Information Processing and Management 48 (2012) 303–325

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

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

Indices of novelty for emerging topic detection

Yi-Ning Tu [a,*], Jia-Lang Seng [b]

[a] Department and Graduate School of Statistics and Information Science, College of Management, Fu Jen Catholic University, New Taipei 242, Taiwan
[b] Department and Graduate School of Accounting, College of Commerce, National Chengchi University, Taipei 116, Taiwan

ARTICLE INFO

Article history:
Received 7 May 2009
Received in revised form 2 March 2011
Accepted 13 July 2011
Available online 31 August 2011

Keywords:
Topic detection and tracking
Text mining
Information retrieval
Novelty index
Published volume index
Aging theory

ABSTRACT

Emerging topic detection is a vital research area for researchers and scholars interested in searching for and tracking new research trends and topics. The current methods of text mining and data mining used for this purpose focus only on the frequency of which subjects are mentioned, and ignore the novelty of the subject which is also critical, but beyond the scope of a frequency study. This work tackles this inadequacy to propose a new set of indices for emerging topic detection. They are the novelty index (NI) and the published volume index (PVI). This new set of indices is created based on time, volume, frequency and represents a resolution to provide a more precise set of prediction indices. They are then utilized to determine the detection point (DP) of new emerging topics. Following the detection point, the intersection decides the worth of a new topic. The algorithms presented in this paper can be used to decide the novelty and life span of an emerging topic in a specific field. The entire comprehensive collection of the ACM Digital Library is examined in the experiments. The application of the NI and PVI gives a promising indication of emerging topics in conferences and journals.

1. Introduction

Tracking the evolution of a discipline and detecting the emergence of a main stream is important to researchers and scholars (Lee, Gosain, & Im, 1997). Knowledge in these areas can be accumulated based on experience and state-of-the-art techniques can be used to investigate trends and identify new research topics. However, before a new research topic can be identified, sometimes years can pass. It is when a topic is already in great demand, in other words, "hot" that it starts to attract the attention of many researchers. This is always a lag behind index and as the number of papers discussing the same topic increases, their influence decreases. Consequently, in this study we develop some novel topic detection indices using automatic approaches which are designed to help researchers detect upcoming topics and make decisions about pursuing them before they become popular.

1.1. Research background

Topic detection and tracking (TDT) is an important field that tracks the evolution of a topic. TDT was developed in 1996 by the Defense Advanced Research Projects Agency (DARPA). A pilot study in by Allan, Carbonell, Doddington, Yamron, and Yang (1998) laid the groundwork for this field, generated a small corpus of information and established a durable system. Although TDT research flourished from 1998–1999 (Lee, Lee, & Jang, 2007), many studies are conducted in new areas.

* Corresponding author.
E-mail addresses: 082435@mail.fju.edu.tw (Y.-N. Tu), seng@nccu.edu.tw (J.-L. Seng).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2011.07.006

Lee et al. have discussed the evolution of topics in information systems (IS) (Lee et al., 1997). They found that the focus in journals and magazines is on different themes, with the former focused on conceptual and abstract models, while the latter is focused on specific applications. It should be noted that academic themes show a tendency to vary more over time. Although Lee et al. discussed trends in topics related to IS, they lacked TDT techniques to deal with the time-consuming task of identification. Swan and Allan addressed the issue of how to automatically identify the timeline for a set of news stories (Swan & Allan, 2000). They used the χ^2 method to identify a burst of feature terms that appear more frequently at some point in time than at other times. Kleinberg proposed a method for analyzing document streams (Kleinberg, 2002). Morinaga and Yamanishi improved Kleinberg's approach (Morinaga & Yamanishi, 2004).

Related work can be roughly divided into three groups, those that use: (1) text mining and data mining approaches (Aurora, Rafael, & Jose, 2007; Chou & Chen, 2008; Clifton, Cooley, & Rennie, 2004; Franz & McCarley, 2001; Hatzivassiloglou, Gravano, & Maganti, 2000; Kollios, Gunopulos, Koudas, & Berchtold, 2003; Kuramochi & Karypis, 2004; Ozmutlu, 2006); (2) those that use time-line burst detection of feature terms and measurements (Chen, Luesukprasert, & Chou, 2007; Manmatha, Feng, & Allan, 2002; Wang, Zhai, Hu, & Sproat, 2007; Yang, Yoo, Zhang, & Kisiel, 2005); and (3) those that use combined content analysis or link analysis (Jin, Myaeng, & Jung, 2007; Jo, Lagoze, & Giles, 2007; Nallapati, Ahmed, Xing, & Cohen, 2008; Ontrup, Ritter, Scholz, & Wagner, 2008; Ozmutlu & Cavdur 2005; Steyvers, Smyth, & Griffiths, 2004; Stokes & Carthy, 2001; Wu, Chen, & Sun, 2004; Yang, Zhang, Carbonell, & Jin, 2002; Zhang, Surendran, Platt, & Narasimhan, 2008). The principal task of time-line burst detection of feature terms and measurement is to determine when or whether a topic is emerging, whereas others focus on detecting the burst of a new topic.

There has been some work extending tracking or detecting techniques to such areas as literature-related discovery (Kostoff, 2008; Kostoff, Bhattacharya, & Pecht, 2007; Kostoff, Briggs, Solka, & Rushenberg, 2008), topological analysis of citation networks (Shibata, Kajikawa, & Matsushima 2007; Shibata, Kajikawa, Takeda, & Matsushima, 2008; Shibata, Kajikawa, Takeda, Sakata, & Matsushima, 2010), and combinations of other techniques and assessment tools (Daim, Rueda, Martin, & Gerdsri, 2006; Kostoff, Briggs, Rushenberg, Bowles, Icenhour, et al., 2007; Kostoff, Briggs, Rushenberg, Bowles, Pecht, et al., 2007; Tran & Daim, 2008). This area of endeavor is called tech mining and has been applied not only to find research topic for papers but also to determine new technologies and new patents for industry, organizations and governments.

1.2. Research Issues

As we know, the lifecycle of a research topic can be expressed as an S-curve with five stages, which are the initial stage, early stage, expansion stage, maturation stage and decline stage (Braun, Schubert, & Kostoff, 2000). Although in the maturation stage and at the peak of its lifecycle a topic is a famous one, this also means that it will soon decline. Thus, it is important for researchers to discover potential research topic as they emerge in the expansion stage of their developmental life time.

This study develops a set of novel indices for identifying such emerging topics to help researchers to determine whether a topic has the potential to become a hot topic in its lifecycle. For example, if there are ten papers in which a topic is discussed, then the impact of a new paper on the same topic can be calculated to be 1/11 = 0.0909. IN contrast, when there are 1000 papers in which a topic is discussed, the impact of an additional paper can be calculated to be 1/1001 = 0.000999.

The concept of novelty as applied by Zhang, Callan, and Minka (2002) and aging theory as developed in a TDT by Chen, Chen, Sun, and Chen (2003) are of assistance when constructing an index for detecting emerging topics. Chen et al. (2007) used aging theory and term frequency to solve the problem of topic detection, and proved that aging theory was the best solution. Although they claimed that they could detect a topic while it was still emerging, but not before it emerged. Here, we try to use these newly developed indices to examine, from the novelty and published volume, the stage in the life cycle (expansion, maturation or decline) and to determine the topic's research potential based on these conditions. This method gives researchers internal as well as external suggestions to consider the potential of an emerging topic.

Based on the proposed indices we can determine whether the topic of a conference or journal paper is representative of a leading trend, and how long the trend will continue. This will help researchers make decisions about whether they should pursue a topic beforehand. Conversely, not all new topics are valuable so should not become a focus of considerable research. This study develops a detection table to solve this problem. Novelty, aging theory and the curve representing accumulative relative frequency are all used to develop an appropriate set of indices for detecting emerging topics.

1.3. Research approach

Inductive learning and deductive prediction methods of machine learning help in constructing predictable indices and determining whether they are feasible. Whether a topic will develop is based on discussion in papers within the same period. That is, the volume of published studies on a topic helps determine whether it is important.

Additionally, whether a topic has potential is partly based on its novelty. Generally, as the number of times a topic is accessed increases, its potential value increases. A valuable topic attracts the attention of many researchers. Novel topics have considerable content that has not yet discussed. Whether in empirical or theoretical studies, novel topics have not yet completely developed and there is large space in the real world for applications. Consequently, the novelty and volume of papers published on a topic are important indices for determining it has potential to become a hot new emerging topic. The novelty index (NI) and published volume index (PVI) are developed and utilized to identify the detection point (DP) of a topic. The DP

in a topic's life cycle will move as later papers continue to be published. The moving DPs form a detection period which represents that the topic is emerging and valuable. Its novelty and popularity (published volume) are both in the highest states in this period of time.

The NI in the period before a specific topic emerges is higher than after emerging. Conversely, the PVI is higher in the subsequent period than the previous period. Hence, when a large volume of works are published on a topic, it is an emerging topic. PVI increases with the publications in the previous period, DP forms and represents a topic at an early stage as an emerging topic. If delays in published volume are significant, the rise in the curve would be delayed until a subsequent period. Regardless of the time the PVI starts increasing, it can be viewed as representative of the topic. While the DP is still floating, It represents a topic continues evolving to emerge for a period of time. All lifecycle records of these topics in journal and conference publications in the ACM database (at least until 2007) are compiled to form a detection table. The DP for a topic as indicated in conferences and in journals can help in determining whether conferences or journals are the leading trend of the topic.

The remainder of this paper is organized as follows. Section 2 discusses the extension of the theory application, and the development of TDT techniques and aging theory utilized in the method. Section 3 describes how to develop and construct indices in order to detect emerging topics. Section 4 presents the experimental design, execution and experimental results. Section 5 discusses the implications and contributions of the study. Section 6 gives Applications and Implications of this work. Section 7 summarizes the concluding remarks.

2. Related works

The availability of large linked document collections, such as the World Wide Web, and specialized literature archives present new opportunities for mining knowledge about community activities. Topic discovery is an example of such knowledge mining that has recently attracted considerable research interest. Topics can be considered semantic units that function as the basic building blocks in knowledge discovery. Once discovered, topics can be utilized in various ways, including for information navigation, trend analysis, and high-level data description (Jo et al., 2007). Topic trend analysis has the following three steps: (1) topic structure identification to identify the main topic types and their importance; (2) topic emergence detection to detect the emergence of a new topic and determine how it grows; and, (3) topic characterization to identify the characteristics of each main topic (Morinaga & Yamanishi, 2004).

2.1. Topic detection and tracking

The dissemination and exchange of documents has become commonplace with the recent growth of the Internet, thus raising the significance of content analysis techniques. Topic analysis of, say, e-mails and news articles is an important research task (Cui & Kitagawa, 2005). Many researchers have focused on techniques for topic discovery, topic tracking, topic-based text segmentation and related issues. In addition to TDT, Malone, McGarry, and Bowerman (2006) utilized data mining to analyze trends (2006). Aurora et al. (2007) developed a topic discovery system that reveals the implicit knowledge in news streams.

TDT is a recently developed information retrieval technology. It was developed in 1996 when the DARPA (Defense Advanced Research Projects Agency) was searching for a technique that could function without human intervention to detect topic structures in news streams. A pilot study in 1997 (Allan, Carbonell, et al., 1998) laid down the essential groundwork, generating a small corpus of knowledge and establishing a durable system. TDT research continued to flourish (Lee et al., 2007).

Makkonen et al. described the prevailing techniques applied in TDT such as formation extraction, retrieval and filtering, text clustering and text categorization and natural language processing. A TDT system that is implemented on-line does not have knowledge of unseen documents, which makes a case for clustering. Some studies have utilized retrospective topic detection and tracking when a system shows all data simultaneously; however, these studies focus mainly on on-line environments (Makkonen et al., 2004).

Allan, Papka, and Lavrenko (1998) described problems related to new event detection and event tracking within a stream of broadcasted news stories. They focused on an on-line setting, i.e., one in which the system makes decisions about one story before analyzing subsequent stories. They used a single-pass clustering algorithm and novel threshold model with event attributes as a major component. Their tracking approach is similar to typical information filtering methods. They determined the value of unique terms that had unusual occurrence characteristics, and applied on-line adaptive filtering to identify the evolution of events in the news. New event detection and event tracking are TDT initiatives.

Subsequent studies have improved TDT techniques. For instance, Walls, Jin, Sista, and Schwartz (1999) developed a system for TDT detection tasks for unsupervised groups of stories in the news and on web pages based on topics. Their system used an incremental *k*-means algorithm to cluster stories. A probabilistic document similarity metric and conventional vector space model was adopted to compare stories (Salton, Wong, & Yang, 1975).

Schultz and Liberman (1999) proposed approaches for detecting and tracking which are based on the well-known *idf*-weighted cosine coefficient similarity metric. They achieved excellent tracking results using a very simple term-selection method that did not involve word stemming or score normalization. However, their detection task results were poor, probably due to the poor performance of the clustering algorithm rather than that of the underlying similarity metric.

Some have found that while existing learning techniques must be adapted or improved to manage difficult situations in which each event has very few positive training instances, most training documents are unlabelled, and most events have short durations. Yang, Ault, Pierce T., and C. W. (2000) combined several supervised text categorization methods, namely, several new variants of the *k*-Nearest Neighbor (KNN) algorithm and the Rocchio approach, to track events. Their approach, based on a traditional parameter optimization solution, significantly decreased variance in the performance of their event-tracking system for different data collections.

Kleinberg (2002) proposed a method for analyzing document streams. Although the main objective was to detect bursts of topics, the method can be adopted for topic activation analysis. However, Kleinberg's method only considers document arrival rates, and disregards document relevance. Furthermore, the Kleinberg method is a "batch-oriented" approach. Cui and Kitagawa (2005) presented a solution to these problems. Although many studies have improved TDT techniques, these techniques are generally applied to time-sensitive documents (e.g., real-time news and e-mails) and have not been widely applied to identify new topics in academic papers.

2.2. Emerging topic detection

An emerging trend is a topic area that is growing in interest and utility over time. For instance, Extensible Markup Language (XML) emerged as a trend in the mid-1990s. Knowledge of emerging trends is particularly important to individuals and companies that monitor developments in a particular field or industry. For example, a market analyst specializing in the biotech industry may want to review technical and news-related literature for recent trends that will impact biotech companies. Manual review of all available data is simply not feasible. Human experts who must identify emerging trends must rely on automated systems as the amount of information available in digital resources is considerable (Berry, 2004).

Zhang et al. (2002) extended an adaptive information filtering system to make decisions regarding the novelty and redundancy of documents. They argued that relevance and redundancy should be modeled explicitly and separately. They developed a set of five redundancy measures which they evaluated in experiments with and without redundancy thresholds. Experimental results demonstrated that the cosine similarity metric and a redundancy measure based on a mixture of language models effectively identified redundant documents. Their research focused on the novelty and redundancy of documents, but did not address research topics.

Yang et al. (2002) proposed a novel two-stage approach that used (1) a supervised learning algorithm to classify on-line document streams into pre-defined broad topic categories and (2) performed topic-conditioned novelty detection for documents in each topic. They also exploited named entities for event-level novelty detection. Their study used the unit of a topic for novelty detection, but did not discuss when a topic is emerging.

Jo et al. (2007) generated unique approach that used correlation between the distribution of a term representing a topic and the link distribution in a citation graph in which nodes are limited to documents containing the term. This tight coupling between a term and graph analysis differed from other approaches such as those using language models. They applied a topic score to each item using the likelihood ratio of binary hypotheses based on a probabilistic description of graph connectivity. Their approach was based on the assumption that if a term is relevant to a topic, documents containing that term have a stronger connection than randomly selected documents. They applied the algorithm to detect a topic represented by a set of terms based on the assumption that if the co-occurrence of terms represents a new topic, the citation pattern should exhibit a synergy. They tested the algorithm on two electronic literature collections, arXiv and Citeseer. Their evaluation results showed that their approach was effective and revealed some novel aspects of topic detection. However, the curve which they only used term frequency to develop was still a lag behind index for detecting.

2.3. Aging theory

Capturing variations in a distribution of key terms on a time line is critical when extracting hot topics. Therefore, tracking terms to determine their lifecycle stage is essential. Previous studies have recognized that topics in a continuous document stream can be identified via a simultaneous temporal burst of related documents. There has been research applying Aging Theory to model the life span of a news event. In this respect a news event can be considered a life form that goes through a lifecycle of birth, growth, decay, and death, which is reflective of its popularity over time. This study utilized the concept of energy to track even lifecycles. The level of energy indicates the stage of a news event in its life span. The energy of an event increases as it becomes popular and decreases as its popularity wanes. Hence, Aging Theory is suitable for tracking variations in term frequency, which we consider critical to successful hot topic extraction (Chen et al., 2007).

2.4. Tech mining

Tech mining has been considered a new and import field since 2006 (Cunningham, Porter, & Newman, 2006). The topic of tech mining concentrates on the fields of science, technology and innovation, also called "ST&I". Tech mining uses text

mining techniques to inform or manage the knowledge obtained from searching electronic science and technology databases. Works in this area include fields such as information retrieval, scientometrics and content analysis (Porter & Cunningham, 2005). These techniques not only improve the decision-making processes of research studies but also can be applied to firms, organizations and institutions with extensive application in industry, government and academia. In this study, we track emerging topics in research work finding the indices for tech mining.

Zhu and Porter (2002) also discussed the automated extraction and visualization of technological intelligence and forecasting problems. They described a process for generating a technology family map and also describe a method to produce their innovation indicators. Their work can help researchers and decision makers realize the situation about the technology they are concerned with. The indications are developed as a composite concept.

Compare to his research, their relative accumulated frequency was set up different from ours, and they did not consider the detection point concept and growth analysis of a topic in this research.

Kostoff's related work (Kostoff, 2008; Kostoff, Bhattacharya, et al., 2007; Kostoff et al., 2008) applies text mining and information retrieval techniques to the field of discovery in science, which is a sub-field of tech mining. Literature-related discovery (LRD) methods involve the linking of two or more concepts that have heretofore not been linked by text mining procedures, in order to produce novel, interesting, plausible, and intelligible knowledge. LRD has two components: (1) Literature-based discovery (LBD) which generates potential discovery through literature analysis alone; (2) literature-assisted discovery (LAD) which generates potential discovery through a combination of literature analysis and interactions among selected literature authors. In turn, there are two types of LBD and LAD: (1) open discovery systems (ODS), where one starts with a problem and arrives at a solution; and (2) closed discovery systems (CDS), where one starts with a problem and a solution, then determines the mechanism(s) that links them.

Compared to this research, Kostoff extends the basic idea and combines the text mining and information retrieval techniques so as to retrieve core literature on target problem, characterize core literature by key researchers or identify centers of excellence through bibliometrics (which they called the thrust area), expand core literature and generate potential discovery by restricting classes of solutions. They try to find new science by looking at the literature connections, which is different from what we are doing in this current study. In this work we develop indices to identify a topic or new science whether it is during an emerging period in its lifecycle.

In related work, Shibata and Kajikawa and others (Shibata et al., 2007, 2008, 2010) analyzed the topology of the citation networks which they collected. They investigated the factors which determine the capability of academic articles to be cited in the future using a topological analysis of citation networks. Their work can also help to find emerging topic. The basic idea of their work is that articles that have many citations were in a "similar" position topologically in the past. They investigated the correlation between future times cited and three measures of centrality, which are clustering centrality, closeness centrality, and betweenness centrality. They also analyzed the effect of aging as well as of self-correlation of times cited. Their works suggested that times cited is the main factor in explaining the near future times cited, and betweenness centrality is correlated with the distant future times cited. The effect of topological position on the capability to be cited is influenced by the migrating phenomenon in which the activated center of research shifts from an existing domain to a new emerging domain.

This is different from the view in this current research which is aimed at discovering a new emerging domain. They used the citation network topology to observe the migrating phenomenon. This point of view looks at the lifecycle of the topic or domain itself, while we are both concerned with the novelty and popularity of a topic. Past research result may run into a risk when a topic is not yet migrating from another topology and just primitive novel data, the previous method has difficulty in identifying.

Daim et al. (2006) claimed that it is very difficult to forecast emerging technologies as there is no historical data available. They suggested that in this circumstance, bibliometrics and patent analysis are useful data resources for researchers or decision makers to generate useful information. They use three emerging technologies of scenario planning, growth curves and analogies, and data sources such as bibliometric, patent analysis and the modeling tool as system dynamics to present forecast tools to determine an emerging technology. Their works demonstrated that integrating multiple methodologies can improve the forecasting work, and that biliometrics and patent analysis can play a role providing historical data missing in the case of emerging technologies.

Daim et al.'s works helped the researcher to integrate multiple techniques to produce an assessment tool for forecasting the emerging technologies. In this current work we hope to give researchers a set of indices that can be used as a selection and assessment tool for choosing problems. Novelty and popularity are two important concepts in these indices which can provide a tool for considering the important properties in emerging technologies.

2.5. Summary

The research focuses on the second task in TDT, emerging topic detection. We attempt to detect the emergence of a new topic and determine its growth stages. The concepts of novelty concepts, aging theory and traditional frequency are applied to academic research topics. Aging theory is used in the construction of the novelty index. This study differs from previous work where only the frequency term was used. Instead we use a curve indicating accumulative relative frequency to develop the PVI to create the emerging topic detection indices.

3. Developing the Indices for detecting emerging topics

The theoretical basis and experimental design for evaluating the effectiveness of the indices for emerging topic detection are clarified in the following sections. Section 3.1 describes the emerging topic detection indices and the theoretical basis underlying the development of the indices. Section 3.2 illustrates their properties. Section 3.3 presents the information produced by these indices. Section 3.4 shows how the emerging topic detection table is constructed using the emerging topic detection indices.

3.1. Novelty of emerging topics

We create an NI and a PVI, which are related to the development of the emerging topic detection indices, to construct an approach for investigating the novelty of a research topic, that is, whether it is emerging.

3.1.1. Term, candidate research topic, research topic, hot topic and emerging topic

Before discussing the NI and the PVI, this study defines a set of terminologies: term, candidate research topic, research topic, emerging topic and hot topic. A term is defined as a set of words or an abbreviation that is mentioned more than three times in the same conference or journal publications in the same year. This threshold is based on the work of Joachims (1998). A term may be a single word, composite word or abbreviation of a proper noun. A candidate research topic is defined as terms that are composite words or an abbreviation of a proper noun as extracted from conference or journal publications. A candidate research topic indicates that a term may be an important research topic.

A research topic is defined as the intersection set between a candidate research topic in conferences and journals. These sets are of assistance when examining the leading relationship between research topics that appear simultaneously in conference and journal papers. Additionally, this study considers overlapping candidate research topics (in conferences and journals) as important research topics. A hot topic is defined as a topic that appears frequently within a given period (Chen et al., 2007). No topic can remain hot indefinitely; restated, each topic has a lifecycle comprised of birth, growth, maturity and death. A hot topic is one in the mature stage. An emerging topic is defined as a research topic that is important and in the growth stage but still not a hot topic. This study extracts emerging topics using the proposed indices.

3.1.2. Novelty index

Before defining the NI, we should define what the potential development year is. The potential development year (PDY) is defined as the period from the first year to the current year when a topic becomes a research topic that does not include any year with zero papers in the following years. Let us consider the research topic "XML" is as an example. Since "XML" is a well known and established topic researchers in the field have reached a consensus in terms of the lifecycle of its emergence.

Table 3-1 presents the datasets collected from the ACM database. The column entitled Type separates conference and journal papers, where "J" represents journal papers and "C" represents conference papers. This study selects and records the published volume of each paper type and each year separately. The value of type J in 2008 is 12, indicating that there were 12 journal papers focused on XML in that year. Comparatively, there were 114 conference papers focused on XML in 2008. One can identify the first paper referring to XML in conferences using Table 3-1, which was published in 1989. However, no paper discussed XML from 1990–1993. There was a paper in 1994 but none from 1996–1998, indicating that if the first paper caught the attention of researchers, it cannot be considered the start of an emerging topic.

Hence, if the PDY is 1989, it does not have the deterministic evidence in the research. However, after 1999, XML was the main topic in a considerable number of papers. Furthermore, we assert that if a topic is not discussed in any paper during whole year, that topic does not have the potential to be an emerging topic at that time. Therefore, this study defines the first year of the PDY as 1999 for conferences and 2001 for journals. The NI is defined as the inverse of the PDY. The NI indicates whether a topic is novel. For example, if the PDY of a topic is 5, the NI of the topic in the 5th year is 1/5 = 0.2. That is, in its nth development year, the NI is $1/n$. We use the proposed Algorithm 3–1 to identify first PDY.

Table 3-1
The volume of published papers on XML in each year: an example.

Type		Year									
		2008	2007	2006	2005	2004	2003	2002	2001	2000	1999
J_j	XML	12	11	20	15	11	7	11	3	0	1
C_i	XML	114	152	155	191	179	147	156	78	30	11
		Year									
		1998	1997	1996	1995	1994	1993	1992	1991	1990	1989
J_j	XML										
C_i	XML	0	0	0	0	1	0	0	0	0	1

Algorithm 3-1: Identifying which year is the C_F for the research topic
Input: C_i, the published volume of conference papers in the ith year
Output: C_F, the first PDY in conference papers for a research topic
1 For i = 2008–1989
2 If $C_i > 0$ then
3 $C_F = i$
4 Else
5 Return $C_F = i + 1$
6 Break
7 End If
8 Next

where

- C_i is defined as the published volume of conference papers in the ith year. For XML, i = 1989, . . . , 2008.
- J_j is defined as the published volume journal papers in the jth year. For XML, j = 1999, . . . , 2008.
- C_F is defined as the first PDY in conference papers for a research topic.
- J_F is defined as the first PDY in journal papers for a research topic.
- $CNI_k(Topic)$ is defined as the NI for the kth year for a research topic in conference papers.
- $JNI_k(Topic)$ is defined as the NI for the kth year for a research topic in journal papers.

We assume that a research topic is new when it is first published; thus, NI should be normalized to 1 = 100%. In its second year, the NI should be 1/2 = 50%. The value of the NI should be normalized to 0–1. By Algorithm 3–1, the formula for CNI_k(-*Topic*) is as follows:

$$CNI_k(Topic) = \frac{1}{k - C_F + 1}. \tag{3-1}$$

Furthermore, the formula for $JNI_k(Topic)$ is

$$JNI_k(Topic) = \frac{1}{k - J_F + 1}. \tag{3-2}$$

Taking XML (Table 3-1) as an example, C_F is 1999 and, using Algorithm 3–1, we start from the year 2008. When $C_{2008} = 114 > 0$, C_F will be 2008 temporarily. Next keep searching using i = 2007 until i = 1998 while $C_{1998} = 0$. The loop will break and return to $C_F = i + 1$ = 1999, indicating that there are no papers in the ith year focused on XML. It was not the first PDY, so actually C_F is next year of i as $i + 1$.

After determining that 1999 is the C_F year for XML, then $CNI_{1999}(XML) = 1$. Comparatively, to compute the conference Novelty Index (CNI) for 2008, we take k = 2008 in Formula (3–1), obtaining a CNI of 2008. $CNI_{2008}(XML) = \frac{1}{2008-1999+1} = \frac{1}{10} = 0.1 = 10\%$. Table 3-2 shows the calculated NI for each year using Algorithm 3–1, and Formulas (3–1) and (3–2).

3.1.3. Published volume index

The NI is a measurement of novelty. This study can determine whether a research topic is emerging or hot based on the volume of papers published in the same period. Conversely, if a topic is discussed over a long period in a vast number of publications, it is likely mature and well developed. It may cross or enter another domain so the focus would turn to a combination and applications with other domains, not only the ontology itself. Consequently, if the focus is solely on the volume of published papers, one cannot determine whether the topic has potential research value. Notably, as the volume of papers increases, topic impact decreases. The conventional frequency curve method for determining the volume of published papers lacks the ability to determine whether a topic is hot or emerging. When the PVI declines, we conclude that this has been a hot topic. The traditional frequency curve is a backward index.

Table 3-2
The NI of XML example in each year.

Type	Year									
	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999
J_j	12	11	20	15	11	7	11	3		
JNI_k(XML)	0.125	0.143	0.167	0.200	0.250	0.333	0.500	1.000		
C_i	114	152	155	191	179	147	156	78	30	11
CNI_k(XNL)	0.100	0.111	0.125	0.143	0.167	0.200	0.250	0.333	0.500	1.000

We not only consider the volume of published papers but also a topic's hotness over time. The accumulative relative frequency curve reflects the variation in published volume. This method is better than the frequency of a topic at detecting topic status based on comparing the growth situation without waiting until it really becomes mature. The PVI is defined as the accumulative relative frequency of the kth development year normalized to 0–1. Algorithm 3–2 is used to compute the PVI for the kth year, and Formula (3–3) comprises the equations for the CPVI.

Algorithm 3-2: Compute the PVI, taking the $JPVI_k(Topic)$ as an example
Input: Sum_J, Sum_i, J_i
Output: $JPVI_k(Topic)$
1 For $i = J_F$ to k
2 $Sum_J = Sum_J + J_i$
3 Next
4 For $i = J_F$ to k
5 $Sum_i = Sum_{i-1} + J_i$
6 $JPVI_i(Topic) = \frac{Sum_i}{Sum_J}$
7 Next

where

- Sum_C is the accumulated number of conference papers from C_F to the kth year.
- Sum_J is the accumulated number of the journal papers from J_F to the kth year.
- Sum_i is the accumulated number of papers from first year to the ith year for the same paper type.
- $CPVI_k(Topic)$ is the PVI for a topic in the kth year in conference publications and formulated as follows:

$$CPVI_i(Topic) = \frac{Sum_i}{Sum_C}. \tag{3-3}$$

- $JPVI_k(Topic)$ is the PVI of a topic in the kth year in journal publications and formulated as follows:

$$JPVI_i(Topic) = \frac{Sum_i}{Sum_J}. \tag{3-4}$$

This study takes XML as an example to illustrate the application of the PVI in Table 3-2. As discussed in Section 3.1.1, the question of concern to researchers is when a topic becomes an emerging topic with no break in subsequent years. The first time a research topic is discussed is not necessarily an important point. 1999 is identified as the first year in which XML appeared in journals using Table 3-1. We use Algorithm 3–1 to find $J_F = 2001$, the real PDY. PDY n in 2001–2008 is 2008–2001 + 1 = 8. Table 3-3 is used with Algorithm 3–2 to compute the PVI for each year. For 2003, $J_{2003} = 7$, $Sum_{2003} = 21$ and $Sum_{2003} = Sum_{2002} + J_{2003} = 14 + 7 = 21$. For 2008, $Sum_{2008} = 90$ is calculated as $Sum_J = 90$, which is the sum from the first year, 2001 to 2008. Hence, $JPVI_{2003}(XML) = \frac{Sum_{2003}}{Sum_{2008}} = \frac{21}{90} \approx 0.233$. Similarly the PVI ofor other years is computed in the same manner and recorded in Table 3-3.

3.1.4. Detection point

According to the two proposed detection indexes, NI and PVI, when the PDY is early compared to its lifecycle, the NI is high (Table 3-2). For example, $J_F = 2001$ and $JNI_{2001}(XML) = 1$. Compared to $JNI_{2002}(XML) = \frac{1}{2} = 0.5$, when the PDY is late relative to its lifecycle, and then the NI decreases. Conversely, Table 3-3 indicates that the situation for the PVI is opposite. Since the published volume reveals the amount of discussion a research topic receives, the PVI reflects the relative degree of growth in volume. The two indices can use the values in Table 3-4 to determine the development of XML.

Using the data in Table 3-4 we can draw the curves for JPVI (Journal PVI), JNI (Journal NI), CPVI (Conference PVI) and CNI (Conference NI). This study discovers that the NI and PVI is the trade-off curve, that is, a new topic lacks the volume needed to be a hot topic, and when a hot topic exists for a period of time, it loses its novelty. Consequently, the maximal NI and PVI

Table 3-3
The PVI of XML example in journals for each year.

	Year							
	2008	2007	2006	2005	2004	2003	2002	2001
J_j	12	11	20	15	11	7	11	3
Sum_i	90	78	67	47	32	21	14	3
$JPVI_j(XML)$	1.000	0.867	0.744	0.522	0.356	0.233	0.156	0.033

Table 3-4
The values of indexes for XML example in emerging-topic detection.

	Year									
	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999
J_j	12	11	20	15	11	7	11	3		
Sum_i	90	78	67	47	32	21	14	3		
$JPVI_j$(XML)	1.000	0.867	0.744	0.522	**0.356**	**0.233**	0.156	0.033		
JNI_j(XML)	0.125	0.143	0.167	0.200	**0.250**	**0.333**	0.500	1.000		
C_i	114	152	155	191	179	147	156	78	30	11
Sum_i	1213	1099	947	792	601	422	275	119	41	11
$CPVI_i$(XML)	1.000	0.906	0.781	0.653	0.495	**0.348**	**0.227**	0.098	0.034	0.009
CNI_i(XML)	0.100	0.111	0.125	0.143	0.167	**0.200**	**0.250**	0.333	0.500	1.000

The bolded value is where the DP decreases.

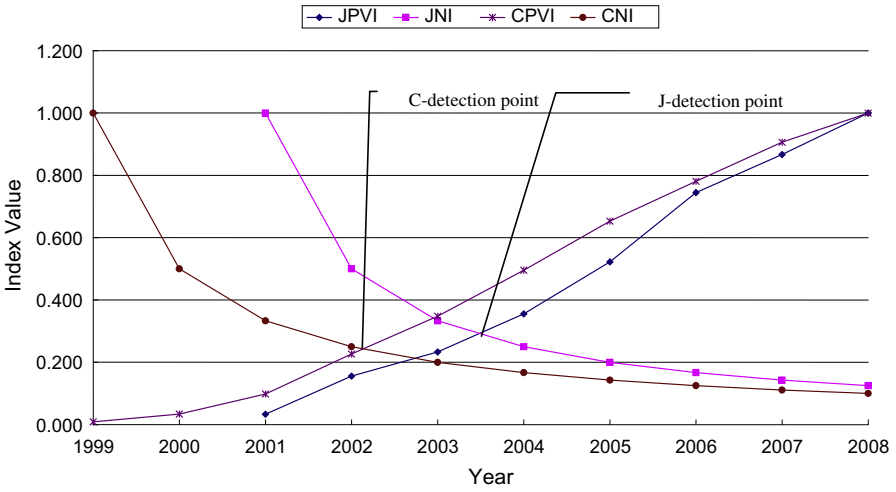

Fig. 3-1. The detection point for the emerging topic detection index for XML.

must be obtained from the intersection of curves, which is called the DP. Fig. 3-1 shows the detection point for the emerging topic detection index for XML.

The DP is defined as the point at which the NI and PVI intersect. We suggest that the DP can be used to determine whether a topic is hot and emerging, as it is the maximal value of two indices for both sides of novelty and hotness. The DP also separates the conference detection point (CDP) and journal detection point (JDP). Algorithm 3–3 shows how to compute the DP for JPVI and JNI.

Algorithm 3-3: How to compute the DP ofor JPVI and JNI
Input: present year, $JPVI_i$, JNI_i, $JPVI_{i+1}$, JNI_{i-1}
Output: *J*-detection point
1 For i = present year To C_F
2 If $JPVI_i = JNI_i$ Then
3 Return to the *J*-detection point = i
4 Else If $JPVI_i > JNI_i$ and $JPVI_{i+1} < JNI_{i-1}$ Then
5 Return to the *J*-detection point = $\frac{i+1}{2}$
6 End If
7 Next

3.2. Information produced by the emerging topic detection indices

According to the DP indices, if a topic has been published in both conference and journal papers, then, based on the published data, four curves can be drawn to obtain the JDP and CDP using Algorithm 3–3. Additionally, the indices also generate the year for the DP (YDP) and value of the DP.

3.2.1. Year of the detection point

The YDP is the *X*-axis value of the DP. The YDP indicates that a topic has reached the emergence threshold in its development. The YDP can also be separated into year for the CDP (YCDP) and year for the JDP (YJDP). The YCDP is the *X*-axis value of the CDP. The YJDP is the *X*-axis value of the JDP. Regardless of the JDP or CDP, there will also be a value on the *X*-axis, which represents the year. Although the YCDP is near 2002, the graph shows that this is not a DP. The topic does not become an emerging topic until 2003. Consequently, this study takes 2003 as the YCDP and 2004 as the YJDP.

3.2.2. The detection point value

The value of the DP (VDP) is the value at which the DP intersects the *Y*-axis. This value indicates both the NI and PVI for the DP at the same time. Since we use the NI and PVI normalized to 0–1 and the *Y*-axis is their representation, the DP value can be expressed by the NI and PVI. Additionally, the VDP also means that the NI and PVI are equal at the DP. However, the VDP is divided into the value of the CDP (VCDP) and value of the JDP (VJDP). The VCDP is the value at which the DP intersects the *Y*-axis for conferences and is the same value as the CPVI and CNI. The VJDP is the value at which the DP intersects the *Y*-axis for journal papers and is the same value as the JPVI and JNI. Since this study uses year as a unit, and if the DP is not exactly at one year, it must be between two years. The VCDP is between 2002 and 2003, and the value is affected by $CPVI_{2002} = 0.227$, $CPVI_{2003} = 0.348$, $CNI_{2002} = 0.250$ and $CNI_{2003} = 0.200$ (Fig. 3-1). The exact value of the VCDP is the center of those 4 points and is calculated as follows:

$$\frac{CPVI_{2002} + CPVI_{2003} + CNI_{2002} + CNI_{2003}}{4} = \frac{0.227 + 0.348 + 0.25 + 0.200}{4} \approx 0.256.$$

Likewise, we compute VJDP = 0.293. Hence, if the YDP is the *i*th year, Formula (3–5) can be used to compute the VDP

$$\frac{CPVI_{i-1} + CPVI_i + CNI_{i-1} + CNI_i}{4}. \tag{3-5}$$

3.3. The properties of emerging topic detection indices

After creating the NI and PVI to construct the emerging topic detection indices and detection table, we can analyze the academic publications and forecast the trend.

3.3.1. Novelty index properties

This study defines NI = $1/n$, where PDY is n. We suggest that this is a curve that can be used even if this is not verified. Since it is supposed that regardless of conferences or journals there exists a relationship between them. Furthermore, the NI will produce the same result for the relationship of conferences and journals with any validated index. Nevertheless, we assert that the NI is a reasonable and convenient index. To determine the entire lifecycle of a topic, one must obtain the termination date at which volume is 0, as well as the novelty for each year based on the termination date. For instance, if one knows that a topic has developed for 10 years, and that NI = 1 in the first year and NI = 0.9 in the second year, this process continues until the last year. However, one cannot determine when a topic terminates until it is terminated. Therefore, using NI = $1/n$ can avoid this lack. We suggest that novelty decreases as the PDY increases. Hence, regardless of the topic, the impact of the NI is $1/n$ at the nth PDY, and the NI is 1 in the first year, and in the second year, it is 1/2 = 0.5.

3.3.2. Published volume index properties

As mentioned, comparing the PVI and the traditional frequency measure can improve the forward effect. Here, XML is used as an example to describe the properties of the PVI. The data in Table 3-5 illustrate how the PVI reflects the emergence of XML.

The curve for Original-2006 in Fig. 3-2 is derived from journal data for XML during 2001–2006. The curve Decrease-2008 indicates that the amount of data decrease after 2006. The amount of data in 2007 is 1/2 of that in 2006 (10) and in 2007 it is 1/2 that in 2008 (5). The other situation is Increase-2008, which indicates that the amount of data increases after 2006; thus, the amount of data in 2007 is 2 times that in 2006 (40) and that in 2008 is 2 times that in 2007 (80). Thus, PVI-2006, PVI-2008-decrease and PVI-2008-increase are the indices for Original-2006, Decrease-2008 and Increase-2008, respectively.

While the volume of PVI increases relative to that in the past, like PVI-2008-increase, an upward opening concave curve forms. Conversely, when the volume of PVI decreases relative to that in the past like PVI-2008-decrease, then a convex curve downward opening curve forms. Consequently, as the volume of PVI is comparatively larger compared to the value in 2006 between PVI-2008-decrease and PVI-2008-increase, and the curve will rise from year 2006, indicating that the topic is becoming a hot topic. Conversely, the proportion of PVI-2008-increase in 2006 is lower than in the past – the largest volume of topic exists after 2006 – so the curve is relatively flat in 2006, indicating that topic has not yet become a hot topic.

3.3.3. Detection point properties

The DP is the intersection of the NI and PVI, and produces the YDP and VDP. The discussions in Sections 3.2.1 and 3.3.2 refer to the properties upon which the YDP is based. The accumulated relative frequency is used to determine the DP properties and validate the effectiveness.

Table 3-5
The PVI table of different situations in XML example.

	Year							
	2001	2002	2003	2004	2005	2006	2007	2008
Original-2006	3	11	7	11	15	20		
Decrease-2008	3	11	7	11	15	20	10	5
Increase-2008	3	11	7	11	15	20	40	80
PVI-2006	0.04	0.21	0.31	0.48	0.70	1.00		
PVI-2008-decrease	0.03	0.16	0.23	0.36	0.52	0.74	0.94	1.00
PVI-2008-increase	0.02	0.08	0.13	0.19	0.28	0.40	0.57	1.00

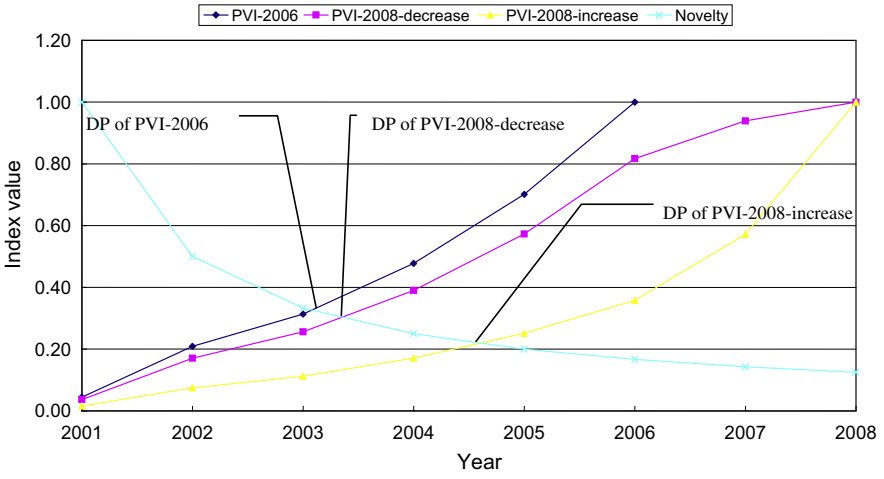

Fig. 3-2. The PVI curves for different situations for XML.

3.3.3.1. As the YDP increases, the DP is delayed. This study compares the curves of PVI-2006 to those of PVI-2008-decrease and PVI-2008-increase. Regardless of whether the amount of data increases or decreases, as long as a topic keeps developing (published volume is not 0), the curve will delay the intersect point. This makes sense because a later YDP means the topic has not yet reached the highest point in its lifecycle and growth stage. Conversely, for PVI-2006, the DP must intersect before 2006 when the YDP is 2004.

3.3.3.2. Increase in frequency for the year means the entire curve will rise. Consider PVI-2008-decrease. The highest value is produced in 2006 and the curve intersects in front of the DP of PVI-2008-increase. For the case in which the topic is in its mature stage, then the curve falls. Conversely, PVI-2008-increase indicates that 2006 was not the highest year in terms of its lifecycle. The highest volume in its lifecycle was reached in 2008. The delayed DP indicates that the topic is not hot.

3.3.3.3. The DP time. When a topic has a high PVI in its early stage, the curve will increase and the DP will form in the former part of the curve. This indicates that the topic is becoming hot at that time. Comparatively, if the PVI is high in a late stage, the curve will delay the DP. However, when the PVI curve starts to increase, it indicates that the topic is being discussed and is an emerging topic. The highest stage is not a point of concern as the topic is already mature. The DP represents the emerging topic produced as the DP is always in front of the present point and is a trade-off between the NI and the PVI.

We use the proposed emerging topic detection indices to examine the relationship between conferences and journals. If the reasoning is correct, regardless of how the NI is defined, the pattern of topics in conferences and journals will be the same. The DP of XML in this database is 2004, which is before the highest amount of data in 2006. Although we cannot determine whether XML has reached the highest volume in its history and could have a higher volume later, the DP is in 2004, which matches the expected date. Hence, the PVI has a better ability to predict an emerging topic than does the traditional frequency method.

The value of the DP, regardless of whether the NI or PVI, is maximal in the trade-off and can be used to detect when a topic is emerging. We assert that the DP must exist before the topic becomes hot. Consequently, the DP must exist during period from the first PDY to the present. Whether a topic becomes hot or not, the DP can still be calculated (as long as the PDY is more than 2 years) using the proposed indices. Hence, this study uses the YDP and VDP indexes to identify the situation in which a topic is hot. The emerging topic detection table is used to detect the value of retaining a research topic.

3.3.3.4. Detection period. Collecting all the DPs in each year for a topic, we can get a period called the detection period. Observing the detection period, we discover that the DP will move while papers on the topic continue to be published and discussed in the last topic's life cycle. This movement represents the growth of a topic. As we know, when a topic reaches maturation or is in the declining stage in its lifecycle, the following published volume will decrease and the PVI cannot continue extending by the potential developed year and the DP cannot make the YDP delay, which means cannot keep developing.

The detection period represents that the topic is emerging. Its value is highest at this point in its lifecycle because its novelty and popularity (published volume) are high during this period of time. When the DP is moving and the intersection delayed by the following published volume, it means that the topic is continuing to grow, and emerging for a period of time.

3.4. The emerging topic detection table

The emerging topic detection table helps in identifying the DP, which includes the YDP and VDP. By comparing conferences and journals, one can determine which reaches the threshold first. If a topic has never been an important topic, the DP year can still be calculated; however, it is worthless. Hence, this study uses the VDPs of each year for conferences and journals, respectively, to develop the emerging topic detection table. Take the XML at conferences as an example. From C_F = 1999 to the present date of 2008, the PDY is 10 years. Thus, YCDP = 2003 and VCDP = 0.256. However, in 2003 or 2004, we still can compute another YDP and VDP at that time. This study uses the properties of VDP to construct the emerging topic detection table. When a topic develops over 10 years, the VDPs for each year can be derived. The study identifies all research topics in the ACM database and computes their VDPs for each year. The median VDP is used to avoid confounding by extreme data and to generate the emerging topic detection table.

The VDP represents a new topic. Although the volume is small, the PVI is high. For instance, if the PDY is 2 years and the first and second year volumes are 1, then $PVI_1 = 0.5$ and $PVI_2 = 1$. The VDP must be 0.5–1, but there is still room for discussion with this topic. A high VDP indicates that the topic is still worth investigating. By increasing the volume regardless of the NI or PVI, the curve will easily delay the DP. For instance, in the third year, 2 papers were published so $PVI_1 = 0.25$, $PVI_2 = 0.5$ and $PVI_3 = 1$, while $NI_1 = 1$, $NI_2 = 0.5$, and $NI_2 = 0.5$; thus, VDP = 0.5. The VDP decreases and the DP is delayed; therefore, as the VDP increases, a topic warrants further research.

Each topic has its own development time—some topics develop slowly, while others may generate considerable discussion when first published. By using the median VDP for each research topic, one can construct a baseline. Therefore, when the lifecycle stage of a topic is unknown, the emerging topic detection table can be used to compare the topic with the baseline. When the VDP of a topic is higher than the baseline, it is worth keeping; when the VDP of a topic is lower than the baseline, the topic is worthless. Comparatively, if the VDP of a topic from the start to end of its lifecycle is never higher than the baseline, it cannot be an important topic.

4. The research experiment

This section describes the experimental design, execution, data collection, and the development of the emerging topic detection indices. This section is organized as follows: this section illustrates the leading trend relationship between conference and journal papers. Section 4.1 gives the introduction of this research experiment. Section 4.2 describes the experimental design. Section 4.3 presents the experimental results. Section 4.4 shows how to apply the emerging topic detection table for predicting worthy topics, and outlines the value of this investigation.

4.1. Introduction

To verify the accuracy and effectiveness of the proposed indices for recognizing and predicting new trends, this study uses data on journal and conference papers listed in the ACM Digital Library and IEEE Computer Society databases. Table 4-1 shows the four descriptors of correlations between conference and journal papers.

Some leading correlation categories between the conference papers and journal papers:

- Conference papers lead conference papers which we call C → C.
- Conference papers lead journal papers which we call C → J.
- Journal papers lead conference papers which we call J → C.
- Journal papers lead journal papers which we call J → J.

Table 4-1
The four categories of correlations between conference papers and journal papers.

Leading following	Conference papers	Journal papers
Conference papers	C → C	J → C
Journal papers	C → J	J → J

What researchers most care about is that the lead time and the leading trend position can be firmly established. The lack with C → C is that the leading trend position can be firmly established based on the conference paper property. With J → C there is a time lag since the turn-around time is always longer for journal papers than conference papers, which we call the time lag. The correlation J → J also has the same problem with time lag. Based on the properties of conferences and journals, the correlation C → J (without the time lag problem and including the journal trend) is most convincing to scholars to establish a research position. If we can make sure that that relation is C → J, it can help detect new trends in conference papers verified by journal papers. Conference papers are the leading indicator.

This study focuses on identifying seed trends for a particular domain of study by exploring the relationship between conference papers and journal papers. A conference paper appears to mark the beginning of a research process. Therefore, we believe that conference papers represent trends. This study focuses specifically on data mining and information retrieval. Computer science includes many sub-domains. Specific sub-domains need to be selected to define the perimeter of this research. Since this study applies the techniques of data mining and information retrieval, these topics are the focus of discussion. These two sub-domains can be described with ten keywords.

The ACM Digital Library and IEEE Computer Society are adopted as databases for conference papers. These are two renowned academic communities within the domain of information systems and computer science, and hold extensive collections of conference papers in all formats. The ACM Digital Library, IEEE Computer Society, ProQuest and ScienceDirect Onsite are the four databases for journal papers. The two extra databases are included to complement the ACM Digital Library and IEEE Computer Society's sparse journal collections, and to generalize the research findings.

The titles of the papers are utilized as indicators for the extraction of information, since they are strongly representative of the entire article. TextAnalyst is utilized to analyze the words used in titles, and single out the ones that are repeated more than three times. The data are presented in matrix form, with columns representing features and rows representing papers. This matrix can be used to calculate the number of papers with each feature. Finally, the Cosine of Similarity between annual conference papers and journal papers can be obtained using years as a unit of measurement.

Similarities in the topics of conference papers in sequential years are identified. The research findings strongly support the assumption that 87.23% of the data nodes in 1990–2007 demonstrate that the topic for one year influences topics in future years. These findings indicate that researchers can recognize new trends in research topics by looking at conference papers. Furthermore, massive amounts of data can be efficiently processed automatically by computing the similarities between conference papers and journal papers, pinpointing the keywords and topics that would most often appear in future journals (Tu & Seng, 2009).

4.2. Experimental design

To verify the accuracy and effectiveness of this method, an experiment is designed to utilize the proposed indices. The experimental results obtained in this study are compared with those obtained in previous work from which one can determine whether the results are consistent.

4.2.1. Choosing the field and data resources

Before determining whether a topic is important we first choose the data field and database, in this case the ACM Digital Library. The ACM is the largest and oldest academic community in the field of education and computer science. It has had a platform for exchanging information, innovation and discoveries since 1947. ACM members belong to the information systems and computer science community, and include professors, professionals, and students in industry, academia and public services in over 100 countries.

The range of the data is defined by using this method to browse journals and transactions not included in the magazine published by the ACM that appear in its digital library. A total of 35 journals exist. The (IEEE)/ACM Transactions on Computational Biology and Bioinformatics (TCBB) and IEEE/ACM Transactions on Networking (TON) are not published by the IEEE, and the range of discussion is far exceeded the conferences held by the ACM. Conference data published by the ACM in 2007 are used. In total, the ACM held 137 conferences. Some conference papers in the database are not formal papers but rather are student papers, short papers, poster papers, keynote speeches, tutorials, and demo abstracts. These papers are not included in this study because they do not present new issues.

4.2.2. Selection of the descriptors

A paper needs descriptors that describe its contents. The research topic of a paper is extracted based on these descriptors. We use the following four descriptors referring to paper content:

1. *Title*: the title of a paper is treated as a condensed description of the entire text. Thus the essence of the paper is captured concisely within a limited number of words. Words are sometimes coined by the authors themselves, meaning new trends are concealed within titles.
2. *Abstract*: when papers are reviewed the abstract can be used to give a rough grasp of the content within a short period of time. Thus abstracts can illustrate the content of a research papers far more explicitly than the titles, but contain many times the words. Consequently, the impact of each word unit as indicative of the meaning of the paper is thus diluted.

3. *Keywords*: keywords have the highest density in knowledge, but cannot describe a new trend. Authors must identify the keywords from the abstract. The paper can then be searched according to its keywords. In other words, keywords are expressive words that are most widely adopted by researchers for a particular concept within the same domain. Therefore, researchers in a particular research domain take a long time and much effort to reach a consensus that enables concepts to be translated into keywords. This process is usually time-consuming. Therefore, keywords in research papers are understood to achieve a high density of knowledge, filtered and crystallized through various researchers to form a single accumulated consensus on a concept, enabling them to express the paper far more precisely than titles. However, keywords can rarely identify new trends. This is because keywords relate to well-known concepts. A long period of time is required for domain experts to reach a consensus about a concept. Therefore, this study concludes that keywords do not describe the content of a paper as well as its title does.
4. *Full Text:* The full text includes every concept the researcher uses concerning the subject, yet individual words embody very little substance. The full text obviously includes the integrity of the content, but it is a compilation of an immense load of information that far exceeds that of titles, abstracts and keywords. Therefore the degree to which each phrase can express the concepts of the paper is small. Using the full text to describe the content of a paper would waste resources and time.

Authors use keywords to characterize their papers. Consequently, when a term is a keyword, it becomes a backward term. Although the full text contains the most information, it is a low knowledge-density descriptor. Many words and terms can be used to represent a research topic. Hence, typical information retrieval techniques without human judgment tend to extract many terms that do not exactly represent a topic. Therefore, two descriptors, the title and abstract, are used in this pilot study. The descriptors are utilized for data mining and information retrieval. We search the ACM's digital library to identify journal and conference papers containing a term. The terms extracted from titles and abstracts in 2007 are used for comparison. Based on previous work we argue that knowledge density in the title is higher than that in the abstract. Additionally, this study finds that just because there are more words in the abstract, the information embedded therein is more complete than in a title. There are 12–25 times more terms without stop-words in the abstract than in the title. Therefore, this work uses the abstract as the study descriptor. And, Stop-words means the words which are needed but not important meaning in a sentence such as "the", "a", "an", "is"... and so on.

4.2.3. Investigating extracted topics

To avoid unnecessarily large term vectors, a word is treated as a term only if it appears in the training data at least three times, and is not a "stop-word" (e.g. "and", "or"), (Joachims, 1998). Candidate research topics are extracted from terms rather than frequently used words. Instead of a single word, a composite word or abbreviation is used as the candidate research topic. Although this approach will overlook topics represented by a single word such as ontology, a single word topic must be identified by a person. Candidate research topics comprised of composite words or abbreviations possess better properties than single word topics which require human judgment. Conferences and journals have their own candidate research topics. To determine which one is a leading trend, we examine the intersection of candidate research topics between conferences and journals. This intersection represents the research topics in this study.

We assume that a hot or important topic can be found in 2007. When a topic is hot or important, discussion will increase regardless of when the topic was introduced. The year 2008 was not chosen because it had not yet ended when this study was carried out meaning that collections for conferences and journals would not be complete. For each journal and conference, we assume that the research position is equal without considering priority and importance. Furthermore, the database only records the volume of papers, not the frequency that the terms occur in the document. Thus, regardless of how many times it is mentioned in a paper, a research topic is counted as one paper. In other words, we do not consider the weight of a topic. The approach mentioned above is used to identify research topics, which are then input into the ACM search engine where their YDP is recorded based on the type of conference or journal. Finally, a table is obtained that has the same format as Table 3-1.

After determining the volume of published papers in each year (see Table 3-1), Algorithm 3–1 is applied to determine which year is the C_F, and Formulas (3–1) and (3–2) are applied to compute the CNI_k and JNI_k. The values are listed in the NI table, which is formatted the same as Table 3-2. Algorithm 3–2 and Formulas (3–3) and (3–4) are used to compute $CPVI_k$ and $JPVI_k$, which are listed in the PVI table, which is formatted the same as Table 3-4. The values can be utilized to generate the emerging topic detection index. Finally, Algorithm 3–3 is used to calculate the JDP using JNI and JPVI, and the CDP using CNI and CPVI. Formula (3–5) is then used to compute the VDP for conferences and journals.

The emerging topic detection index helps in detecting the DP of a research topic in the YDP. We continue by using the YDP for conferences and journals to develop am emerging topic detection table for detecting whether a topic warrants further research. Each research topic that has a PDY exceeding 3 years is used to compute the VDP; the median of VDPs for the same year is used to construct the table. The reason we do not begin in the second year is if a topic has developed for only 2 years, its DP must be between the first and second year. After the third year, the VDP and DP will vary and NI_k and PVI_k fall into different blocks. The emerging topic detection table uses the median VDP for each year. However, if a topic only develops for 3 years, then the VDP for the fourth year and later will not use the value of the research topic.

4.3. Experimental results

This study examined 35 journal issues and 137 conferences, altogether 689 journal papers and 5154 conference papers from 2007. TextAnalyst is used to extract those terms mentioned more than 3 times in the same publication in a year. Single-word terms are deleted and composite words and abbreviations retained as candidate research topics. The number of candidate research topics from conferences is 1791 and that from journals is 311. There are 89 topics in the intersection set, which are viewed as the research topics. The intersection ratio for conferences is almost 5% and for journals is near 29%, indicating that there is more convergence in journal topics. Although the range is broader for conferences than journals, it is easier to discover new topics in conference papers and there is still more divergence than journals.

The study suggests that some journals may concentrate on well defined topics while conferences in the same year may look forward to new topics, so that two subsets are not very high. Another reason for this might be that there is more divergence for conferences than journals so that topics are scattered. There are many different fields that cannot be covered since the volume of journals per year is less.

There is only topic "cutaway illustrations" that started in 2007 that appears in conferences or journals. We assume that if a research topic is valuable, it will survive into 2007. The proportion of research topics that match this assumption is 98.88%. "Cutaway illustrations" is an exception since it started in 2007 with no data existing before 2007. Thus, cutaway illustrations cannot be viewed as an important and valuable research topic.

Using the CDP produced by *CNI* and *CPVI*, and the JDP produced by *JNI* and *JPVI*, we obtain the YCDP, which is the year of the CDP; YJDP is the year of the JDP. For the same research topic, the CDP and the JDP have a sequential relationship, which represents which type of curve generates the DP first. The YCDP and YJDP can be the determining point for which type of papers comprise the leading tread. Generally, we assume that the first paper published at a conference or in a journal will have the first DP. However, this is not exactly correct if we refer to the NI and PVI in the research.

Of the 89 research topics, only 5 are published in journals before conferences; 11 are published first at conferences. Moreover, 1 topic has a DP later than that for conferences, indicating that the first publication year is not the only factor to consider when determining the lead position. When the NI and PVI are also considered the outcome changes. In total, 87.64% of research topics appear first in conference publications. On average, conference papers are published 4.26 years ahead of journal papers, while if the journal has the leading topic, the journal papers are published 3.5 years ahead of conference papers.

This investigation confirms that researchers can discover new trends for research topics from conference papers. The research findings strongly support the hypothesis. 85.42% of the data nodes collected from 1990 to 2007 show that topics in conference papers influenced the topic of journal papers in the same year and for the following two years. In other words, researchers can mine new issues from conference papers.

Results from this study and previous studies can be used to validate the leading topic relationship. Here we investigate 89 research topics, while previous work focused on similarities over 3 years from 1991–2007. Although the data units are different, we find that conferences lead journals by 87.64% as compared to the findings in previous work (85.42%). This verifies the effectiveness of this method and indicates that the indices are useful and accurate.

4.4. How to use the emerging topic detection table to predict whether a topic warrants further research

The CDP generated from the *CNI* and *CPVI* and the JDP from the *JNI* and *JPVI* are used to generate the VCDP from CDP and the VJDP from JDP. This work computes the VDP of each topic based on its PDY. If a topic's PDY is 5 years, one can calculate

Table 4-2
The emerging-topic detection table in the experiment.

The year of the development	Journal	Conference	The year of the development	Journal	Conference
3rd year VDP	0.583	0.572	23rd year VDP	0.208	0.134
4th year VDP	0.465	0.458	24th year VDP	0.205	0.137
5th year VDP	0.396	0.385	25th year VDP	0.195	0.149
6th year VDP	0.355	0.338	26th year VDP		0.148
7th year VDP	0.326	0.304	27th year VDP		0.144
8th year VDP	0.299	0.271	28th year VDP		0.140
9th year VDP	0.291	0.257	29th year VDP		0.138
10th year VDP	0.277	0.236	30th year VDP		0.135
11th year VDP	0.255	0.203	31st year VDP		0.131
12th year VDP	0.256	0.191	32nd year VDP		0.126
13th year VDP	0.250	0.178	33rd year VDP		0.122
14th year VDP	0.234	0.168	34th year VDP		0.115
15th year VDP	0.218	0.150	35th year VDP		0.114
16th year VDP	0.213	0.153	36th year VDP		0.106
17th year VDP	0.197	0.138	37th year VDP		0.099
18th year VDP	0.208	0.160	38th year VDP		0.092
19th year VDP	0.229	0.146	39th year VDP		0.105
20th year VDP	0.227	0.144	40th year VDP		0.102
21st year VDP	0.224	0.139	41st year VDP		0.099
22nd year VDP	0.218	0.132			

Table 4-3
How to use the detection table: an example of virtual environment.

The year of the development	Journal	Virtual environments JVDP	Conference	Virtual environments CVDP
3rd year VDP	0.583	0.700	0.572	0.675
4th year VDP	0.465	0.592	0.458	0.625
5th year VDP	0.396	0.433	0.385	0.589
6th year VDP	0.355	0.330	0.338	0.433
7th year VDP	0.326	0.295	0.304	0.381
8th year VDP	0.299		0.271	0.229
9th year VDP	0.291		0.257	0.165
10th year VDP	0.277		0.236	0.167
11th year VDP	0.255		0.203	0.131
12th year VDP	0.256		0.191	0.113
13th year VDP	0.250		0.178	0.135
14th year VDP	0.234		0.168	0.120
15th year VDP	0.218		0.150	0.107
16th year VDP	0.213		0.153	0.102
17th year VDP	0.197		0.138	0.099

the VDP for each year for years 3–5. The 3-year VDP focuses on the first, second and third year, and ignores data for the fourth and fifth years. The 4-year VDP is then calculated and year 5 data is ignored. This process continues until the full PDY is obtained, and the development of each topic in its lifecycle can be determined for each year. Table 4-2 shows the baseline for the 89 research topics.

Each topic has its own start and emerging time. Consequently, the median VDP can be used to generate the baseline for a research topic in each year. If an unknown topic exists then one cannot determine whether it has high worthiness for continued investigation or not. This study calculates the VDP for a topic for each year in its development. If the VDP is higher than the baseline in any year, the topic should be investigated further as it may be a novel topic. Comparatively, if a topic is old and mature, it will have a low VDP. Furthermore, if a topic, from the beginning to the end of its history, never has a VDP higher than the baseline, then that topic is worthless. Based on a topic's development, one can determine whether the topic is valuable.

Take Virtual Environments (VEs) as an example. The VDP for each year is compared to the baseline (Table 4-3). The emerging topic detection table is used to determine whether VEs should be investigated further. The JVDP in the third year is 0.583. The JVDP of VEs is 0.7 higher than the baseline, meaning that VEs is valuable in its third year. The JVPD for the fourth year is 0.465, and 0.592 higher than the baseline, meaning it is a valuable topic. In year 6, the JVDP of VEs is 0.33, lower than the baseline value of 0.355, indicating that the topic of VEs is mature. Similarly, the CVDP in the table in the third year is 0.572 and the VCDP of VEs is 0.675, which is higher than that in the table. In its eighth year, the CVDP of VEs is lower than that in the table (0.271), indicating that the topic of VEs in conferences in the eighth year is mature.

5. Validating the accuracy and effectiveness of the emerging topic detection indices

In this section, we describe an experiment using this research model to validate the proposed indices. We survey previously published related works to find information about topics which have already been validated using other methods. We also survey some experts on the topic selected to validate the indices. The validation procedure for author impact power is illustrated in Fig. 5-1.

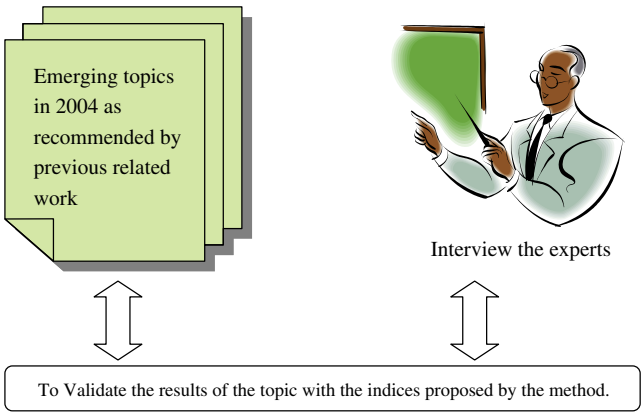

Fig. 5-1. Validation procedure for emerging topic detection indices.

5.1. Comparing the experimental results with previous work

In order to validate the accuracy and effectiveness of the emerging topic detection indices, we look for related work where emerging topics have also been detected but within a different research time range and field. The most similar work that we found is that of Jo et al. (2007) in SIGKDD (ACM SIGKDD Conference on Knowledge Discovery and Data Mining). Their approach is based on the intuition that documents related to a topic should be more cohesively connected in the citation graph than a random selection of documents. They used the Citeseer data which contains 716,771 papers, with 1740,326 citations. This amounts to 2.43 citations per paper. For each paper, we use the combined title and abstract for documentation. The number of bigrams in the corpus after pruning out the low-frequency bigrams and 35 stop words is 631,839. The majority of papers are from the years 1994 to 2004.

Besides the research approach, the database and the time range of the dataset are different from those used this study, but if we use the same time range to examine the research results, this approach can predict with precision the correct proposed emerging time of each topic. Their work contains both conference papers and journal papers. In order to map their work we selected topics which they claimed to have emerged from their study. These topics are image retrieval, sensor network, semantic web, support vector. The indices are then applied to find out the potential developed year and the published volume in each year for these topics in the ACM digital library. The NI and PVI of each topic proposed by their work are computed afterwards. A comparison of the results is discussed below.

5.1.1. Sensor networks

The topic sensor network is one of the most common emergent topics in their work. The emerging time falls in the year 2004. Fig. 5-2 shows the topic evolution of sensor networks over time. The original published conference and journal volumes are shown in Figs. 5-3 and 5-4, respectively. The year of the conference detection point is 2004. This is the same as in Jo et al. (2007) where the year of journal detection point is 2006. The results are shown in Fig. 5-5.

5.1.2. Semantic web

The topic of "semantic web", which emerged in the year 2004, is another common emergent topic that was discussed in their work. Fig. 5-6 shows the evolution of the topic semantic web over time. The original published volume of conferences and journal are shown in Figs. 5-7 and 5-8, respectively. The year for the conference detection point is 2004 which is the same as in Jo et al. (2007) and the year for the journal detection point is 2007. The results are shown in Fig. 5-9.

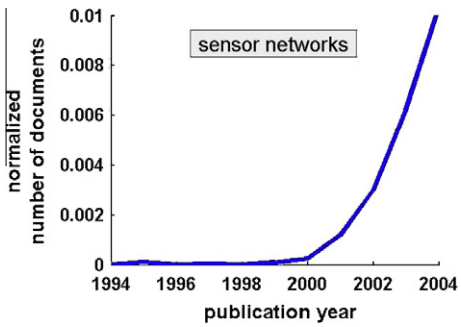

Fig. 5-2. Topic evolution of sensor networks over time from 1994 to 2004 (Jo et al., 2007).

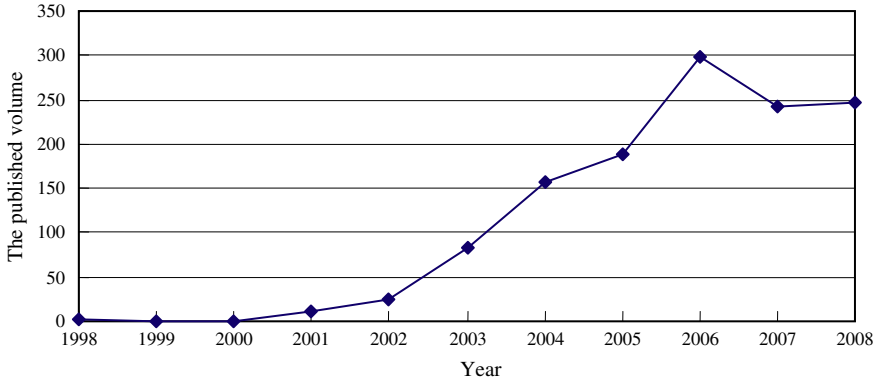

Fig. 5-3. Conference papers published volume for sensor networks in the ACM database.

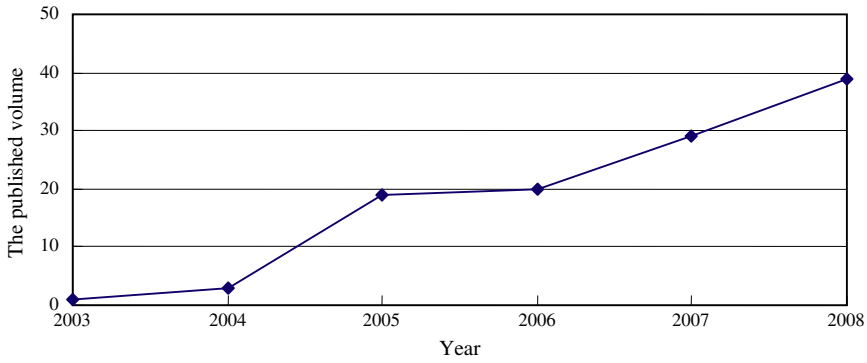

Fig. 5-4. Journal papers published volume for sensor networks in the ACM database.

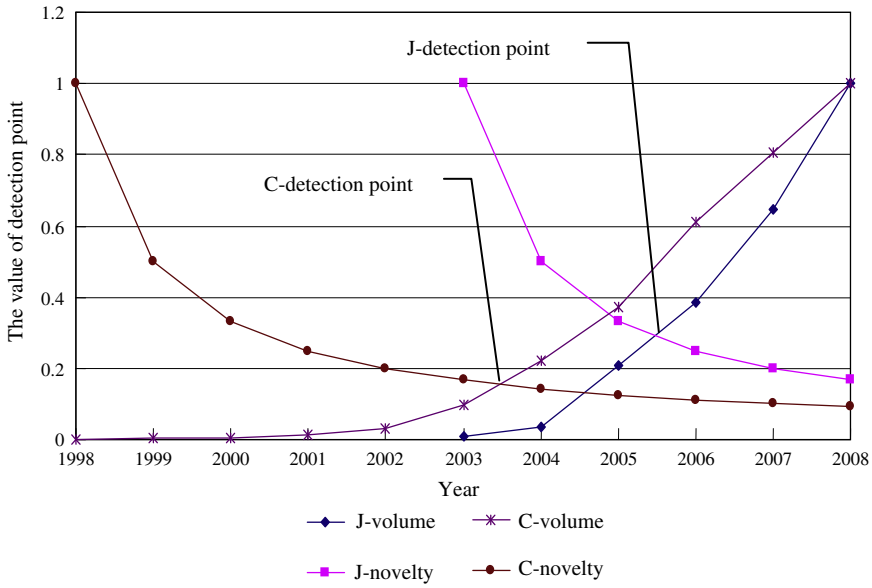

Fig. 5-5. Detection point of sensor networks in the ACM database.

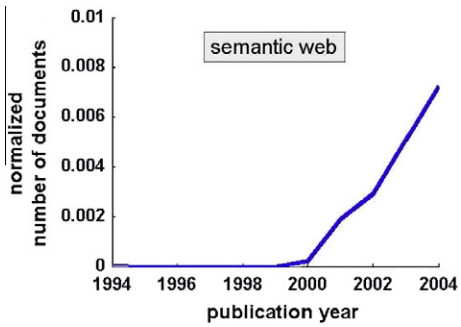

Fig. 5-6. Evolution of the topic semantic web over time from 1994 to 2004 (Jo et al., 2007).

5.1.3. Support vector

The topic "support vector" is another topic that emerged in the year 2004. However, they claimed that the topic support vector is not as obvious as for the previous two topics. The curve obtained is not as inclined as for the previous two topics, sensor network and semantic web. Fig. 5-10 shows the evolution of the topic of support vector over time. In this current

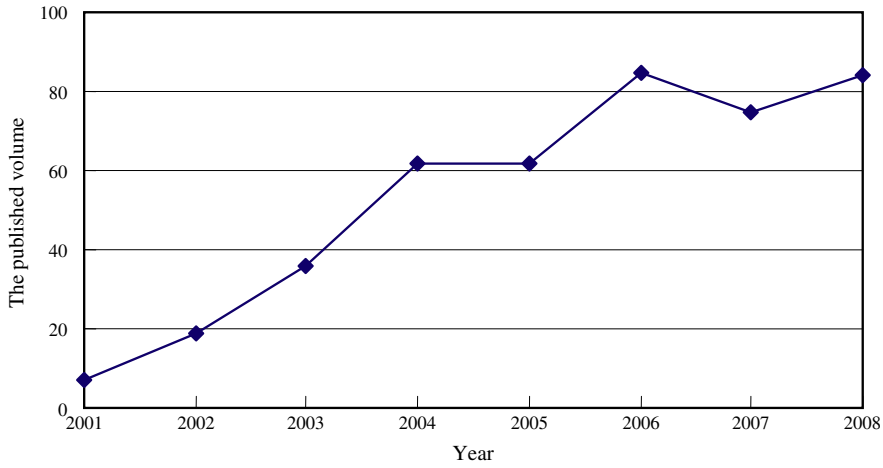

Fig. 5-7. Conference papers published volume for sensor networks in the ACM database.

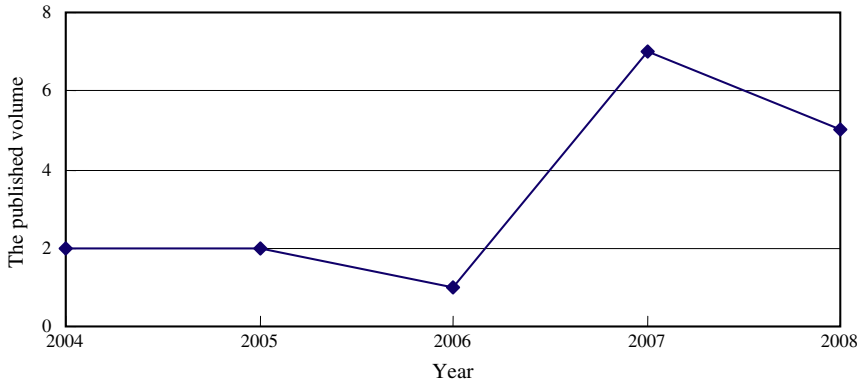

Fig. 5-8. Journal papers published volume for sensor networks in the ACM database.

research, the conference detection point is in 2004 which is the same as in Jo et al. (2007) and the journal detection point is in 2005. The results are shown in Fig. 5-11. All of these four topics which are considered by Jo et al. (2007) as emerging topics in 2004, are detected correctly using the emerging topic detection indices.

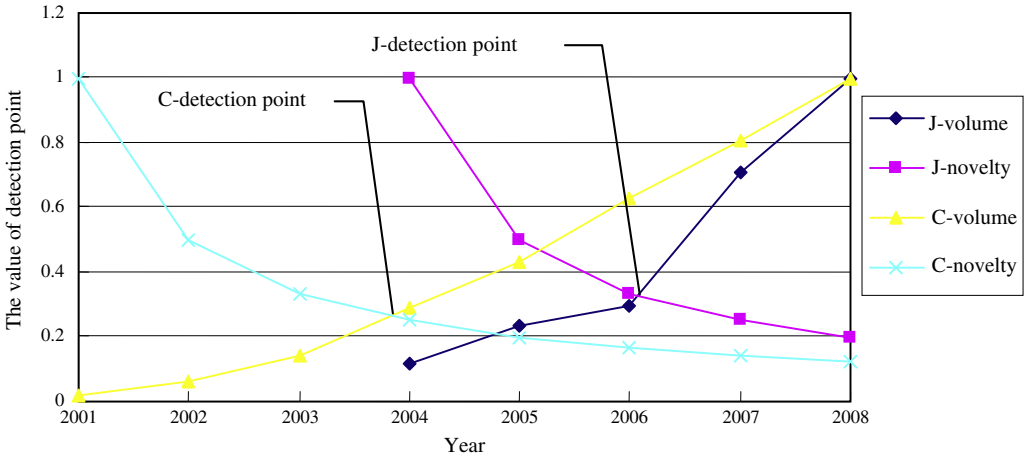

Fig. 5-9. Detection point of semantic web in the ACM database.

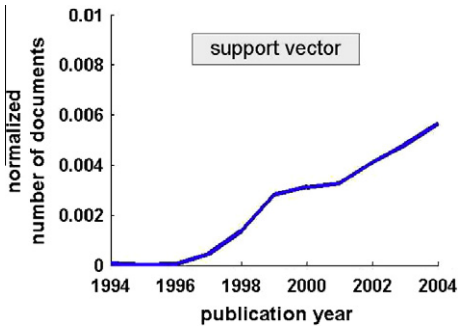

Fig. 5-10. Evolution of support vector over time from 1994 to 2004 (Jo et al., 2007).

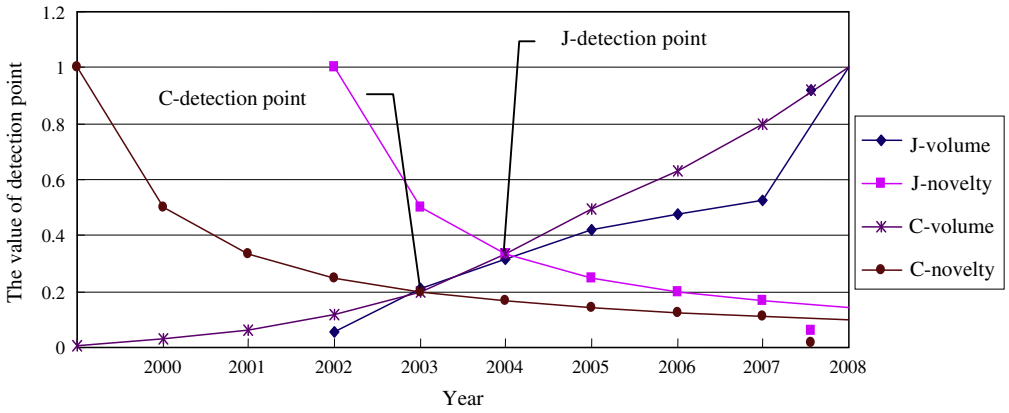

Fig. 5-11. Detection point of support vector in the ACM database.

Table 5-1
Expert survey results on emerging time.

No.	Questions	E1	E2	E3	E4	E5	Points
1	Do you agree the topic "Sensor Network" is emerging during the year 2004?	Y	Y	Y	Y	Y	5
2	Do you agree the topic "Semantic Web" is emerging during the year 2004?	Y	Y	Y	N	No opinion	3
3	Do you agree the topic "Support Vector" is emerging during the year 2004?	N	N	Y	Y	Y	3

5.2. Comparing the experimental results with the expert survey results

Besides comparing the experimental results with those obtained from previous works as noted in the literature review, we also survey five experts who have investigated the topic of data mining. We ask them to give opinions, including yes, no and no opinion to reflect whether each topic is emerging during a given period of time. These experts included two assistant professors, one associate professor and two full professors, three from the department of Management of Information Systems and two from Computer Science and Engineering. E1 means Expert 1, and Grade represents the votes which a topic receives. The expert survey results for 3 topics also validate these indices as shown in Table 5-1.

We give each topic a grade when one of the experts considers it to really be an emerging topic during the given period of period of time. We find that all topics all received more than 3 votes, which is more than half the number of experts' opinions. The results for the topic "sensor network", obviously emerging in 2004, are consistent. The topic "support vector" did not get more support because some of the experts were not sure whether support vector referred to the support vector machine or not. This is a limitation because of the meaning of the term. We only can extract terms from previous work and can not exactly make sure of their meanings.

6. Applications and Implications

The NI helps researchers to exam research topics from the view point of novelty and aging theory. The novelty and aging concepts should be considered while we discuss the emerging topic detection not only the hot topic detection. Emergency implies new and urgent.

The PVI differs from past simple frequency lines such as in the work of Jo et al. (2007) which only can tell how much the frequency is in each year. The PVI however, adopts the concepts of accumulated relative frequency for each year based on the volume of different PDYs. One can tell which topic is an emerging topic by the rising curve and which is a mature topic by the falling curve.

The combination of the NI and PVI can draw the detection point and the VDP since the DP means the NI and PVI are both at the highest value. The DP possessing the characteristics of being both novel and hot matches the expectation of emerging topics. The most important finding is the development of a method and indices which helps researchers to construct their own field topic detection tables and examine new topics in their own field.

The database used in this study includes various journals and conferences related to computer science from the ACM database of publications of journals and transactions, and conference proceedings. There were 689 journal papers and 5154 conference papers published in 2007. Conferences account for 1791 research topics, and journals for 311 topics. The intersection is 89 research topics. From the research topics intersection ratio, the ratio of conferences is almost 5% and the journal ratio is nearly 29%. Thus, there is more convergence in the topics discussed in journals and more divergent in conferences.

We demonstrate the development of emerging topic detection indices. The YCDP and YJDP can determine the lead relationship. The first DP in the curve, regardless of for the CDP or JDP, indicates the maximal value of the NI and PVI. Consequently, we suggest that the first DP is a leading position. The YDP can help to determine which type of paper is in the lead position. This study also proves that the first publication time is not the critical factor when determining the YDP. The NI and PVI affect the DP directly. In total, 87.64% of research topics represent that conference topics lead the development of journals. We use different research methods and different databases than in previous works which only focus on the leading relationship to generate a similar lead relationship results. The experimental results verify the accuracy and effectiveness of the proposed indices.

This study uses the NI and PVI to develop the emerging topic detection indices. The concepts of terms, topics, and candidate research topics are used to investigate topics, and we also discuss the CNI and CPVI for conferences and the JNI and JPVI for journals. The DP is produced even though the CDP or JDP creates the YDP and VDP to represent the year and the value when the topic emerges. The YCDP and YJDP can be used to determine which type of papers stand in the leading position. The VDP can be used to determine whether to investigate a topic further. Based on the NI and PVI, one can show that when the published volume for the present year is large, the curve will rise, indicating the DP in an early stage of its lifecycle, and decline to make the DP delay. Based on the NI and PVI and the properties of the DP, the emerging topic detection indices can help to determine whether a research topic is worth further investigation. A high VDP indicates that topic novelty is sufficiently high for further research.

Finally, the emerging topic detection indexes detect the DP and obtain the YDP and VDP. By comparing conferences and journals, one can determine which reaches the threshold first. However, if a topic has never been important, the topic DP is useless. The emerging topic detection table can be used to examine whether a topic warrants further research. Each topic has its own value; therefore, the value of researching a topic indicates that the topic has not reached the highest point in its lifecycle. When a topic is hot and mature, its potential worthy for further research will decrease in the future. The VDP is the basis of the detection table.

Even when the published volume is low, the PVI will be high since it compares to itself the total number that can decide the PVI. A high VDP represents a large development space for additional effort. Hence, a high VDP indicates that a topic warrants further investigation. Consequently, when one does not know whether a topic is important or worthy, one can compute its VDP and then compare this value with that in the emerging topic detection table. If the VDP is lower than the baseline, the topic is mature or worthless. Comparatively, if a topic's lifecycle is never higher than the VDP for the same year in the detection table, it has never garnered popular attention; thus, the topic is worthless.

The method presented in this study can also be applied in bibliometrics and patent analysis as in previously related works on tech mining. The method can be used by business analysts, inventors and governmental agencies and so on. For example, in publishing, novelty is a very important. The indices can help the publisher to know whether a series of reports related to some issue will increase or decrease. Are the customers getting tired of the same issue? The indices can help for stock price prediction by the financial analyst, to know which area of the market is over hot and where more investment is possible. Organizations and governments can use the indices to determine and realize novelty and the number of the inbound competitors in their enterprise. Governments also can use the indices to observe the development of social phenomenon such as economics and to make sure that they balance supply and demand in marketplace.

7. Concluding remarks

This study addresses the inadequacy of topic detection and tracking to develop a set of novel indices for emerging topic detection. The novelty concept is used in combination with aging theory to develop the novelty index (NI). The published volume index (PVI) is an improvement over traditional frequency methods to reflect the growth of a discussion topic. The DP and YDP help determine the relationship between conference topics and journal topics and how long they lead ahead. The VDP is created to construct the detection table to determine whether a topic warrants further research. The NI and PVI can be applied to other fields to determine new trends, for example, the news or stock price predictions.

The indices also have some limitations based on the units of the data sets which we can collect. For example, the dates of research papers are based on the year instead of months, and the DPs at year 2022.1 and year 2002.9 will be the same given these indices.

Future work will extend the NI and PVI with more diversified experiments. The set of novelty indices can be improved using other areas of training and testing models. A more complicated detection table can be generated.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, T. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription an understanding workshop*.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 37–45).

Aurora, P. P., Rafael, B. L., & Jose, R. S. (2007). Topic discovery based on text mining techniques. *Information Processing & Management, 43*, 742–768.

Berry, M. W. (2004). *Survey of text mining-clustering, classification, and retrieval*. Springer, pp. 185–224.

Braun, T., Schubert, A. P., & Kostoff, R. N. (2000). Grwth and trends of fullerence research as reflected in its journal literature. *Chemical Reviews, 100*(1), 23–27.

Chen, C. C., Chen, Y. T., Sun, Y., & Chen, M. C. (2003). Life cycle modeling of news events using aging theory. In *Proceeding of 14th European conference machine learning (ECML '03)* (pp. 47–59).

Chen, K. Y., Luesukprasert, L., & Chou, S. C. (2007). Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Transactions on knowledge and data enginerting, 19*(8), 1016–1025.

Chou, T. C., & Chen, M. C. (2008). Using incremental PLSI for threshold-resilient online event analysis. *IEEE Transactions on Knowledge and Data Engineering, 20*(3), 289–299.

Clifton, C., Cooley, R., & Rennie, J. (2004). Topcat: Data mining for topic identification in a text corpus. *IEEE Transactions on Knowledge and Data Engineering, 16*(8), 949–964.

Cui, C., & Kitagawa, H. (2005). Topic activation analysis for document streams based on document arrival rate and relevance. In *Proceedings of the 2005 ACM symposium on applied computing* (pp. 1089–1095).

Cunningham, S. W., Porter, A. L., & Newman, N. C. (2006). Introduction – Special issue on tech mining. *Technological Forecasting & Social Change, 73*, 915–922.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting & Social Change, 73*, 981–1012.

Franz, M., & McCarley, J. C. (2001). Unsupervised and supervised clustering for topic tracking. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 310–317).

Hatzivassiloglou, V., Gravano, L., & Maganti, A. (2000). An investigation of linguistic features and clustering algorithms. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 224–231).

Jin, Y., Myaeng, S. H., & Jung, Y. (2007). Use of place information for improved event tracking. *Information Processing & Management, 43*, 365–378.

Jo, Y., Lagoze, C., & Giles, C. L. (2007). Detecting research topics via the correlation between graphs and texts. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 370–379).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the EMNLP conference*.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 91–101).

Kollios, G., Gunopulos, D., Koudas, N., & Berchtold, S. (2003). Efficient biased sampling for approximate clustering and outlier detection in large data sets. *IEEE Transactions on Knowledge and Data Engineering, 15*(5), 1170–1187.

Kostoff, R. N. (2008). Literature-related discovery (LRD): Introduction and background. *Technological Forecasting and Social Change, 75*, 165–185.

Kostoff, R. N., Bhattacharya, S., & Pecht, M. (2007). Assessment of China's and India's science and technology literature-introduction, background, and approach. *Technological Forecasting and Social Change, 74*, 1538–1591.

Kostoff, R. N., Briggs, M. B., Rushenberg, R. L., Bowles, C. A., Icenhour, A. S., Nikodym, K. F., et al (2007). Chinese science and technology – Structure and infrastructure. *Technological Forecasting and Social Change, 74*, 1539–1573.

Kostoff, R. N., Briggs, M. B., Rushenberg, R. L., Bowles, C. A., Pecht, M., Johnson, D., et al (2007). Comparisons of the structure and infrastructure of Chinese and Indian science and technology. *Technological Forecasting and Social Change, 74*, 1609–1630.

Kostoff, R. N., Briggs, M. B., Solka, J. L., & Rushenberg, R. L. (2008). Literature-related discovery (LRD): Methodology. *Technological Forecasting and Social Change, 75*, 186–202.

Kostoff, R. N., Johnson, D., Bowles, C. A., Bhattacharya, S., Icenhour, A. S., Nikodym, K., et al (2007). Assessment of India's research literature. *Technological Forecasting and Social Change, 74*, 1574–1608.

Kuramochi, M., & Karypis, G. (2004). An efficient algorithm for discovering frequent subgraphs. *IEEE Transactions on Knowledge and Data Engineering, 16*(9), 1038–1051.

Lee, C., Lee, G. G., Jang, M. (2007). Dependency structure language model for topic detection and tracking. *Information Processing & Management, 43*, 1249–1259.

Lee, Z., Gosain, S., & Im, I. (1997). Topics of interest in IS: Evolution of themes and differences between research and practice. *Information & Management, 36*, 233–246.

Malone, J., McGarry, K., & Bowerman, C. (2006). Automated trend analysis of proteomics data using an intelligent data mining architecture. *Expert Systems with Applications, 30*, 24–33.

Makkonen, J., Ahonen-Myka, H., & Salmenkivi, M. (2004). Simple semantics in topic detection and tracking. *Information Retrieval, 7*(3–4), 347–368.

Manmatha, R., Feng, A., Allan, J. (2002). A critical examination of TDT's cost function. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 403–404).

Morinaga, S., & Yamanishi, K. (2004). Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 811–816).

Nallapati, R., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 542–550).

Ontrup, J., Ritter, H., Scholz, S. W., & Wagner, R. (2008). Detecting, assessing and monitoring relevant topics in virtual information environments. *IEEE Transactions on Knowledge and Data Engineering, 20*(7).

Ozmutlu, S. (2006). Automatic new topic identification using multiple linear regression. *Information Processing & Management, 42*, 934–950.

Ozmutlu, H. C., & Cavdur, F. (2005). Application of automatic topic identification on excite web search engine data logs. *Information Processing & Management, 41*, 1243–1262.

Porter, A. L., & Cunningham, S. W. (2005). *Tech mining: Exploiting new technologies for competitive advantage*. New York: Wiley-Interscience.

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

Schultz, J. M., & Liberman, M. (1999). Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of the DARPA broadcast news transcription an understanding workshop*.

Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2010). Detection emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting & Social Change*.

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology, 58*(6), 872–882.
Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detection emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation, 28*, 758–775.
Steyvers, M., Smyth, P., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 306–315).
Stokes, N., & Carthy, J. (2001). Combining semantic and syntactic document classifiers to improve first story detection. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 424–425).
Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 49–56).
Tran, T. A., & Daim, T. (2008). A taxonomic review of methods and tools applied in technology assessment. *Technological Forecasting & Social Change, 75*, 1396–1405.
Tu, Y. N., & Seng, J. L. (2009). Research intelligence involving information retrieval–An example of conferences and journals. *Expert Systems with Applications, 36*(10), 12151–12166.
Walls, F., Jin, H., Sista, S., & Schwartz, R. (1999). Topic detection in broadcast news. In *Proceedings of the DARPA broadcast news transcription an understanding workshop*.
Wang, X., Zhai, C., Hu, X., & Sproat, R. (2007). Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 784–793).
Wu, K., Chen, M., & Sun, Y. (2004). Automatic topics discovery from hyperlinked documents. *Information Processing & Management, 40*, 239–255.
Yang, Y., Ault, T., Pierce T., & Lattimer, C. W. (2000). Improving text categorization methods for event tracking. In *Proceedings of the 23th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 65–72).
Yang, Y., Zhang, J., Carbonell, J., & Jin, Chun. (2002). Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 688–693).
Yang, Y., Yoo, S., Zhang, J., & Kisiel, B. (2005). Robustness of adaptive filtering methods in a cross-benchmark evaluation. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 98–105).
Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 81–88).
Zhang, Y., Surendran, A. C., Platt, J. C., & Narasimhan, M. (2008). Learning from multi-topic web documents for contextual advertisement. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1051–1059).
Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting & Social Change, 69*, 495–506.