., concepts) (Grobelnik, Mladenic, & Milic-Frayling, 2000). Moreover, TM involves IE, which is the task of extracting named entities and factual assertions from texts (Wilks, 1997). IE facilitates the transformation from unstructured document space to structured concept space, paving the way to analysis of interactions between concepts extracted from a textual corpus.

There is a fairly extensive body of literature on co-word analysis (e.g., Callon, Law, & Rip, 1986; Courtial, 1994). Feldman, Klbsgen, Ben-Yehuda, Kedar, and Reznikov (1997) provided an early seminal work on concept co-occurrence relationships in a corpus of documents. He (1999) considers co-word analysis a powerful and proven quantitative tool for knowledge discovery in a research field. According to Rapp (2002), concepts that co-occur tend to be related, demonstrating relatedness association. Therefore, co-occurring concepts have been considered as carriers of meaning across different domains in studies of science and technology, and general indicators of activity in textual document sets (Leydesdorff & Hellsten, 2006). Co-word clustering is a process that begins by assessing the strength of the link value between two concepts, as based on their co-occurrence in a given record or document, and ends with the grouping of strongly linked concepts into clusters. The definition used in this study for the co-occurrence measure is the Similarity Link Value (SLV), also known as Equivalence Index (E), which is defined by Callon et al. (1986) as:

$$\mathrm{SLV}_{ij} = \frac{C_{ij}^2}{C_i \times C_j}, \quad 0 \le \mathrm{SLV}_{ij} \le 1, \quad C_{ij} = C_{ji} \ge 0$$

In this definition, C_{ij} is the number of co-occurrences of terms i and j (i.e., the number of documents in which both terms co-occur), and C_i and C_j —respectively, count the term occurrence (i.e., the number of documents in which the term appears) of term i and term j. A concept map is a common method for representing the relationships among a set of concepts, with vertices/nodes (e.g., named entity concepts, such as person, company, location) capturing concepts, and edges/links capturing the relations between concepts (Ruiz-Primo & Shavelson, 1996). More specifically, a concept map is a dynamic graphical map that visually presents concepts and relevant relationship clusters, which can then be portrayed as an undirected graph $G = (V, E)$ consisting of a set of vertices V and a set of edges E. Novak and Canas (2008) argue that the relationships between concepts indicated by a connecting edge potentially represent creative leaps (i.e., meaningful learning) in the creation of new knowledge. Thus, the most challenging aspect of constructing a concept map is linking the concepts into a meaningful and coherent structure that reflects understanding of a specific domain. For any decision making purpose, including the use case of technology assessment, conventional concept mapping has an inherent and crucial flaw —i.e., the unreliable measure of the contextual distance between co-occurring concept pairs. This weaknesses is amplified when the textual corpus of the initial concept map is accompanied by a large amount of noise and overload of irrelevant contextual concept relationships. Indeed, a web-based corpus of textual data, such as is implemented in the current study, is often accompanied by a large amount of noise. This may result in an inaccurate or incomplete concept map where existing relations might not be discovered, discovered relations might not be the result of actual relations, or a given link might have a spurious or missing relationship.

3. Research model

The research model in this study meets the challenge of measuring the contextual distance between co-occurring concept pairs in conventional concept mapping by using webometric-based co-word analysis to measure relatedness proximity. First, a-priori calculation of each SLV is carried out. Second, a bibliometric SLV based on webometric hit count estimates (HCEs) or web counts is calculated. Third, these two SLV values are the basis for a combined SLV value. The research model thus includes three steps (Fig. 1):

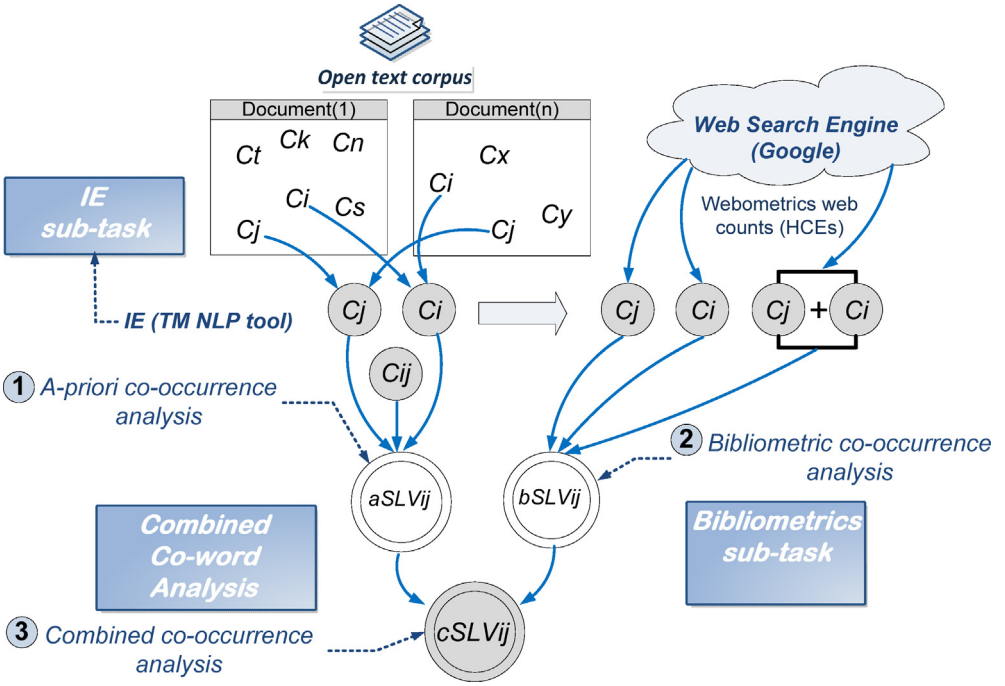

Fig. 1. A conceptual workflow describing the combined co-word analysis process.

1. *A-priori* co-occurrence analysis yielding $a\text{SLV}_{ij}$
2. *Bibliometric* co-occurrence analysis yielding $b\text{SLV}_{ij}$
3. *Combined* co-occurrence analysis yielding $c\text{SLV}_{ij}$.

In Step 1, using NLP-based TM to complete the IE task, significant semantic concepts (named entities, such as person, company, location, product) are extracted from the time-tagged corpus of text documents (e.g., TXT files, PDF, HTML files etc.), and an a-priori $a\text{SLV}_{ij}$ co-occurrence value is calculated for each relation between Concept i and Concept j. In Step 2, the bibliometric analysis task uses the same exact concept pairs in a series of webometric queries to a web search engine, acquiring after Concept i, Concept j, as well as their conjunctive Concept i + Concept j, using the AND Boolean operator. The bibliometric $b\text{SLV}_{ij}$ co-occurrence value is derived from the web counts of search results (i.e., number of web pages) retrieved for each concept pair. In Step 3, both a-priori and bibliometric SLVs (i.e., $a\text{SLV}_{ij}$ and $b\text{SLV}_{ij}$) are synthesized into a combined $c\text{SLV}_{ij}$ relatedness value for each concept pair, measuring relatedness proximity for the Concept-pair i, j.

Using this combined co-word analysis, weak or strong signals obtained in Step 1 are improved via weighted synthesis with the co-occurrence values, obtained by applying the webometric web counts in Step 2. It is safe to assume that concept relatedness, appearing in un-indicative context in the a-priori co-occurrence analysis (i.e., Step 1), may appear in a very obvious context in the bibliometric co-occurrence analysis (i.e., Step 2) and vice versa. Gledson and Keane (2008) consider web-searching an important part of measuring concept relatedness, as it provides up-to-date information on word co-occurrence frequencies in large available collection of documents such as the web. Similarly, Cilibrasi and Vitanyi (2007) assert that relative frequencies of web pages (e.g., web counts or HCEs) containing search terms give objective information about the relationship between the terms. Finally, the two types of SLVs were combined into the following additive formulation:

$$c\text{SLV}_{ij} = f\left(a\text{SLV}_{ij}, b\text{SLV}_{ij}\right) = f\left(\left(\frac{aC_{ij}^2}{aC_i \times aC_j}\right), \left(\frac{bC_{ij}^2}{bC_i \times bC_j}\right)\right) \approx \left(\frac{\left(aC_{ij} + bC_{ij}\right)^2}{\left(aC_i + bC_i\right) \times \left(aC_j + bC_j\right)}\right)$$

Assuming that,

$$cC_i = aC_i \cup bC_i$$

and similarly,

$$cC_j = aC_j \cup bC_j$$

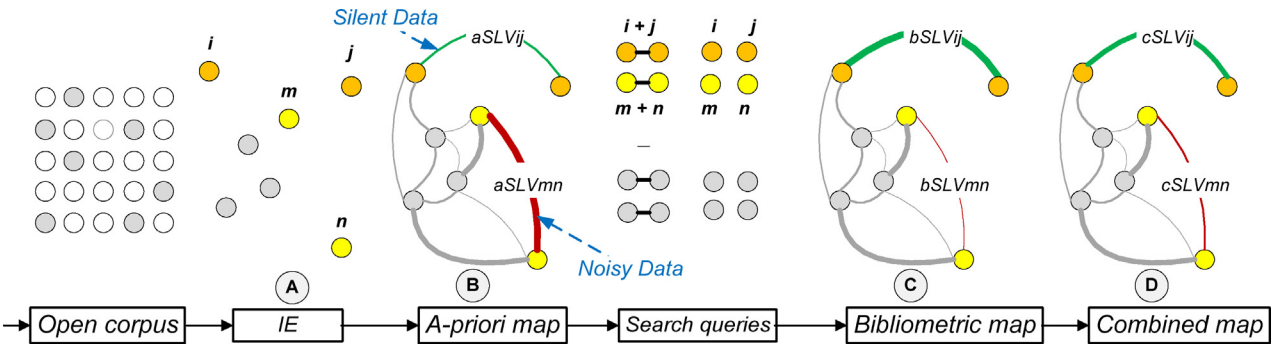

Fig. 2. Adding contextual knowledge to a concept map. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Provided that,

$$a\mathrm{SLV}_{ij} = \frac{\left(\#\left(aC_i \cap aC_j\right)\right)^2}{\#\left(aC_i\right) \times \#\left(aC_j\right)}$$

and,

$$b\mathrm{SLV}_{ij} = \frac{\left(\left(bC_i \cap bC_j\right)\right)^2}{\left(bC_i\right) \times \left(bC_j\right)}$$

Similarly,

$$c\mathrm{SLV}_{ij} = \frac{\left(\#\left(cC_i \cap C_j\right)\right)^2}{\#\left(cC_i\right) \times \#\left(bC_j\right)}$$

where by definition,

$$c\mathrm{SLV}_{ij} = f\left(a\mathrm{SLV}_{ij}, b\mathrm{SLV}_{ij}\right)$$

Thus,

$$c\mathrm{SLV}_{ij} = \frac{\left(\#\left(\left(aC_i \cup bC_i\right) \cap \#\left(aC_j \cup bC_j\right)\right)\right)^2}{\left(\#aC_i + \#bC_i\right) * \left(\#aC_j + \#bC_j\right)} \cong \frac{\left(\#\left(aC_i \cap aC_j\right) + \#\left(bC_i \cap bC_j\right)\right)^2}{\left(\#aC_i + \#bC_i\right) \times \left(\#aC_j + \#bC_j\right)}$$

and, as previously defined,

$$c\mathrm{SLV}_{ij} = \left(\frac{\left(aC_{ij} + bC_{ij}\right)^2}{\left(aC_i + bC_i\right) \times \left(aC_j + bC_j\right)}\right)$$

Hence, the basic premise underlying the bibliometric method for measuring 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, where silent information is emphasized and noisy information (e.g., outliers) is reduced, as illustrated in in Fig. 2(A)–(D), which for the sake of simplicity describes four steps for only two concept pairs.

A. The IE task extracts from the time-tagged corpus two concept pairs: Concept-pair m, n (colored in orange in Fig. 2) and Concept-pair i, j, (colored yellow in Fig. 2) with important co-occurrences.
B. The constructed initial concept map may potentially exclude silent data for Concept-pair i, j ($a\mathrm{SLV}_{ij}$ is erroneously low as illustrated by the thin green line connecting the two concepts), or include noisy data for Concept-pair m, n ($a\mathrm{SLV}_{mn}$ is erroneously high as illustrated by the bold red line connecting the two concepts).
C. The bibliometric concept map *is* generated by retrieving the webometric web counts from the search queries $(i, j, i \wedge j, m, n, m \wedge n)$ to, respectively, obtain the bibliometric co-occurrence index values $b\mathrm{SLV}_{ij}$ and $b\mathrm{SLV}_{mn}$ which, apparently, seem different compared to the a-priori index values, suggesting that the conventional concept map needs to be modified and improved accordingly (as illustrated by the adjusted thickness of the lines connecting the two concepts).
D. The extended relatedness proximity measure, adding contextual knowledge to the initial concept map, is generated by calculating the combined $c\mathrm{SLV}_{ij}$ and $c\mathrm{SLV}_{mn}$.

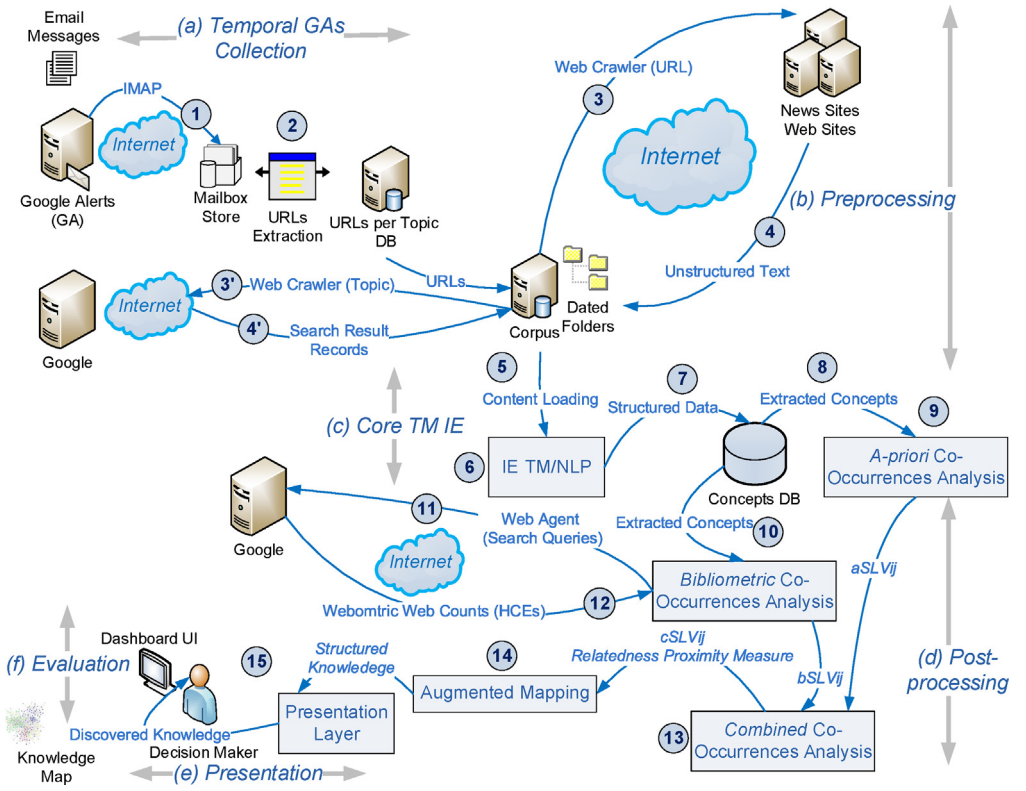

Fig. 3. The stages and tasks of the TASK research instrument.

Given the scale and noise characteristic of web data, it is possible that two concepts may appear on some web pages purely accidently. In order to reduce the adverse effects attributable to random co-occurrences, $b\text{SLV}_{ij}$ is set to zero by the model if the web count for the query bC_{ij} $(i \cap j)$ is under the $\varepsilon = 5$ threshold, similar to Bollegala, Matsuo, and Ishizuka (2007). The predefined threshold ensures that the support for the co-occurrence is high enough (Mika, 2005).

$$b\text{SLV}_{ij} = \begin{cases} 0 & \text{if } bC_{ij} \leq \varepsilon \\ \dfrac{bC_{ij}^2}{bC_i \times bC_j} & \text{otherwise.} \end{cases}$$

The extended relatedness proximity measure is accomplished by synthesizing webometric web counts (i.e., HCE) as bibliometric indicators derived as keyword search in a web search engine in the model, with the conventional co-word analysis to determine the relationship between co-occurring concepts. This method of improving the measurement of relatedness proximity features: (a) filtering out outlier concept co-occurrences (noise) that the conventional co-word analysis process extracts from the corpus and, inaccurately (due to resource-constrained corpus), presents them as significant to the decision maker, and (b) amplification filtering that enables discovery of elusive relationships not detected by the conventional co-word analysis process due to the weak signal yielded by the algorithm and, thus, missing out important hidden concept co-occurrences to the decision maker. The result of improving the measurement of relatedness proximity is a robust knowledge-added concept map for identification and selective extraction of significant concept co-occurrences. Decreasing the number and dimensionality of extracted concept pairs and displaying only significant key ones also improves the visualization of the resulting augmented concept map.

4. Methodology

A web-based research instrument (entitled TASK and shown in Fig. 3) was developed to demonstrate and validate the research model developed in the current study. This research instrument allows data collection and processing toward concept mapping that yields a time-tagged textual document. Then, via an advanced interactive dashboard-oriented user interface (UI), TASK allows decision makers to automatically generate an augmented concept map and derive relevant propositions.

Implementation of TASK followed the general leading methodology CRISP-DM (Cross Industry Standard Process for Data Mining) model, considered as the de facto standard for developing data mining and knowledge discovery projects (Marbán,

Mariscal, & Segovia, 2009). The instrument is divisible into the following six main stages (as proposed in the CRISP-DM model) and 16 tasks:

(a) ***Temporal GA collection*** tasks involve collecting a repository of Google Alert (GA) email updates which include one or more URL links to domain-specific (i.e., IT topic) web documents (e.g., HTML, XML) from various source types of web sites (news, web, blogs, and discussion groups' sites). The setting for the delivery rate of GA messages was defined on the basis of 'as-it-happens'. Steps 1 and 2 in Fig. 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 Fig. 3 depict this stage.
(c) ***Core TM and IE*** NLP-based tasks are routines and processes for concept discovery in the document corpus yielded by Stage (b), which is categorized, keyword-labeled and time-stamped, to extract and store concepts and their relevant metadata for further analysis (e.g., time stamp, total number of appearances, average concept distribution etc.). Steps 5 to 7 in Fig. 3 depict this stage.
(d) ***Post-processing*** analysis tasks include all procedures and methods required for conducting relatedness proximity measurements toward concept mapping. Steps 8 to 15 in Fig. 3 depict this stage.
(e) ***Presentation*** tasks and browsing functionality include graphical user interface and listing capabilities. *Presentation layer* components display the augmented concept map with references to co-occurrence weights calculated at each step. Step 16 in Fig. 3 depicts this stage.
(f) ***Evaluation*** tasks are carried out by decision makers while evaluating and interpreting acquired results, and are therefore not depicted in Fig. 3. Generalization, pruning or additional collection of textual data may be implemented by the user to enrich the corpus.

4.1. Corpus building and information extraction

As mentioned above, technology assessment is harnessed to demonstrate how the research model and instrument can improve similarity measures of relatedness proximity, serving as a use case with relevance to other applications as well. Datasets (39,724 URLs) used for building the time-tagged corpus about five IT topics were created using Google Alerts (GA) collected throughout 190 days in 2011. Each alert is an email message in HTML format; it includes aggregation of links (i.e., URLs) to the latest news articles about each of the five technologies used to demonstrate and validate the research model and instrument in this study from various source types (news, web, blogs, and discussion group sites). In planning the corpus, the goal was to use an IT array with a spectrum of IT types in lifecycle maturation stages sufficiently diverse for model demonstration and validation. Thus cloud computing, much-hyped at the time corpus building commenced in August 2010 (and expected to substitute grid computing), in addition to Business Process Management (BPM), which attracted a lot of new attention at the time, were both included. Semantic Web, regarded as new and particularly promising, and Service Oriented Architecture (SOA), already considered then a de-facto standard on the web, were also included in the IT array. The number of documents in the corpus for each IT topic was: cloud computing—12,535, grid computing—6470, BPM—8908, semantic web—6030, and SOA—5781.

To accomplish the IE process, NLP-based TM analysis was applied to the TXT files (following conversion of HTML file to text files) in the time-tagged corpus, using IBM's SPPS/PASW Text Analytics Version13 (former SPSS Text Mining Modeler) and AlchemyAPI for rigor and robustness, although each of these tools autonomously provides all functions necessary for the IE process. The IE process employs a named-entity processor which allows identification of multi-gram (i.e., NLP phrases) concepts, such as person names, location names and names of organizations. Moreover, to improve the accuracy of text extraction, it is common to apply domain-dependent linguistic resource files during extraction phase to refine the rules and dictionaries applied in the course of IE. In this work, to yield high-quality IT assessment results, the dictionary (i.e., domain-specific resource file) employed in the IE process was sensitive to IT knowledge to allow 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. Finally, a sparse document-by-concept (NLP phrases) occurrence matrix, demonstrating concept presence by 'T' (true) and concept absence by 'F' (false), was computed and uploaded to SQL database for further analysis. The number of extracted concepts (n) posed a computational-complexity challenge of $O(n^2)$ in the co-word analysis, while generating the concept-by-concept relatedness matrices.

To cope with the scalability challenge and based on similar studies (Leydesdorff & Hellsten, 2006; Hutchins & Benham-Hutchins, 2010), 100 top concepts were used as the maximum number of concepts to be included in the computation of the concept-by-concept relatedness matrices, so that an optimized number of k concepts yields near-linear time complexity. Moreover, a large concept dataset also impedes visualization of concept maps since the effectiveness of tools for presenting maps is impaired by too many concepts and associations and, essentially, the number of items that a user can take in, at any one time, is rather limited (Feldman & Snager, 2007; Novak & Canas, 2008). Liu (2007) asserts that in most specific-domain

Table 1
Co-occurring concepts with high $cSLV_{ij}$ values.

Topic	Concept 1	Concept 2
Cloud computing	Government	Web Services
IaaS (Infrastructure as a Service)	Oracle	Amazon EC2
SOA (service oriented architecture)	SOA	Amazon
	IBM	IaaS
Grid computing	Sun	Parallel Processing
	IBM	Finance
	Parallel Processing	Utility Computing
	Google	Parallel Processing
Semantic web	RDF	Social Web
RDF (resource description framework)	HTML 5	Mobile applications
SEO (search engine optimization)	Microformats	Google
	SEO	RDF
Service oriented architecture	Software AG	Information Technology
PaaS (Platform as a Service)	PaaS	Cloud
WSDL (web service definition language)	Middleware	Financial services
UDDI (universal description discovery & integration)	WSDL	UDDI
Business process management	BPO	Simulation
BPO (business process optimization)	Aris	Java
BPMN (business process markup notation)	Microsoft	BPMN
	IBM	BPO

concepts extraction applications, the decision maker is only interested in some specific concepts in order to simplify and to speed up the knowledge acquisition process.

The extended relatedness proximity measurement was conducted disjointedly for each of the five investigated IT topics used to demonstrate and validate the research model. Table 1 presents top 4 co-occurring IE-extracted (NLP phrases) concepts (i.e., Concept 1 and Concept 2) with high $cSLV_{ij}$ values for each IT topic.

4.2. Validation methodology and tools

Validation of the proposed model is a two-fold process. First, the relatedness proximity measurement is validated, including validation of the HCE webometric web counts used in the bibliometric co-occurrence analysis. Second, propositions derived from the augmented concept maps for each IT topic are validated.

Relatedness proximity analysis is validated as a targeted web-based survey ($n = 136$) aimed for domain experts, such as IT practitioners and IS scholars. Survey questionnaires were distributed internationally for each IT topic, targeting a database of domain experts obtained from two major sources: LinkedIn and a leading global IT consulting firm. The majority of respondents (89%) have over four years of experience in a specific IT topic. The survey questionnaire was comprised of questions about 20 pairs of co-occurring concepts for relatedness proximity validation, asking the respondent to determine weighted relationship scores.

This study implements a visual analogue scale (VAS) as a continuous evaluation device rather than categorical scales which only reach ordinal-scale level. A VAS device consists of a line and two anchors, one at each end, highlighted with verbal material that mark opposite ends of the relatedness proximity semantic addressed in a similarity question. This type of measurement requires survey participants to express their opinion in a visual form, placing a mark at an appropriate position on a continuous line (McClure, Sonak, & Suen, 1999; Ruiz-Primo & Shavelson, 1996); (2) questions about the respondent's demographic characteristics as nationality, position, and number of years of experience regarding an assessed IT topic.

The web counts (i.e., HCE) validation process is based on seeking logical consistency among multiple related search queries, also known as *Metamorphic Relations*, as proposed by Chen, Tse, and Zhou (2010). To define the metamorphic validation (i.e., MR for conjunctions of the type 'a AND b'), let X be a search criterion and $\#(X)$ be the number of pages returned for search criterion X. A useful general MR is then defined as:

$$\mathrm{MR}_{\mathrm{AND}} : \text{if } A_2 \equiv (A_1 \text{ AND } B), \quad \text{then } \#(A_2) \leq \#(A_1)$$

The latter should hold because any page satisfying the A_1 AND B co-existence condition must also satisfy the A_1 condition, but not vice versa. The actual number and percentage of failed MR tests observed for all five topics is a low overall total of 2.9%, suggesting that HCE values are fairly reliable and consistent.

5. Results

5.1. Relatedness proximity measurement validation

To compare $cSLV_{ij}$ results with the human rankings for validating the improved similarity measures of relatedness proximity measurements, inter-rater reliability measures and correlation coefficient measures were statistically analyzed. The

Table 2
ICC values.

Topic	ICC
Cloud computing	0.983
Grid computing	0.920
Semantic web	0.972
Service oriented architecture	0.978
Business process management	0.943

Table 3
Pearson correlations.

Topic	Expert' ratings vs. $cSLV_{ij}$ Pearson's correlation coefficient	Expert' ratings vs. $aSLV_{ij}$ Pearson's correlation coefficient
Business process management	0.879	0.423
Cloud computing	0.951	0.250
Grid computing	0.939	0.763
Semantic web	0.949	0.541
Service oriented architecture	0.913	0.325

reliability for all the expert raters averaged together is a measure of internal consistency, providing an index of homogeneity of responses based on the Intraclass Correlation Coefficient (ICC), also termed (in SPSS) the average measure intraclass correlation is equal to the value of Cronbach's alpha for continuous variables.

As seen in Table 2, presenting obtained ICC values for each topic, homogeneity and similarity of responses indicate a high degree of inner resemblance of expert rankings for all five topics. A Pearson's correlation coefficient was used to compare the average ranking produced by human subjects (i.e., raters) with two model-generated values $aSLV_{ij} \wedge cSLV_{ij}$. Table 3 presents Pearson's correlations, suggesting that all measures perform well for all five topics, with high correlations for $cSLV_{ij}$ values (left side of Table 3) and lower correlations for $aSLV_{ij}$ values (right side of Table 3). The only exception is grid computing, for which a relatively high correlation was found between expert rankings and $aSLV_{ij}$ (0.763), compared with the low correlation obtained for the four other IT topics; this is due to the fact that $aSLV_{ij}$ values are attributed to conventional concept mapping without added knowledge. Low correlation values between rater rankings and $aSLV_{ij}$ are expected for non-mature IT topics (as opposed to the more mature grid computing), since conventional co-word analysis based on a time-tagged corpus frequently lacks the contextual background available on the web.

By discovering and assimilating contextual knowledge in the form of webometric web counts, the research model thus elevates the conventional concept map in the technology assessment use case to an augmented concept map, as long as the technology is not as mature as grid computing. According to Google Trends, a service which indicates frequency of topic searches over time, the interest in grid computing by the worldwide IT community is diminishing, as indicated by an ongoing decrease of *search value index* in 2004–2011 from 3 to 0.4.

5.2. Validation of model-based propositions

Proposition validation is accomplished by comparing two sets of propositions: (1) propositions derived by technology savvy decision makers from augmented concept maps generated by the TASK instrument, and (2) propositions extracted from assessments reported by leading IT consulting firms, such as Gartner, and complemented by studies and scholar research. The validation revealed that the former propositions (#1), derived based on pairs of highly correlated co-occurring concepts

Table 4
Validation of Cloud Computing propositions based on concepts co-occurrence.

No.	Derived proposition	Extracted proposition validation statements
1	Cloud Computing and **Web Services** are infrastructure of open public data cloud for various **e-government** services	'The importance of **government G-clouds** as models for large cities, creating urban clouds sponsored by local government hold the promise to reduce IT costs, providing platforms for small business applications and **e-services**.' (Townsend, Maguire, Liebhold, & Crawford, 2011; Komninos, Schaffers, & Pallot, 2011). *'Smart cities* consider Cloud Computing and **Web Services** as a fundamental layer of open public data cloud from various **gov**ernment sources and agencies.' (Ovum, 2011)
2	Self-managed **Oracle** set of technologies are provided by **Amazon EC2** (elastic compute cloud) as hosting service on the Cloud	'On September 2010 Oracle and Amazon announced at the OpenWorld conference the deployment of a range of Oracle software on Amazon's Elastic Compute Cloud (EC2) service in production form.' (http://www.computerworld.com/s/article/9186840/Oracle_will_support_apps_running_on_Amazon_EC2)
3	**SOA** referred to the use of public cloud services plays a vital role in **Amazon** offering on the cloud	'**Amazon** excels at addressing IT operations audience on the path towards data center transformation. . .by migrating to **Amazon** cloud. . .' (Gartner, 2011)

presented on the TASK's augmented concept maps, were found to be compatible with the latter propositions (#2), extracted from respective reports by a leading IT consulting firm (i.e., Gartner) and scholar assessment studies. For the sake of brevity of this manuscript, only one snap shot of propositions for Cloud Computing is presented in Table 4.

6. Conclusion

The research model proposed in this work is an analytics framework that embraces and synthesizes unstructured textual data from web sources. Since the use of textual information accounts for the vast majority of traffic flowing over the web, this study's approach holds the potential promise to improve decision making processes toward acquiring and maintaining competitive advantage. The textual data-driven decision making model developed in the current study relies on research areas such as information extraction, text mining, web mining, concept and knowledge mapping as well as visualization.

As a demonstrative use case, the research instrument used to augment the conventional concept mapping was found valuable in assisting decision makers in assessing emerging and existing ITs but less appropriate for more mature ones. The newly computed similarity measure to improve the relatedness proximity measurement was found to be highly correlated with experts subjective ratings ($n = 136$): $r > 0.879$. Also, high inter-rater reliability scores were found based on Intraclass Correlation Coefficient (ICC) > 0.92. Moreover, in model validation processes, propositions derived based on the augmented concept maps were found to be valuable.

The first challenge faced by the current research is that some dynamically created web pages are difficult to find or to access due to not being indexed by commercial search engines, thus hidden in the 'Invisible Web' ('Deep Web'). Another major challenge left for future research is the aspiration to automate the whole decision making process, not just the concept mapping part. Moreover, future research should also develop full-map validation methods, including a detection of erroneous automatically-generated propositions. Although, it is safe to assume that the current work that validated the proposed research model and instrument in the technology assessment case, can most probably be generalized to other domains for which ample textual data is posted on the web, yet it is advised to address this issue by future research. Finally, it is recommended that the novel contextual augmentation of concept mapping in support of managerial decision process in the enterprise be enhanced by temporal augmentation.

References

Anderson, J. (1997). On using concept maps to assess the comprehension effects of reading expository text. In *ERIC document reproduction service*.

Ashrafi, N., Xu, P., Kuilboer, J., & Koehler, W. (2006). Boosting enterprise agility via IT knowledge management capabilities. In *Proceedings of the 39th annual Hawaii international conference on system sciences, HICSS'06* (p. page46), vol. 2.

Bolshakov, I. A., & Gelbukh, A. (2004). *Computational linguistics: Models, resources, applications*. Mexico: Center for Computing Research (CIC) of the National Polytechnic Institute, the Economic Culture Fund Press.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th international conference on World Wide Web* (pp. 757–766).

Brockhoff, K. (1991). Competitor technology intelligence in German companies. *Industrial Marketing Management*, *20*(2), 91–98.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47.

Callon, M., Law, J., & Rip, A. (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. Washington, D.C: Macmillan Press.

Chen, T. Y., Tse, T. H., & Zhou, Z. Q. (2010). Semi-proving: An integrated method for program proving, testing, and debugging. *IEEE Transactions on Software Engineering*, *37*(January/February (1)), 109–125.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(3), 370–383.

Courtial, J. P. (1994). A coword analysis of scientometrics. *Scientometrics*, *3*(1), 251–260.

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

Feldman, R., Klbsgen, W., Ben-Yehuda, Y., Kedar, G., & Reznikov, V. (1997). Pattern based browsing in document collections. In *Principles of data mining and knowledge discovery: First European symposium, PKDD'97*.

Feldman, R., & Snager, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.

Feldman, S. (2004). The high cost of not finding information. *KMWorld*, *13*(March (1)). http://www.kmworld.com.

Filho, L., Dos Santos, D. M., Coelho, G. M., & Santos, M. D. E. M. (2005). Future studies in Brazil: CGEE approach for bio-and nanotechnology. *Journal of Business Chemistry*, *2*, 126–137.

Garcia, M. L., & Bray, O. H. (1997). *Fundamentals of technology road mapping, Sandia National, Laboratories and United States*. Albuquerque, NM: Dept. of Energy.

Gartner. (2011). *Hype cycle for cloud computing 2011*. Gartner.

Gledson, A., & Keane, J. (2008). Using web-search results to measure word-group similarity. In *Proceedings of the 22nd international conference on computational linguistics* (pp. 281–288), vol. 1.

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* (vol. 2:2).

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

Hauber, R. P., Vesmarovich, S., & Dufour, L. (2012). The use of computers and the Internet as a source of health information for people with disabilities. *Rehabilitation Nursing*, *27*(4), 142–145.

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

Henselewski, M., Smolnik, S., & Riempp, G. (2006). Evaluation of knowledge management technologies for the support of technology forecasting. In *Proceedings of the 39th annual Hawaii international, conference on system sciences, HICSS'06*.

Hutchins, C. E., & Benham-Hutchins, M. (2010). Hiding in plain sight: Criminal network analysis. *Computational & Mathematical Organization Theory*, *16*(1), 89–111.

Komninos, N., Schaffers, H., & Pallot, M. (2011). Developing a policy roadmap for smart cities and the future internet. In *eChallenges e-2011 conference proceedings, IIMC International Information Management Corporation*.

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

Leake, D., Maguitman, A., & Canas, A. (2001). Assessing conceptual similarity to support concept mapping. In *Proceedings of the fifteenth international Florida artificial intelligence research society conference* (pp. 172–186).
Lee, S., Baker, J., Song, J., & Wetherbe, J. C. (2010). An empirical comparison of four text mining methods. In *2010 43rd Hawaii international, conference of system sciences (HICSS)* (pp. 1–10).
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.
Li, Y., & Zhong, N. (2004). Web mining model and its applications for information gathering. *Knowledge-Based Systems*, *17*(5), 207–217.
Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data*. Berlin Heidelberg: Springer-Verlag (2007).
Marbán, Ó., Mariscal, G., & Segovia, J. (2009). Data mining and knowledge discovery process model. In J. Ponce, & A. Karahoc (Eds.), *Data Mining and Knowledge Discovery in Real Life Applications*. Vienna: I-Tech.
McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, *36*, 475–492.
Mika, P. (2005). Flink: Semantic web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, *3*(2), 211–223.
Naisbitt, J. (1996). *Megatrends 2000*. New York, NY: Smithmark Publishers.
Narin, F., Olivastro, D., & Stevens, K. A. (1994). Bibliometrics/theory practice and problems. *Evaluation Review*, *18*(1), 65–76.
Novak, J. D., & Canas, A. J. (2008). *The theory underlying concept maps and how to construct and use them*. Florida: Florida Institute for Human and Machine Cognition Pensacola.
Ovum. (2011). *Is your city smart enough?* London: Ovum Publications.
Plotnick, E. (1997). Concept mapping: A graphical system for understanding the relationship between concepts: An ERIC digest. In *ERIC Digest*. EDO-IR-97-05.
Porter, A. L. (1994). Technology opportunities analysis: Integrating technology monitoring, forecasting, and assessment with strategic planning. *SRA Journal*, *26*(2), 21–31.
Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change*, *49*(3), 237–255.
Porter, A. L., Ashton, W. B., Clar, G., Coates, J. F., Cuhls, K., Cunningham, S. W., et al. (2004). Technology futures analysis: Towards integration of the field and new methods. *Technological Forecasting and Social Change*, *49*, 287–303.
Prado, H. A., & Ferneda, E. (2007). *Emerging technologies of text mining: Techniques and applications, information science reference*. New York, NY: Hershey.
Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th international conference on computational linguistics* (pp. 1–7), vol. 1.
Rousseau, D. M. (1979). Assessment of technology in organizations: Closed versus open systems approaches. *The Academy of Management Review*, *4*(4), 531–542.
Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concepts maps in science assessment. *Journal of Research in Science Teaching*, *33*, 569–600.
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.
Townsend, A., Maguire, R., Liebhold, M., & Crawford, M. (2011). A planet of civic laboratories. In *The future of cities, information and inclusion*. Palo Alto: Institute for the Future.
Varian, H. R. (2006). The economics of internet search. *Rivista di Politica Economica SIPI Spa*, *96*, 9–23.
Waltman, L., van Eck, N. J., & Noyons, E. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(October (4)), 629–635.
White, C. (2005). Consolidating, accessing, and analyzing unstructured data. In *Business intelligence network article*. ⟨http://www.b-eye-network.com/view/2098⟩.
Wilks, Y. (1997). *Information Extraction as a Core Language Technology, Lecture Notes in Computer Science*. Springer-Verlag.
Zhu, D., Porter, A., Cunningham, S., Carlisie, J., & Nayak, A. (2004). *A process for mining science & technology documents databases, illustrated for the case of knowledge discovery and data mining*. Atlanta, GA: Technology Policy & Assessment Center Georgia Institute of Technology.