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

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).