Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2014 Proceedings

TEXT MINING AND TEMPORAL TREND DETECTION ON THE INTERNET FOR TECHNOLOGY ASSESSMENT: MODEL AND TOOL

Elan Sasson
Ben Gurion University of the Negev, Beer Sheva, Israel, sasson.elan@gmail.com

Gilad Ravid
Ben-Gurion University of the Negev, Beer-Sheba, Israel, rgilad@bgu.ac.il

Nava Pliskin
Ben-Gurion University of the Negev, Beer-Sheva, Israel, pliskinn@bgu.ac.il

Elan Sasson, Gilad Ravid, and Nava Pliskin, 2014, "TEXT MINING AND TEMPORAL TREND DETECTION ON THE INTERNET FOR TECHNOLOGY ASSESSMENT: MODEL AND TOOL", Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014, ISBN 978-0-9915567-0-0
http://aisel.aisnet.org/ecis2014/proceedings/track08/3

TEXT MINING AND TEMPORAL TREND DETECTION ON THE INTERNET FOR TECHNOLOGY ASSESSMENT: MODEL AND TOOL

Complete Research

Sasson, Elan, Ben-Gurion University of the Negev, Beer-Sheva, Israel, elans@bgu.ac.il

Ravid, Gilad Ben-Gurion University of the Negev, Beer-Sheva, Israel, rgilad@bgu.ac.il

Pliskin, Nava, Ben-Gurion University of the Negev, Beer-Sheva, Israel, pliskinn@bgu.ac.il

Abstract

In today's world, organizations conduct technology assessment (TAS) prior to decision making about investments in existing, emerging, and hot technologies to avoid costly mistakes and survive in the hyper-competitive business environment. Relying on web search engines in looking for relevant information for TAS processes, decision makers face abundant unstructured information that limit their ability to assess technologies within a reasonable time frame. Thus the following question arises: how to extract valuable TAS knowledge from a diverse corpus of textual data on the web? To cope with this question, this paper presents a web-based model and tool for knowledge mapping. The proposed knowledge maps are constructed on the basis of a novel method of co-word analysis, based on webometric web counts and a temporal trend detection algorithm which employs the vector space model (VSM). The approach is demonstrated and validated for a spectrum of information technologies. Results show that the research model assessments are highly correlated with subjective expert (n=136) assessment (r > 0.91), and with predictive validity value above 85%. Thus, it seems safe to assume that this work can probably be generalized to other domains. The model contribution is emphasized by the current growing attention to the big-data phenomenon.

Keywords: Technology Assessment, Information Extraction, Co-word Analysis, Temporal Trend Detection.

1 Introduction

In today's harsh global business arena, the pace of events has increased rapidly, with technological innovations occurring at ever-increasing speed and considerably shorter life cycles. This reality means that any company wishing to remain economically viable must engage in technology assessment (TAS) prior to making any crucial decision regarding investments, mainly in new and emerging technologies. The process thus begins anew, with information collected from a myriad of sources, analysed and finally acted upon. One obvious treasure trove of information is of course the internet, where the exponential growth of textual content is an established phenomenon (Varian, 2006; Prado and Ferneda, 2007; Hauber et al., 2012). However, this excellent source has inherent limitations; its vast scope of data makes it difficult to extract valuable insights. Naisbitt's (1996) statement seems particularly apt in this context - that we are drowning in information but thirsty for knowledge, especially given that nearly 80% data is unstructured (White, 2005). When relying on web search engines in looking for relevant information for TAS processes, decision makers face an abundance of information that restricts their ability to assess technologies quickly and efficiently.

There are already many managerial analytical applications that exploit the massive amounts of available textual documents. Some of these applications, which are of great importance to decision makers, intelligence analysts, and marketing analysts, perform text mining and co-occurrence analysis for generating concept maps (Porter and Detampel, 1995; Plotnick, 1997; Budanitsky and Hirst, 2006; Waltman et al., 2010). Generally speaking, concept maps capture concepts and concept relationships within a knowledge domain, using a two-dimensional visually-based graphic representation of concepts and their relationships (Leake et al., 2001). Concept maps (co-occurrence networks) may deal with various questions, such as: what are the underlying relationships among concepts in a specific technology domain? Or what are the most relevant concepts and relationships in a concept map associated with that technology domain? It is assumed in the current study that automatically-generated concept maps, while responding to the challenge of extracting useful information for TAS purposes, leave technology assessors still wondering how closely concept pairs on the map are *contextually* related and *temporally* linked.

In this research, the goal was to overcome this limitation and improve concept mapping for TAS by using webometrics synthesized with co-word analysis and pair-wise temporal measures, thereby quantifying the distance between two concepts contextually and temporally. Application of webometrics, i.e., quantitative bibliometric counts on the web (also known as hit count estimate - HCE), improves concept mapping by enhancing the relatedness proximity measure between concepts. Expressing the temporality value of the relationship between two co-occurring concepts improves concept mapping via pair-wise temporal analysis, distinguishing between pairs of emerging and hot concepts using a novel unsupervised and automatic algorithm. The addition of contextual and temporal information upgrades the traditional concept map to a knowledge map, based upon which a technology-savvy decision maker is able to derive TAS insights. These insights provide a comprehensive picture not only of the general landscape of the evaluated technology, but also of important indications of trends, providing actionable knowledge that might not otherwise be cost-effectively achievable.

This research leverages a unique synergy of several well-established research fields. The TAS approach in this work begins by amassing a corpus of unstructured textual data about a specific technology from diverse web-based sources. Then, to uncover hidden patterns in the corpus and generate a conventional concept map (co-occurrence network), information extraction (IE) is applied to the corpus to create a concept map, using a text mining (TM) technique based on natural language processing (NLP) followed by co-word analysis. However, the generated concept map can contribute very little confident knowledge about concept relatedness. To bridge the knowledge gap, the initial concept map is then processed further in this work into a knowledge map, and improved in two novel

consecutive phases: relatedness proximity measurement and pair-wise temporal analysis, depicting the extent to which concept pairs on the map are contextually related and temporally linked.

This research makes novel theoretical and practical contributions. From the theoretical perspective, to the best our knowledge, this study represents the first attempt in jointly measure relatedness proximity and conduct pair-wise temporal analysis, thus upgrading concept maps based on traditional co-word analysis algorithms into knowledge maps. From the practical perspective, this study contributes to the development of a decision-support research model and research instrument for mangers engaged in TAS decision making processes.

The remainder of this paper is structured as follows: Section Two introduces literature review and theoretical background. The research model and methodology are discussed in Sections Three and Four, respectively. Section Five presents study results, and the paper is summarized by a section discussing limitations and offering a conclusion.

2 Literature Review

The ability of decision makers to foresee technological advances and assess new and current technologies is essential for anticipating future developments, understanding market position vis-a-vis competitors, identifying upcoming innovations, and finally applying these insights to strategic business planning (Halsius and Lochen, 2001). 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 et al., 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 et al., 2004). Quantitative bibliometric indicators use information, such as word counts, date information, word co-occurrence information and citation information, to track activity in a subject area (Kontostathis et al, 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, potentially providing a powerful source of information on emerging technologies.

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 et al. (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 et al., 2000). Moreover, TM involves IE, which is the task of extracting named entities and factual assertions from text (Wilks, 1997). IE allows the transformation from the unstructured document space to the 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 et al, 1986; Courtial, 1994). Feldman et al. (1997) provided an early seminal work on concept co-occurrence relationships in a corpus of documents. He (1999) considers co-word analysis as 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 and 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), defined by Callon et al. (1986) as:

$$SLV_{ij} = \frac{{C_{ij}}^2}{C_i * C_j} \, , 0 < \; SLV_{ij} \leq 1, C_{ij} = \; C_{ji} \; \geq 0$$

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 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 and Shavelson, 1996). More specifically, a concept map is a dynamic graphical map that visually presents concepts and relevant relationship clusters, which can be portrayed as an undirected graph *G= (V, E)* consisting of a set of vertices *V* and a set of edges *E*. Novak and Canas (2008) argue that the relationships between concepts indicated by a connecting edge often represent creative leaps (i.e., meaningful learning) in the creation of new knowledge. The most challenging aspect of constructing a concept map is linking the concepts into a meaningful, coherent structure that reflects understanding of a specific domain (Canas and Novak, 2004). For TAS purposes, conventional concept mapping suffers from one major drawback - i.e., the unreliable measure of the contextual distance between co-occurring concept pairs. This weaknesses is amplified when the textual corpus upon which the initial concept mapping 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 a missing relationship.

In many TM applications, one encounters a stream of text documents, each of which contains some meaningful time stamp (e.g., publication date), with an underlying temporal and evolutionary structure (Mei & Zhai, 2005). Temporal Text Mining (TTM), concerned with discovering temporal patterns in text information collected over time (such as emerging concepts and hot concepts), are crucial factors in the TAS decision-making process and can be depicted as carriers of current and progressive tendencies of new developments in a specific technology or scientific field under investigation (Porter & Cunningham 2005). Pottenger and Yang (2001) present the process of detecting emerging conceptual content in regions of semantic locality in concept maps, analogous to the operation of a radar system, which aims to effectively assist in the differentiation of mobile (i.e., dynamic) and stationary (i.e., static) objects (e.g., concepts). Thus, the challenge of detecting emerging and hot concepts in textual data applications depends on mapping from the digital domain (portrayed as a web-based corpus) to the semantic domain (presented as concept map) in a temporally sensitive environment. Emerging trend detection (ETD) systems may help address the temporal challenge of concept mapping, as ETD systems help identify concepts that are either novel or are growing in importance within a textual data corpus. In most conventional ETD applications, a human reviewer is required to subjectively finalize concept classification (Porter and Detampel 1995; Nowell et al., 1997; Blank et al., 2001; Roy et al., 2002; Havre et al., 2002; Blank et al., 2002; Chen, 2006), imposing limitations such as comparability and completeness. Another limitation of most ETD systems is related to the use of unitary static monolithic text corpus from human-maintained indexed databases, such as INSPEC, TDT or COMPENDEX (Nowell et al., 1997; Lent et al., 1997; Swan and Jensen, 2000; Wong et al., 2000; Kumaran & Allan, 2004; Mei and Zhai, 2005; Zhang et al., 2007; Subasic and Berendt, 2010; Chen and Chundi, 2011). A closed static textual data corpus suffers from limited diversity, variety and richness, and it must be periodically refreshed, all of which constitute major drawbacks in terms of data coverage (Alexa, 1997; Zweigenbaum et al., 2001; Banko and Brill, 2001;

Keller and Lapata, 2003) and indexer effect (King, 1987; Law and Whittaker, 1992). Moreover, many ETD research projects aim at performing single node (stand-alone concept) analysis for detecting changes (Desikan and Srivastava, 2004), such as identifying time periods with a burst of activities related to a stand-alone topic based on analysing time-stamped documents. To improve the temporal dimension of concept mapping in the current study, detection of temporal attributes of co-occurring concepts is needed at a two-item level. The aim in this study is to conduct pair-wise temporal analysis regarding two co-occurring concepts, thereby expanding the scope of concept maps to present decision makers with far improved, more practical knowledge maps. This approach differs from tracing changes over time in ETD for frequency of one *single concept* with no reference to other concepts. For example, the study of one concept at a certain time period may have influenced or stimulated the study of another concept during the same time period. Discovering such evolutionary relationships between concepts can reveal not only hidden concepts as a single semantic entity, but most importantly the latent inter-linkage of timely synchronized emerging and hot co-occurring concepts.

3 Research Model

As presented, this study involved adding contextual and temporal knowledge to the initial concept map, modelled to yield a knowledge map, in order to overcome the drawbacks of conventional concept mapping for TAS. Creation of the web-based time-tagged textual corpus mainly involves gradual iterative accumulation over time of text documents about a specific technology targeted for assessment, retrieved from various web sources using Google Alerts (GA) service and stored in a database. GA is Google's content change-detection and notification service that automatically notifies subscribers when new Internet content matches a set of search terms (e.g., topic). This corpus building novel method allows collecting relevant documents without the need to subjectively evaluate the feed sources, since the service supplier determines the source validity.

3.1 Relatedness Proximity Measurement via Co-Word Analysis

This study aimed to meet the challenge inherent in the existing contextual distance between co-occurring concept pairs of conventional concept mapping for TAS by using webometric-based co-word analysis to measure relatedness proximity. Consequently, contextual knowledge was added to the initial concept map, and a-priori calculation of each SLV was carried out, followed by calculation of a bibliometric SLV based on webometric hit count estimates (HCEs) or web counts. Then, the two previously calculated SLV values were combined to an extended SLV value, thus following the three steps illustrated in Figure 1:

1. *A-priori* co-occurrence analysis yielding $aSLV_{ij}$
2. *Bibliometri*c co-occurrence analysis yielding $bSLV_{ij}$
3. *Combined* co-occurrence analysis yielding $cSLV_{ij}$

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 $aSLV_{ij}$ 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, and the bibliometric $bSLV_{ij}$ co-occurrence value is derived from the web counts of the search results retrieved for each concept pair. In Step 3, both a-priori and bibliometric SLVs (i.e., $aSLV_{ij}$ and $bSLV_{ij}$) are synthesized into a combined $cSLV_{ij}$ relatedness value for each concept pair, measuring relatedness proximity for the Concept-pair *i, j*.

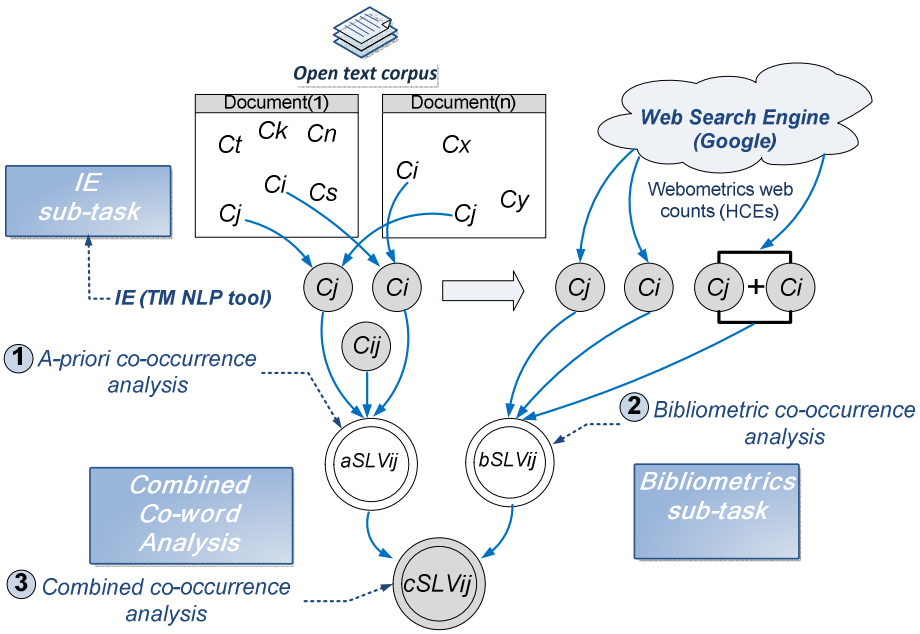

Figure 1. A conceptual workflow describing the combined co-word analysis process

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 unindicative 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. Finally, it seems promising to combine the two types of SLVs into one $cSLV_{ij}$ (Step 3) based on the following additive formula:

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

This method of improving the measurement of relatedness proximity includes two key features relevant to decision makers: (a) de-noising filtering of outlier concept co-occurrences that the conventional co-word analysis process extracts from the corpus, to avoid the problem of mistakenly presenting insignificant data as significant (b) amplification filtering, enabling discovery of elusive relationships left undetected by conventional co-word analysis process due to the weak signal yielded by the algorithm, thereby overlooking hidden concept co-occurrences. The resulting improvement in 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 knowledge map.

3.2 Pair-wise Temporal Analysis via Trend Detection

When adding temporal knowledge to the initial concept map by means of pair-wise temporal analysis, temporal measures were used to determine concept categorization, based on the time dimension of co-

occurring hot concepts, and on co-occurring emerging concepts. These temporal measures (i.e., temporal fingerprints), defined as quantitative pair-wise temporal operators, are based on objective time properties derived from time-tagged textual corpus. These are as follows:

1. Age - Age of relevant documents where concepts co-occur
2. Frequency - Publishing frequency rate (i.e., activity ratio) of relevant documents where concepts co-occur in a given time interval.

In order to classify a pair of concepts as hot or emerging concepts, they should be semantically richer at a later time than they were at an earlier time and hence Age is chosen as the first measure (Pottenger and Yang, 2001; Goorha and Ungar, 2010). The choice of Frequency as the second measure is 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 (Wayne 1997; Chen, 2006).

To present temporal values in concept maps in a novel way, the current study enhances scalability of the Vector Space Model (VSM) by conducting a *pair-wise temporal analysis* (Salton 1988; Salton *et al.*, 1975), exploiting the cosine similarity measure for expressing the temporality value of the relationship between two co-occurring concepts in a concept map. For example, the relationship can be described either as *old* or *young*. Given that Concept *i* and Concept *j* co-occur in *n* documents, a Vector $\vec{y}$ with *n* dimensions is created, where each coordinate reflects the number of days since creation of each document, and documents are accordingly chronologically ordered. The cosine similarity measure is then applied to a reference Vector $\vec{x}$ with *n* dimensions (i.e., $\vec{y}$ and $\vec{x}$ are two Vectors with equal magnitude), whereby all coordinate values are ones (1s) which express 'fresh' temporal notions and chronological proximity to present time. With this definition we have the angle between Vector $\vec{x}$ and Vector $\vec{y}$, $0 \leq cos(\vec{x},\vec{y}) \leq 1$, where the cosine similarity measure will be close to one (1) in cases that $\vec{x}$ and $\vec{y}$ are nearly identical, thus indicating temporal relationship between the co-occurring concepts regarded as *young*. In contrast, in cases where vectors have little in common, the cosine similarity measure will be close to zero, indicating that the temporal relationship between the co-occurring concepts is *old*.

The following example (Figure 2) demonstrates this approach. Suppose two Concepts i and j co-occur in four (n =4) different documents (See 'a' in Figure 2), published on four different dates: one year ago, four months ago, one month ago, and one week ago. The two corresponding variable-size date ordered vectors ($x_i\, y_{i_,} \in R^n$) are: $\vec{y}$ = (7,30,120,365) and $\vec{x}$ = (1,1,1,1). The first quantitative pair-wise temporal operator of the pair-wise temporal analysis $Age(\vec{x},\vec{y})$: marked 'b' in Figure 2 is defined as:

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

Consecutively, the second quantitative pair-wise temporal operator, $Frequency(\vec{x},\vec{y})$ of the pair-wise temporal analysis, describing "activeness" (i.e., publishing frequency rate, activity ratio - marked 'c' in Figure 2) is calculated. The sequentially ordered coordinates of the vector describing $Frequency(\vec{x},\vec{y})$ are calculated by subtracting the values of two subsequent coordinates; $|y_k - y_{k-1}|$ of the corresponding Vector $\vec{y}$. The subtraction should yield minimal values, as high publication ratio indicates an imaginary notion of documents published on a daily basis. Thus, the cosine similarity measure is then applied to a reference Vector $\vec{x}$ with $n-1$ dimensions, where all coordinates assume the value one (1) to reflect *active* temporal value (0 is not applicable). The two corresponding variable-size date ordered vectors ($x_i\, y_{i_,} \in R^n$) are: $\vec{y}$ = (23,90,245) and $\vec{x}$ = (1,1,1). Accordingly, $Frequency(\vec{x},\vec{y})$ is defined as

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

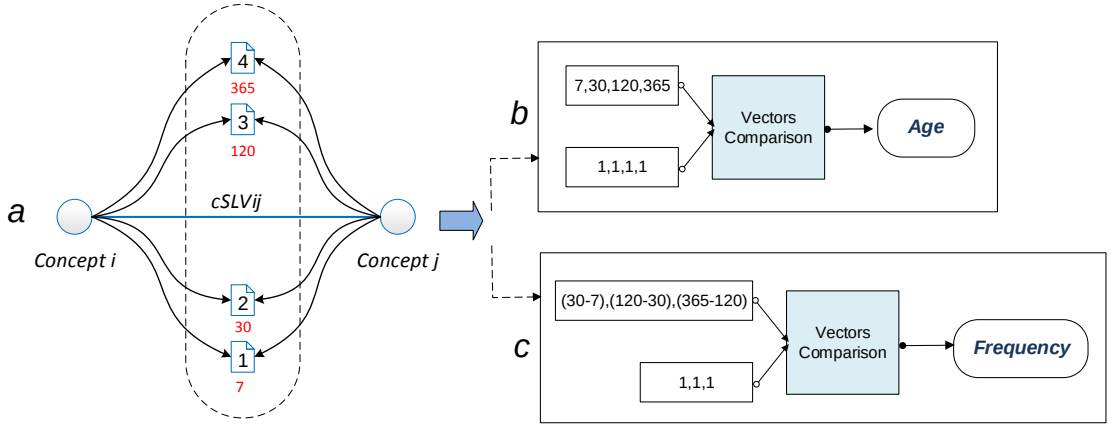

Figure 2. Concept co-occurrences in four different documents

For example, to automatically detect co-occurring emerging concepts in the current research model, the following pair-wise temporal analysis (PTA) value is defined as a linear equation in Definition 1.1.

Definition 1.1

$$\underbrace{PTA}_{emerging}{}_{ij} = f\{(\omega_1 * Age_{ij}(\vec{x},\vec{y})),\ (\omega_2 * Frequency_{ij}(\vec{x},\vec{y}))\}$$

Given that PTA is between 0 and 1, a threshold value δ for classification of co-occurring emerging concepts can either be empirically determined or manually set. In the present study, an exploratory survey (n = 38) of technology experts was conducted for establishing the actual values of ω_i (i.e., $\omega_1 = 0.33$ and $\omega_2 = 0.66$).

4 The TASK research instrument

A web-based research instrument was developed for mapping TAS knowledge (TASK) in order to demonstrate and validate the research model developed in the current study. This research instrument allows data collection and processing prior to knowledge mapping that yields a time-tagged textual document. Then, via an advanced interactive dashboard-oriented user interface (UI), the research instrument allows technology savvy decision makers to automatically generate a knowledge map and explore the extracted knowledge toward derivation of TAS propositions. Implementation of the TASK research instrument (shown in Figure 3) followed the general CRISP-DM model. The instrument is divisible into the following six main stages and 16 tasks:

a) ***Temporal GAs collection*** tasks involve collecting a repository of Google Alert (GA) email updates which includes one or more URL links to domain-specific (i.e., IT topic) web documents (e.g., HTML, XML) in diverse web sites. The setting for the delivery rate of GA messages was defined on the basis of 'as-it-happens'. 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 GAs 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 by Stage (b), which is categorized, keyword-labelled and time-stamped, toward extracting and storing for further analysis concepts and their relevant metadata (e.g., time stamp, total number of appearances, average concept distribution etc.). 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 knowledge mapping. Steps 8 to 15 in Figure 3 depict this stage.

e) ***Presentation*** tasks and browsing functionality include graphical user interface and listing capabilities. *Presentation layer* components display the knowledge map with references to co-occurrence weights calculated at each step, as well as the detected co-occurring emerging concepts and co-occurring hot concept. Step 16 in Figure 3 depicts this stage.

f) ***Evaluation*** tasks are carried out by the decision maker while valuating and interpreting the acquired results, and are therefore not depicted in Figure 3. Generalization, pruning or requiring collection of additional textual data in order to enrich the corpus, may be implemented by the user.

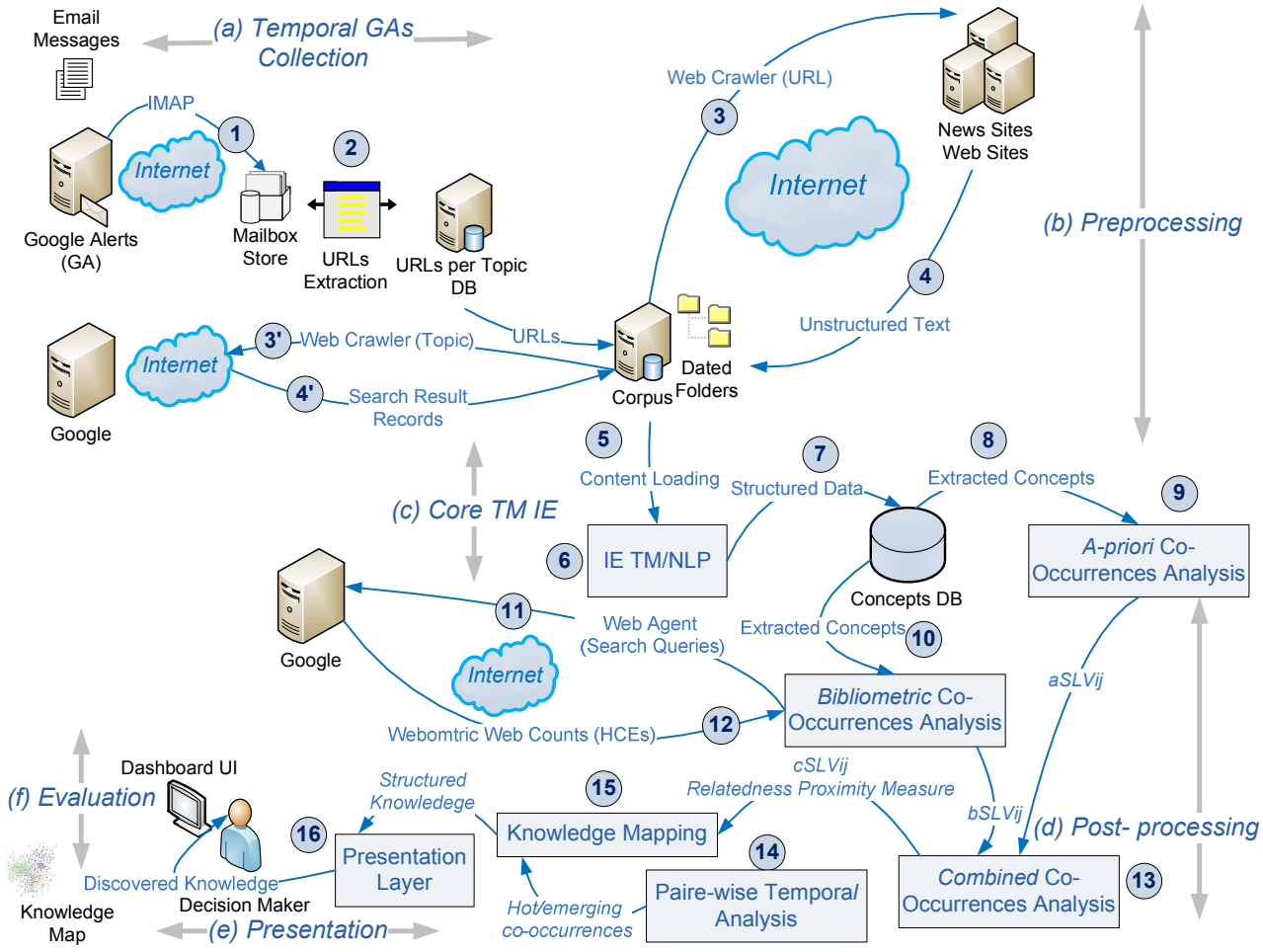

Figure 3. The stages and tasks of the TASK research instrument

4.1 Corpus Building and Information Extraction

Datasets used for building the time-tagged corpus were created using Google Alerts (GA) collected throughout 190 days (August 12, 2010 to February 16, 2011). Each alert is an email message in HTML format; it includes aggregation of links (39,724 URLs) to the latest news articles about each technology 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, which was over-hyped when corpus building commenced in August 2010 and expected to substitute grid computing, in addition to Business Process Management (BPM), which attracted a lot of a 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 - 6,470, BPM - 8,908, semantic web - 6,030, and SOA - 5,781.

To accomplish the IE process, NLP-based TM analysis was applied to the TXT 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. Finally, a sparse document-by-concept occurrence matrix, showing concept presence in a document marked 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 and Hellsten, 2006; Hutchins and 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. The extended relatedness proximity measurement, as well as the pair-wise temporal analysis (as previously described) were conducted disjointedly for each of the five investigated IT topics used to demonstrate and validate the research model.

4.2 Validation methodology and tools

Validation of the proposed model is a three-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, the pair-wise temporal analysis of detecting co-occurring hot concepts is validated. Finally, TAS propositions derived from the knowledge maps for each IT topic are validated.

Relatedness proximity and pair-wise temporal analyses were 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 more than four years of experience in a specific IT topic. The survey questionnaire included two major parts: (1) questions about 20 pairs of co-occurring concepts for relatedness proximity validation, asking the respondent to determine weighted relationship scores (McClure et al., 1999; Ruiz-Primo and Shavelson, 1996) ; (2) questions about 10 pairs of co-occurring concepts for pair-wise temporal analysis validation, asking respondents to determine whether the concepts are hot or not.

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 Zhou et al. (2010). 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.

The TAS proposition validation is accomplished by comparing manually-derived propositions from knowledge maps generated by the TASK tool, with propositions extracted from assessments reported by leading IT consulting firms, such as Gartner, complemented by studies and scholar research.

5 Results

5.1 Relatedness proximity measurement validation

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

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 1. ICC values

As seen in Table 1, presenting obtained ICC values for each topic, the 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: $aSLV_{ij}$ and $cSLV_{ij}$. Table 2 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 2) and lower correlations for $aSLV_{ij}$ values (right side of Table 2). The only exception is grid computing, for which a relatively high correlation between the expert's rankings and the $aSLV_{ij}$ (0.763), in comparison to the low correlation obtained for the four other IT topics, is because the $aSLV_{ij}$ values are attributed to conventional concept mapping without added knowledge. Low correlation values between raters' 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 knowledge background available on the web. By discovering and assimilating that knowledge in the form of webometric web counts the research model augments the concept map to a knowledge map as long as the technology is not as mature as grid computing. According to Google Trends, a service which shows frequency of topic searches over time, the interest in grid computing by the worldwide IT community is diminishing, as indicated by the ongoing decrease of the *search value index* in 2004-2011 from 3 to 0.4. Thus, from the perspective of relatedness proximity measurement, the research model and instrument seem more promising and

valuable for assessing current, innovative and evolving technologies rather than mature ones, which are understandably less relevant in the TAS context.

Topic	Expert' Ratings vs. $cSLV_{ij}$		Expert' Ratings vs. $aSLV_{ij}$	
	Pearson's correlation coefficient	**P_value***	**Pearson's correlation coefficient**	**P_value***
Business Process Management	0.879	0.000	0.423	0.063
Cloud Computing	0.951	0.000	0.250	0.289
Grid Computing	0.939	0.000	0.763	0.000
Semantic Web	0.949	0.000	0.541	0.014
Service Oriented Architecture	0.913	0.000	0.325	0.162

Table 2. Pearson correlations

5.2 Pair-wise temporal analysis validation

To test validity of the mechanism for detecting co-occurring hot concepts, predicative validity and inter-rater reliability (IRR) were analysed in a typical case-by-variable statistical data structure, with the cases being the responders and variables being their subjective ratings. Responder answers (n = 136) showed high values of Fleiss Kappa reliability-of-agreement measures for all IT topics (Table 3), indicating general substantial agreement and average predictive validity higher than 85%. The predictive validity is based on a percentage-agreement measure, which seems to be the most prevalent method for calculating the consensus estimate $A=O/P$, where the agreement rate A is the division of the observed agreement O by the possible agreement P (Grayson and Rust, 2001).

Topic	Percentage agreement measure	Fleiss Kappa Coefficient*
Business Process Management	85.52%	0.725
Cloud Computing	87.83%	0.745
Grid Computing	87.83%	0.765
Semantic Web	86.07%	0.728

Service Oriented Architecture	81.85%	0.698
Total	**85.69**%	

*Table 3. Predicative validity and Fleiss Kappa coefficient [*Values in a range of **0.61-0.8** indicate **substantial agreement** (Landis and Koch, 1977)].*

5.3 Validation of model-based TAS propositions

Manual derivation of major propositions, based on map-centric views of knowledge maps generated by the research instrument for each assessed IT, either by technology-savvy professionals or scholars (like the authors), was pursued in the present study for the five IT topics. The TAS propositions derived, as based on pairs of highly correlated co-occurring concepts presented on the knowledge maps, were found to be compatible with the respective TAS reports by a leading IT consulting firm (i.e., Gartner) complemented with scholar assessment studies.

6 Conclusions

The knowledge mapping model described here is comprised of two integrated, novel, theoretical components: relatedness proximity measurement - reflecting contextual distance between concepts, and pair-wise temporal analysis - reflecting temporal distance between concepts. A technology savvy decision maker is able to manually derive TAS propositions using the TASK research instrument, which automatically acquires a textual data corpus and generates the relevant knowledge map that improves the conventional concept mapping process contextually and temporally. The research instrument used to implement the knowledge-mapping research model was found valuable in assisting decision makers in assessing emerging and existing ITs, but less appropriate for more mature ones. The first challenge faced by this research is that some dynamically created web pages are difficult to find or to access, as they are not indexed by commercial search engines, and therefore hidden in the 'Invisible Web'. The second challenge is the language challenge, as commercial search engines display language biases in site coverage, favouring English as a site language. The third challenge of generalization of the research model to other domains beyond the five assessed IT topics used for demonstration and validation is left to be addressed by future research. Moreover, it remains for future research to extend the research instrument to a variety of applications other than technology assessment, such as augmenting web-search experience. Another challenge left for future research is the aspiration to automate the whole TAS process, not just knowledge mapping. Finally, it can be concluded that the novel decision support system provided by this research model and instrument has the potential to morph into the realm of managerial decision processes.

References

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

Blank, G.D. and Pottenger, W.M. and Kessler, G.D. and Herr, M. and Jaffe, H. and Roy, S. and Gevry, D. and Wang, Q. (2001). Cimel: Constructive, collaborative inquiry-based multimedia e-learning. SIGCSE Bulletin, 33(3), 179.

Blank, G.D. and Pottenger, W.M. and Kessler, G.D. and Roy, S. and Gevry, D.R. and Heigl, J.J., and Sahasrabudhe, S.A. and 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.

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

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

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

Callon, M. and Law, J. and Rip, A. (1986). Mapping the dynamics of science and technology: sociology of science in the real world. Macmillan Press.

Canas, A.J. and Novak, J.D. and Gonz'alez, F.M. and Carvalho, M. and Arguedas, M. and Cognition, M. (2004). Mining the web to suggest concepts during concept map construction. Universidad P'ublica de Navarra.

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.

Chen, W. and Chundi, P. (2011). Extracting hot spots of basic and complex topics from time stamped documents. Data and Knowledge Engineering, 70(7), 642- 660.

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

Desikan, P. and Srivastava, J. (2004). Mining temporally evolving graphs. In the Proceedings of the Sixth WEBKDD Workshop in conjunction with the 10th ACM SIGKDD conference, 22.

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

Goorha, S. and Ungar, I. (2010). Discovery of significant emerging trends. The Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 57-64.

Grayson, K. and Rust, R. (2001). Interrater reliability. Journal of Consumer Psychology, 10(1), 71-73.

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

Halsius, F. and Lochen, C. (2001). Assessing Technological Opportunities and Threats – An introduction to Technology Forecasting. Division of Industrial Marketing, Lulea University of Technology.

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

Havre, S. and Hetzler, E. and Whitney, P. and Nowell, L. (2002). Themeriver: Visualizing thematic changes in large document collections. IEEE transactions on visualization and computer graphics, 9-20.

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

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

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. Computational linguistics, 29(3), 459-484.

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

Kontostathis, A. and Galitsky, L.M. and Pottenger. W.M. and Roy, S. and Phelps, D.J. (2004). A Survey of Emerging Trend Detection in Textual Data Mining, In: Berry, M., (ed.), Survey of Text Mining: Clustering, Classification, and Retrieval. Springer.

Kumaran, G. and Allan, J. (2004). Text classification and named entities for new event detection. In the Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 297-304.

Law, J. and Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. Scientometrics, 23, 417-461.

Leake, D. and Maguitman, A. and 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.

Lee, S. and Baker, J. and Song, J. and Wetherbe, J.C. (2010). An Empirical Comparison of Four Text Mining Methods. The Conference of System Sciences (HICSS), 2010 43rd Hawaii International, 1-10.

Lent, B. and Agrawal, R. and Srikant, R. (1997). Discovering trends in text databases. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD), 227 – 230.

Leydesdorff, L. and Hellsten, I. (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies,' 'Franken foods ' and 'stem cells'. Scientometrics volume. 67(2), 231–258.

McClure, J.R. and Sonak, B. and Suen, H.K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. Journal of research in Science Teaching, 36, 475-492.

Mei, Q. and 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.

Naisbitt, J. (1996). Megatrends 2000.Smithmark Publishers, New York.

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

Novak, J.D. and Canas, A.J. (2008). The theory underlying concept maps and how to construct and use them. Florida Institute for Human and Machine Cognition Pensacola.

Nowell, L.T. and France, R.K. and Hix, D. (1997). Exploring search results with Envision. CHI'97 extended abstracts on Human factors in computing systems: looking to the future, 14-15.

Plotnick, E. (1997). Concept mapping: A graphical system for understanding the relationship between concepts: An ERIC digest. Clearinghouse on Information and Technology.

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

Pottenger, W.M. and Yang, T. (2001). Detecting emerging concepts in textual data mining. Computational information retrieval, 1-17.

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

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

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, 1 -7.

Roy, S. and Gevry, D. and 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.

Ruiz-Primo, M.A. and 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.

Subasic, I. and Berendt, B. (2010). From bursty patterns to bursty facts: The effectiveness of temporal text mining for news. Citeseer.

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

Salton, G. and Wong, A. and C. S. Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.

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

Varian, H.R. (2006). The economics of Internet search. University of California at Berkeley.

Waltman, L. and van Eck, N.J. and Noyons, E. (2010). A unified approach to mapping and clustering of bibliometric networks. Journal of Informetrics.

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

White, C. (2005). Consolidating, Accessing, and Analyzing Unstructured Data. 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, 1- 9.

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

Zhang, K. and Zi, J. and 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.

Zhou, Z.Q. and Zhang, S.J. and Hagenbuchner, M. and Tse, T.H. and Kuo, F.C. and Chen, T.Y. (2010). Automated functional testing of online search services. Software Testing, Verification and Reliability.

Zhu, D. Porter, A. Cunningham, S. Carlisie, J., and Nayak, A. (2004). A process for mining science & technology documents databases, illustrated for the case of "knowledge discovery and data mining," Technology Policy & Assessment Center Georgia Institute of Technology, Atlanta, GA.

Zweigenbaum, P. and Jacquemart, P. and Grabar, N. and Habert, B. (2001). Building a text corpus for representing the variety of medical language. Studies in health technology and informatics, 290-294, IOS Press.