

A comparison of methods for detecting hot topics

YUEN-HSIEN TSENG,^a YU-I LIN,^b YI-YANG LEE,^c WEN-CHI HUNG,^d CHUN-HSIANG LEE^d

^a Information Technology Center, National Taiwan Normal University,
No.162, Sec. 1, Heping East Road, Taipei City Taiwan

^b Taipei Municipal University of Education, Taipei, Taiwan

^c Biotechnology Industry Study Centre, Taiwan Institute of Economic Research, Taipei, Taiwan

^d Science & Technology Policy Research and Information Center, National Applied Research Laboratories,
Taipei, Taiwan

In scientometrics for trend analysis, parameter choices for observing trends are often made ad hoc in past studies. For examples, different year spans might be used to create the time sequence and different indices were chosen for trend observation. However, the effectiveness of these choices was hardly known, quantitatively and comparatively. This work provides clues to better interpret the results when a certain choice was made. Specifically, by sorting research topics in decreasing order of interest predicted by a trend index and then by evaluating this ordering based on information retrieval measures, we compare a number of trend indices (percentage of increase vs. regression slope), trend formulations (simple trend vs. eigen-trend), and options (various year spans and durations for prediction) in different domains (safety agriculture and information retrieval) with different collection scales (72500 papers vs. 853 papers) to know which one leads to better trend observation. Our results show that the slope of linear regression on the time series performs constantly better than the others. More interestingly, this index is robust under different conditions and is hardly affected even when the collection was split into arbitrary (e.g., only two) periods. Implications of these results are discussed. Our work does not only provide a method to evaluate trend prediction performance for scientometrics, but also provides insights and reflections for past and future trend observation studies.

Introduction

Monitoring research trends has always been a concern of policy makers of science and technology, since it helps resource allocation and technology forecast. Increasingly important research topics are of particular interest to those policy makers. These topics have been also termed as *hot topics*, *upward trends*, or *emerging trends*. Although delicate distinctions among these terms can be made (such as degree of increase and/or lateness of existence), we treat them synonymously in this work. As suggested by their literal meaning, they seem to highly correlate with the strength of attention received over time. However, such topics can not be identified simply by their increasing trend of publications, because genuine hot topics are more than their publication numbers can reveal (as we may see in our experiments later).

Received April 27, 2008

Address for correspondence:

YUEN-HSIEN TSENG

E-mail: samtseng@ntnu.edu.tw

0138–9130/US \$ 20.00

Copyright © 2009 Akadémiai Kiadó, Budapest

All rights reserved

Therefore, domain experts are often consulted for this purpose because they are good at identifying interesting research trends based on their knowledge and experience accumulated over time. However, their observations do not generalize effectively to the fields beyond their expertise. Besides, when a large number of research topics need to be prioritized, inconsistent decision may result from different experts' opinions. Thus automatic mechanism for monitoring research trends, especially increasingly important topics, would be of great help.

In a project, funded by Science & Technology Policy Research and Information Center in Taiwan, to analyze a large set of scientific publications in the agricultural areas, we have the opportunity to detect upward trends for a group of experts and have their feedback in trend type labelling. We then analyze the effectiveness of our detection and prediction methods based on this feedback. To better utilize the valuable expert feedback, we propose a computerized approach to evaluate different methods based on the evaluation measures in information retrieval to know which one is better, quantitatively.

This work reports our methods and findings about this evaluation. In particular, a number of trend prediction indices, including those used in scientometrics and in basic statistics, are examined and compared. In the next section, the trend indices used by several scientometrics studies are introduced. Following is the description of the methods for automatic upward trend detection. The metrics for evaluating their effectiveness are introduced. We then describe the data collections for trend analysis and show how domain experts helped in labelling the trend type for each identified topic. With the experts' feedback, various comparisons are made and the results are shown. Finally, we conclude this paper with the discussion of the implication of our findings.

Previous work

Previous studies on trend watch based on quantitative time series data, such as the document distribution over a period of time for a particular topic, have used various methods and options. Noyons et al. [NOYONS & AL. 1999; NOYONS & VAN RAAN, 1998] often split the publications to be analyzed into two periods and a percentage of increase (or decrease) in the second period relative to the first period was used to suggest the trends of the identified topics.

To suggest rapidly developing research areas, the National Institute of Science & Technology Policy in Japan has used a similar measure [STFC, 2004], except that this percentage of increase was calculated on a per year basis and averaged over the watched years.

Chen in his CiteSpace tool [CHEN, 2006] allows users to vary the time slice for interactive analysis and visualization of emerging trends.

In order to determine what topic areas are appearing in the papers at the SIGIR conferences from the last 25 years, SMEATON & AL. [2003] mapped these trends in a topic-document matrix. They listed the clustered topics in rows and number of documents per year in columns and sorted the rows approximately in order of a combination of the year of the first appearance of the topics, and the number of papers published. With this vision friendly map, they then predicted an ideal paper title, with all the hot topics in it, to appear in the SIGIR conference of the following year.

Recently, CHI & AL. [2006] challenged the validity of the simple accumulation of published documents over time (adopted in all the studies just mentioned) as a temporal series for trend analysis. They argued that different sources (such as journals or countries) may contribute to a research topic differently. To cope with such an uneven contribution, they proposed the use of singular value decomposition (SVD) for ideal trend analysis.

As can be seen from the above survey, various methods were used among researchers without referencing to each others. Their application to different situations makes effectiveness comparison among individual methods even difficult. This motivates us to conduct this research to hopefully suggest a best practice for future scientometrics studies.

Method

Trend identification

Given a collection of publications, our method to identify upward trends involves two steps. First, the knowledge structures of the collection are detected based on a multi-stage clustering approach [TSENG & AL., 2007]. Ideally, terms or documents are grouped into concepts, based on a co-word or co-citation analysis [BRAAM & AL, 1989; CALLON & AL., 1983; RIP & COURTIAL, 1984]. These concepts can be further clustered into topics, which in turn can be clustered into categories or domains. For each identified cluster, a cluster title generation algorithm [TSENG & AL., 2006] is applied to obtain a set of cluster descriptors for helping human analysts in interpreting the results.

Second, the time series of each structure (i.e., hierarchy of concepts, topics, or categories) is created and an index to predict its tendency is calculated. Specifically, the publication dates of the documents are grouped by a specified interval. Number of documents in the same intervals from all sources is counted to yield the frequency sequence over time. This result can be expressed in a time series or a trend as $[d_1, d_2, \dots, d_n]$, where n is the number of intervals, and d_i is the number of documents in the i th interval.

The above simple count does not distinguish the uneven contribution from each source. To take differences among individual sources into consideration, CHI & AL.

[2006] proposed the use of eigen-trends, which are time series derived through singular value decomposition [LATHAUWER & AL., 2000] of the breaking down matrix from the above simple trend. Specifically, if there are m sources (journals or countries), the number of documents contributed by each source in each interval can be expressed in the following matrix D as:

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix}$$

Through singular value decomposition, matrix D can be recast into the multiplication of three matrices $D=USV^T$ or as follows:

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mm} \end{bmatrix} \times \begin{bmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{mn} \end{bmatrix} \times \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{bmatrix}$$

Multiple eigen-trends can then be derived from the above decomposition. However the most important one (contain major and most trend information) is the first eigen-trend computed by the multiplication of the first eigen-value s_{11} in matrix S with the first column in matrix V^T . The authority of each source, i.e., the relative importance of individual source in contribution to the overall trend, can also be derived from the decomposition and is simply the first column of matrix U for the first eigen-trend. In comparison, the simple trend and simple authority are all computed from the matrix D , which are just the column sum and row sum of D , respectively. The following formula sum up the above discussion:

Simple Trend: $[d_1, d_2, \dots, d_n]$, where $d_j = \sum_{i=1}^m d_{ij}$ for each time interval j

Simple Authority: $[a_1, a_2, \dots, a_m]$, where $a_i = \sum_{j=1}^n d_{ij}$ for each source i

(First) Eigen-Trend: $[s_{11}v_{11}, s_{11}v_{21}, \dots, s_{11}v_{n1}]$

(First) Eigen-Authority: $[u_{11}, u_{21}, \dots, u_{m1}]$

Chi et al showed that, by suddenly changing the contribution patterns of some sources at a certain interval, eigen-trends are more sensitive to the contribution changes of those high-authority sources while less affected by noise (or those low-authority sources). Their examples illustrated that in eigen-trends, for a source to have high impact on a topic, a track record is needed to be built over time, and one-time shot does not count very much.

Trend indices

From the above time series (simple trends or eigen-trends), a trend prediction index can be calculated to summarize their tendencies. There are a number of such trend indices used in previous studies. They were examined and compared in the following to know which one is the best predictor.

The first one is the average percentage of increase (*api*):

$$api = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{d_{i+1} - d_i}{d_i}$$

This index has been used in a foresight survey in Japan [STFC, 2004]. It is also the trend indicator used by NOYONS & AL [1998] when $n=2$.

The second index, denoted as *slp*, is the slope of the linear regression line that best fits the data in the time series.¹

$$slp = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2}}, \text{ where } x_i = i - \frac{1}{n} \sum_{i=1}^n i \text{ and } y_i = d_i - \frac{1}{n} \sum_{i=1}^n d_i$$

This is the most commonly used prediction method in statistics [MENDENHALL & SINCICH, 2003]. However, it has a large variation among sequences of different frequency scales, which may make it difficult for humans to predict its trend type based on its value.

Thus the third index is to convert all the numbers in the time series into their z scores (computed as $z_i = (d_i - avg) / stderr$) and calculate the slope of the regression line based on the new sequence. This is denoted as *slp_z*.

The fourth index, denoted as *slp_{pi}*, is a combination of the first and the second index. The original time series is converted into a time series in which each data point is the percentage of increase as defined in the first index. Thus the length of the new series is shorter by one than that of the original series. The slope of the linear regression line that best fits the data in the new series is then calculated as the index. To get positive values for this index, the series should have a tendency that the percentage of increase is greater and greater in each successive interval. Thus it may be ideal for sharp increasing trend detection.

The fifth index is denoted as *slp_c*. It is the slope of the (first) eigen-trend of the original series when breaking down by countries.

The sixth index is denoted as *slp_j*. It is the slope of the (first) eigen-trend of the original series when breaking down by journals.

¹ A module in Perl programming language called Statistics::Regression was used for regression computation.

Trend evaluation

To identify upward trends for experts, the identified structures are sorted by a trend index in decreasing order. If the trend index is a good predictor, the inspection in this order should be efficient in finding upward trends. To know which orderings perform better than the others, the tool called `trec_eval` [BUCKLEY, 2007] is used.² This tool is widely adopted in many large information retrieval evaluation tasks, such as those in TREC,³ NTCIR,⁴ and CLEF.⁵ Two metrics can be read from the output of this tool, i.e., NAP (Non-interpolated Average Precision rate) and Pre@R (Precision rate at Recall position). NAP is defined as:

$$NAP = \frac{1}{R} \sum_{i=1}^R \frac{i}{Rank_i},$$

where R is the number of all relevant items (true upward trends) and $Rank_i$ is the position of the i -th relevant item in the ordering. For Pre@R, it denotes the precision rate at the R -th position in the ordering, i.e., r/R , where r is the number of relevant items in the top R items.

To better understand these two metrics, Table 1 shows examples of the two metrics applied to three different orderings. Assume that A-E and V-Z are ten items to be sorted and A-E (in boldface) are the five items that are relevant (to our interest), while V-Z are not. As shown in Table 1, ordering S1 has all the five relevant items sorted in the top 5 positions, ordering S2 has all of them in the last 5 positions, and ordering S3 evenly distributes these 5 relevant items. Obviously, ordering S1 is the best, S3 the second, and S2 the worst. The values of the two metrics for these three orderings are:

$$\begin{aligned} NAP(S1) &= (1/1+2/2+3/3+4/4+5/5)/5=1.0, & Pre@R(S1) &= 5/5=1.0 \\ NAP(S2) &= (1/6+2/7+3/8+4/9+5/10)/5=0.3547, & Pre@R(S2) &= 0/5=0.0 \\ NAP(S3) &= (1/1+2/3+3/5+4/7+5/9)/5=0.6787, & Pre@R(S3) &= 3/5=0.6 \end{aligned}$$

As can be seen, the above two metrics reflect this precedence perfectly, although in different values. Note that these two metrics require that the relevant items be known in advance. The Pre@R measure is equivalent to counting the number of genuine hot topics above a threshold figure which happens to be the number of all genuine hot topics. The NAP delays the decision of the threshold until the last relevant item is found. Up to that last position, the weighted density of relevant items is then calculated. To sum up, the Pre@R measure is a coarser metric but is simpler to interpret, while the NAP measure is more complicated but can distinguish slightly different orderings.

² A similar tool `trec_eval.perl` rewritten in Perl provided by the host of the NTCIR workshop was actually used.

³ Text REtrieval Conference, <http://trec.nist.gov/>

⁴ NTCIR (NII Test Collection for IR Systems) Project, <http://research.nii.ac.jp/ntcir/>

⁵ CLEF (The Cross-Language Evaluation Forum), <http://www.clef-campaign.org/>

Table 1. Examples of NAP and Pre@R applied to three orderings

Rank	S1	S2	S3
1	A	V	A
2	B	W	V
3	C	X	B
4	D	Y	W
5	E	Z	C
6	V	A	X
7	W	B	D
8	X	C	Y
9	Y	D	E
10	Z	E	Z
NAP	1.00	0.35	0.68
Pre@R	1.00	0.00	0.60

Data

Two data sets were used for trend index evaluation: one from the so-called safety agriculture domain; the other from information retrieval. The experimental setup is described as follows.

Safety agriculture: Collection and topic extraction

A set of six research domains regarding safety agriculture were enumerated by a group of experts from the Science & Technology Policy Research and Information Center, Taiwan. They are: food security, crop protection, livestock, fishery, agroforestry, and environment. For each domain a query was formulated to search the ISI's Web of Science database. By limiting the year range between 1996 and 2005, 72500 records were downloaded, parsed, and saved into a database management system. We identify 13 relevant fields for each downloaded record for our analysis, including AU (author), AB (abstract), TI (paper title), SO (journal title), SC (source category), DE (descriptor), ID (identifier), C1 (main author address, converted into country code), CR (citations), NR (number of citations), TC (times cited), PY (publication year), UT (paper identifier number). Table 2 shows the number of records for each year in this set. As can be seen, the overall trend is obviously increasing.

Table 2. Number of records from 1996 to 2005

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Total
Rec.	5448	6056	6363	6211	6773	7475	8028	7700	9178	9268	72500

To identify topics for trend observation in this set, several approaches were tried. The first is document clustering based on bibliographic coupling. Initial analysis showed that there are 2,765,938 citations in total (among these 72500 papers) and

11,248,898 coupling pairs. Due to this large volume, our clustering program failed to result in any clusters. After removing those pairs having less than 5 common citations, 145,471 document pairs remained and they resulted in 20,795 clusters with zero similarity between any clusters. This set of clusters exhibited highly skewed size distribution, i.e., most clusters (93.85%) are small containing less than 4 documents, only 16 clusters have more than 10 documents, and the largest one have only 16. As such, these clusters were not used in later evaluation.

The next approach is a co-word analysis based on the free texts in the title and abstract fields. Keywords (and key-phrases) were extracted alone with their co-words (commonly co-occurred terms in the same sentences) by Tseng's algorithm (Tseng, 2002). Terms, after stemming, occurs in more than a specified document number were clustered based on common co-words they share. This resulted in 423 clusters or more (depending on the similarity threshold) and their distribution is less skewed than that from bibliographic coupling. However, the quality of these term clusters as topics varies from cluster to cluster. The low precision in cluster quality may waste experts' efforts in providing feedback for trend detection. Although this set of clusters may contain most comprehensive topics in these documents, we decide to leave it for future studies after we obtained reliable techniques in upward trend detection.

Our final approach for topic identification applied the co-word analysis based on controlled terms, i.e., terms from the field of SC and DE. An initial analysis showed that there are 1.8 SC terms for each record on average and only 2.8% of these SC terms occur in their title or abstract, and there are 5.52 DE terms on average and about 46.35% of them can be found in these two free text fields. In total there are 179 SC terms each occurs in more than 10 documents and this number is 3632 for DE terms. In short, these controlled terms exhibit manageable size and different facets from the free texts for analysis.

For their co-word analysis, terms from each field (SC or DE) co-occurred in the same records were counted. This pair-wise count was normalized by the individual occurrence of each term. Similarity based on this calculation was used in a complete link clustering algorithm. From the SC field, a total of 80 clusters (out of 179 SC terms) were found and from the DE field, there were 1617 clusters (out of 3632 DE terms).

It is noted that the SC terms are broader terms, each of which alone represents a topic category of a field. Clustering of them leads to new topics that may be the intersection or union of the topics they contains. For DE terms, they are more like free text terms, defining concepts in a narrower topic. The quality of their clusters may be affected by their coverage. For example, Figure 1 shows the similar concepts defined by two clusters, one from the free text terms (i.e., from the second approach), the other from the DE terms. As can be seen, the former covers more similar terms and is more complete for that topic. Nevertheless, compared to the free text terms, DE term clusters lead to more valid topics due to less noise.

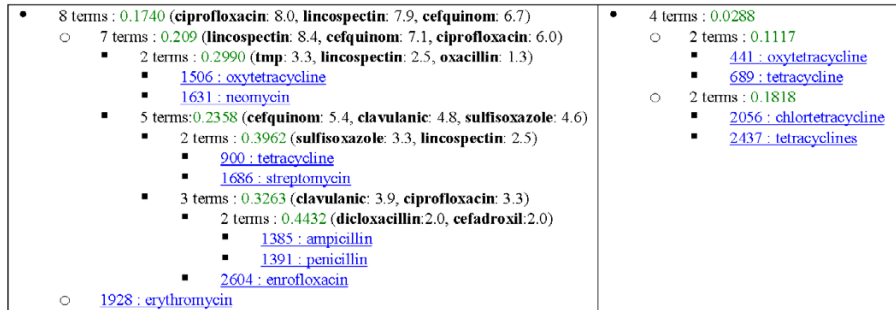


Figure 1. Result examples from co-word analyses. Left figure is based on title and abstract, where common co-words are shown in the parentheses. Right figure is from the DE field.

Safety agriculture: Trend type labelling by experts

To reduce the size of the above clusters to a manageable amount for manual inspection, we sampled 50% of clusters from SC and 10% from DE and removed those clusters or sub-clusters having less than 30 documents. The resultant clusters with their time series were then presented to nine experts for trend type labelling so as to create a benchmark for evaluating the proposed trend indices. These judges, including six professors, two researchers, and one administrative manager invited by the Science & Technology Policy Research and Information Center, are all domain experts having a concern of future agriculture development.

When labelling the trend type, the experts were advised to judge the type of each cluster based on their knowledge about the terms in that cluster. If this did not help, the time series of the cluster can be consulted. If this did not help either, the documents in the cluster can be examined. If all these efforts failed, the cluster was labelled *inconclusive*, denoted as ‘?’. Five types of trend can be labelled if it is decidable, i.e., *sharp increase*, *increase*, *fluctuation*, *decrease*, and *sharp decrease*, denoted as ‘++’, ‘+’, ‘=’, ‘-’, and ‘--’, respectively.

Table 3 shows examples of the trend type labelling made by experts against some DE clusters. In this table, the first column is the cluster ID, the second is the number of terms in the cluster or sub-cluster, the third is the minimum similarity between any terms in the same group, and the rest of columns are self-explained by their column titles.

Table 4 shows the statistics of the experts’ judgment. As can be seen, 43 term groups (43/155=27.74%) from SC were inconclusive in their trend type. This percentage for DE is 22.89%. Since the overall tendency is upward, there is no sharp decreasing case for SC and DE. The decreasing cases are also rare.

Table 3. Examples of experts' feedback on trend type labelling

cid	nt	Sim	Trend	DE Terms	df	96	97	98	99	00	01	02	03	04	05
9	6	0.025	=	osmoregulation; chloride cell; metamorphosis; thyroid hormone; flounder; flatfish	147	3	9	17	14	16	26	17	12	14	19
9	4	0.066	=	osmoregulation; chloride cell; metamorphosis; thyroid hormone	106	2	4	16	11	13	16	11	10	9	14
9	2	0.178	-	osmoregulation; chloride cell	74	1	2	13	8	10	14	7	4	7	8
9	2	0.192	+	metamorphosis; thyroid hormone	36	1	2	3	3	4	3	4	6	2	8
9	2	0.120	?	flounder; flatfish	46	1	5	1	4	5	11	6	2	5	6
28	5	0.032	++	food allergy; anaphylaxis; IgE; gelatin; allergen	102	4	6	4	5	10	15	15	14	15	14
28	2	0.138	++	food allergy; anaphylaxis	67	3	5	4	4	5	12	10	8	9	7
28	3	0.164	++	IgE; gelatin; allergen	46	1	2	1	2	6	5	6	8	7	8
28	2	0.196	+	IgE; gelatin	34	1	2	1	1	6	4	3	8	4	4

Table 4. Statistics of experts' judgment

Field	(sub-)clusters	++	+	=	-	--	?
SC	155	18	57	37	0	0	43
DE	249	20	61	97	14	0	57

Information retrieval: Collection, clusters, and hot topics

For the second data set, we use the collection prepared by Smeaton et al. They collected the titles, author names, and abstracts of all the 853 papers published from the first ACM SIGIR conference to the 25th. Using a commercial software package called Clustan Graphics, 29 non-overlapping clusters were generated for the document set. They then inspected each cluster manually and assigned a topic description to reflect the theme of the majority of the papers in each cluster. To add some structure to this clustering, and see how topics were spread over the 25 SIGIR conferences, they mapped the documents in each cluster to the year of the SIGIR in which they appeared, as shown in Table 5. In the table, the rows, representing clusters or topics, are sorted approximately in order of a combination of the year of their first appearance, and the number of papers published. The ID column denoting the sorting order was inserted by us for the convenience of later discussion.

Apart from attempting to track the evolution of topic areas in the information retrieval field, they extrapolated and predicted the “hottest” topics for the following year. The ideal paper title expected by Smeaton et al to appear in SIGIR 2003 is “**Evaluation of a Language Model Implementation of a Topic-Based, Cross-Lingual Question-Answering and Summarisation System**”, an aggressive title containing all the hot topics at that time.

Table 5. Clustering and ordering of SIGIR papers by topics made by SMEATON & AL.

Cluster \ Year	ID	71	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	Total
Databases, NL Interfaces	29	8	4	1	6	5	10	1	3	5	2	5	2	4	1	3	1	1	2	2								66
General	28	5	2	9	2	9	5	7	10	10	6	10	6	2	5	8	6	2	2	4	3	1		4	2	5	1	126
Models	27	1			2	1	1		4	1	2	1	2	1	2		2	2	2	2	3	1						30
Question answering	26	1			1	1	1						1				1	1		1				4	4	1	17	
Syntactic phrases & SDR	25	1				1	1		2	1	6	3	3	2	3	2	1	1	2	1	1	3	1	1	1	1	1	37
Conceptual IR, KB IR	24	1			4	4	1	3	3	4	3	5	7	5	1	6	3	5	3	2	3	4	1	3	2	1	1	75
Compression	23	1							1	2	2	1	1	1	3	1	1	1	1				2			1	18	
Clustering	22		2		1	1		2		3	3	2				1	2		1	1	2	1		1		3	26	
Relevance feedback	21		1	1	1		2		1	1		1			1	2	4	3		1	2	1	1	1	1	1	25	
Inverted files & Implementations	20		1				1		1				2	1	3	1			2	1				1		1	3	18
Term weighting	19			1	3	2	1	2	1	1	5	3	3			1		2	1	1	1	1				1	1	31
Message understanding & TDT	18			1	1						1					3	2			3	4	2	4	5	5	5	31	
Filtering	17			1					1			1			1			1		4	1	1	1	1	2	3	18	
Hypertext IR, Multiple evidence	16										1	3	1	1	2	1	2	2	2	1	4	3	1	5	2	2	2	33
Image retrieval	15				1			1			1			1	1				2	1	1						9	
Probabilistic & Language models	14				1	1	1					3	1		3	4	2	2	3	2	1	3	1		3	3	34	
Boolean & extended Boolean	13						1		2	1				1			1	1			1	1	1				10	
Japanese & Chinese IR	12							1						1			2		3	2	3	1	1				14	
DBMS & IR	11				1	1	1											1	1								5	
Users & Search	10					2	3	3	2	2	4		3	2	2	3	1	3	3	1	1	1	2	1			38	
Visualisation	9									1	1	1	1		1		1		2	1	1	2			1		12	
Signature files	8							1	1	1	2	2		1	1												9	
Distributed IR	7					1	2	1	2		2	1	1					3	1	1	3	4	2	1	1	1	24	
Evaluation	6																3	4	4	2	1	7		2	3	8	34	
Topic distillation & Linkage retrieval	5																					1		3	3	2	9	
Latent semantic indexing	4											1			1		1						2	1			6	
Text categorisation	3												1					3	3	3	1	3	1	3	3	2	23	
Document summarisation	2																		2				2	2	3	3	12	
Cross lingual	1																			1	3	3	1	1	3	4	16	

Based on their expectation, we listed three sets of relevant topics for evaluation in Table 6. The reason for enumerating three judgment sets is due to our uncertainty about the word “topic-based” in their ideal paper title. We excluded this term in judgment set J5 and gradually added back its relevant topics (relevant to the word “topic-based”) in J8 and J10.

To know exactly what topics appeared in the SIGIR 2003 conference, we analyzed the table of contents in that year’s proceedings. The session titles are listed in the first column of Table 7, with some topical terms added by ourselves in parentheses to better reveal their contents. The second column lists the number of papers in that year.

The third and fourth column lists the corresponding topic IDs in Table 5. Again, these two sets of correspondence represent two possible judges of the relevant topics in Table 5 based on the session titles in Table 7.

Table 6. Three set of hot topics for SIGIR 2003 based on the of ideal paper title expected by SMEATON & AL.

J5	J8	J10
Question answering Probabilistic & Language models Evaluation Document summarisation Cross lingual	Question answering Probabilistic & Language models Evaluation Document summarisation Cross lingual Clustering Filtering Text categorisation	Question answering Probabilistic & Language models Evaluation Document summarisation Cross lingual Clustering Filtering Text categorisation Message understanding & TDT Topic distillation & Linkage retrieval

Table 7. Fourteen session titles (topics) in the SIGIR 2003 conference

Topics	df	S13	S10
Retrieval Models (Language Models, Evaluation)	3	14	14
Question Answering (Evaluation)	3	26	26
Web (Hyperlink, Classification)	3	5	
Human Interaction	6	10	
Text Categorization	6	3	3
Multimedia Information Retrieval	3	15	15
Structured Documents (XML)	2		
Text Representation (Term Modelling)	2	19	
IR Theory (LSA)	3	4	4
Filtering and Retrieval Models (LSA)	3	17	17
Clustering	3	22	22
Distributed Information Retrieval (Source Selection, Topic Segmentation)	3	7	7
Novelty and Topic Change (Text Segmentation)	3	18	18
Cross-Lingual Information	3	1	1

Results

Safety agriculture

It is our concern to distinguish the sharp increasing or increasing cases from the others. The performance in this regard for each of the above trend index is shown in Table 8. Based on the NAP and Pre@R measures, slp and slp_z perform best, the two eigen-trends are the second, and the percentage-of-increase type indices are the worst.

The fact that the linear regression indices are the best verifies their widely usage for trend prediction in statistics. The fact that the eigen-trends did not perform significantly well suggests that the effect of authorities or uneven contribution of individual sources does not prevail in this collection. In fact, when we examined the data, we found that

the simple authority vectors are almost the same as the eigen-authority vectors and the slopes of the simple trends coincide with the slopes of the eigen-trends. As such, no advantage was gained from the singular value decomposition of the break-down trend matrix. The fact that the percentage-of-increase type indices perform worst, when the year span is 1, is somewhat surprising to know, since these indices are intuitively simple to use and interpret. Especially, that the index slp_{pi} performs poorly for sharp increasing trends suggests the complication of trend judgement.

Table 8. Performance of different trend indices.
The Avg rows are the averages of the values in the SC and DE rows

Yearly Index	Judgment Field	+ or ++		++	
		NAP	Pre@R	NAP	Pre@R
<i>api</i>	SC	0.6093	0.6533	0.1733	0.1111
	DE	0.4454	0.5062	0.1587	0.1500
	Avg	0.5273	0.5798	0.1660	0.1306
<i>slp</i>	SC	0.9521	0.8800	0.4552	0.3333
	DE	0.8254	0.7407	0.4293	0.5500
	Avg	0.8887	0.8104	0.4423	0.4417
<i>slp_z</i>	SC	0.9524	0.8933	0.2992	0.2778
	DE	0.8424	0.7654	0.5689	0.5500
	Avg	0.8974	0.8294	0.4340	0.4139
<i>slp_{pi}</i>	SC	0.7250	0.7733	0.2051	0.1667
	DE	0.3807	0.3704	0.0867	0.0000
	Avg	0.5528	0.5719	0.1459	0.0833
<i>slp_e</i>	SC	0.9295	0.8533	0.4457	0.2222
	DE	0.7846	0.6914	0.4574	0.4500
	Avg	0.8570	0.7723	0.4516	0.3361
<i>slp_j</i>	SC	0.9214	0.8267	0.4722	0.4444
	DE	0.8041	0.7284	0.4084	0.3500
	Avg	0.8627	0.7775	0.4403	0.3972

Previous studies have used different intervals or year spans to create the frequency sequence for trend observation. This was evaluated in Figure 2 for index *api*, *slp*, and *slp_z* where the year span varies from 1, 2, to 5. A year span of 5 corresponds to dividing the 10-years collection into two periods. It is again surprising to know that even with only two periods, the performance of the *slp* index remains virtually the same. Also interesting is that the performance of *api* increases as the year span increases. In contrast, the *slp_z* index drops in performance drastically. This is because it yields only three values +2, 0, and -2 when there are only two periods, which can be verified by calculating its value from its definition.⁶ The low variation on its values makes it difficult to sort the trends in a meaningful way.

⁶ Let the sequence be (x_1, x_2) . Its Z-sequence would be $((x_1 - \text{avg}) / \text{stderr}, (x_2 - \text{avg}) / \text{stderr}) = ((x_1 - x_2) / 2 / |(x_1 - x_2) / 2|, (-x_1 + x_2) / 2 / |(x_1 - x_2) / 2|)$. Thus only 3 values result from the Z-sequence: $(-1, 1), (1, -1), (0, 0)$, which in turn yield only 3 possible slopes: +2, -2, and 0.

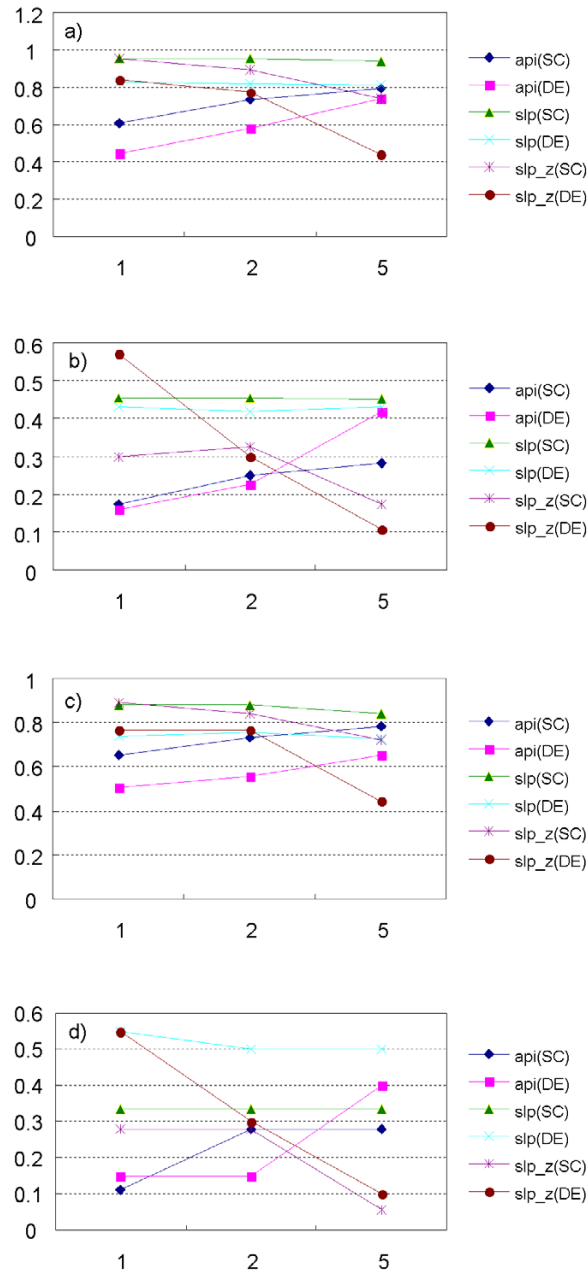


Figure 2. Prediction effectiveness when year span varies from 1, 2, to 5.
 a) NAP for + or ++; b) NAP for ++; c) Pre@R for + or ++; d) Pre@R for ++

The next evaluation is to know how performance would be affected if we only use the first n years of data for prediction. As shown in Figure 3, for the slp index the NAP performance may drop as low as 25% for the sharp increase cases when only the first 8 years of data were used. In other words, when we are at the 8th year to predict the trends in the 10th year, the performance is 75% of that when we have the full 10 years of data.

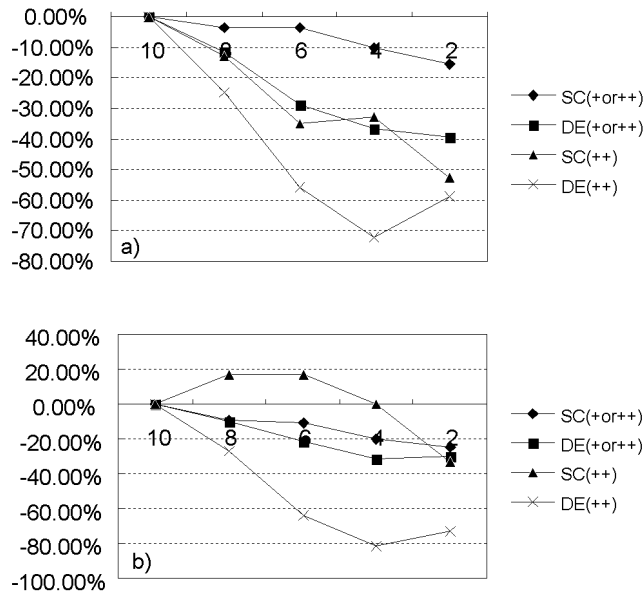


Figure 3. Percentage of performance drop for slp using only the first n years of data, where $n=10, 8, 6, 4,$ and 2 . a) NAP; b) Pre@R

Information retrieval: Collection, clusters, and hot topics

Our final evaluation compared the topic orderings sorted by the above indices with the ordering sorted by Smeaton et al as shown in Table 5. The performance for this set of data for each applicable index is shown in the first six columns in Table 9.

As shown in this table, the index slp_{pi} performs unexpectedly well. This may be due to the tendency of successive increase in trend for relevant topics in this data set. The first ten topics sorted by this index are shown in Table 10, which match 9 out of 10 topics of J10 in Table 6. The second is the normal slope index slp . The ordering of Smeaton et al is only good enough for judgment set J5, but worse than the other two best indices for the other sets.

Compared to the exactly topics appeared in the SIGIR 2003 conference, the performance for different orderings is shown in the last four columns in Table 9. Although the best performing ordering slightly changes for these sets,⁷ the simple slope slp remains effective compared to the others. The ordering of Smeaton et al is not as good as expected.

Table 9. Performance of different trend indices on the data of Smeaton et al and on the actual session topics in SIGIR 2003

Judgment	J5		J8		J10		S13		S10	
Index	NAP	Pre@R	NAP	Pre@R	NAP	Pre@R	NAP	Pre@R	NAP	Pre@R
api	0.2527	0.4000	0.4824	0.3750	0.5724	0.5000	0.5432	0.5714	0.3691	0.4286
slp	0.5852	0.4000	0.6584	0.7500	0.8234	0.8000	0.5404	0.5714	0.3994	0.5000
slp_z	0.3958	0.4000	0.5597	0.6250	0.7340	0.8000	0.5582	0.6429	0.4023	0.5714
slp_{pi}	0.8083	0.6000	0.7698	0.7500	0.9625	0.9000	0.5727	0.5714	0.3808	0.5000
Smeaton	0.5956	0.4000	0.6253	0.5000	0.6712	0.5000	0.5910	0.5000	0.3854	0.3571

Table 10. Top 10 topics predicted by slp_{pi} based on the data in Table 5

Rank	Topic	slp_{pi}	Rank	Topic	slp_{pi}
1	Evaluation	0.0462	6	Cross lingual	0.0173
2	Question answering	0.0287	7	Filtering	0.0156
3	Topic distillation & Linkage retrieval	0.0235	8	Probabilistic & Language models	0.0151
4	Document summarisation	0.0213	9	Text categorisation	0.0138
5	Message understanding & TDT	0.0213	10	Compression	0.0113

Implications and conclusions

We have learned from the scientometrics literature that different year spans may be used to create the time sequence and different indices may be chosen for trend observation. However, the effectiveness of these parameters was hardly known, quantitatively and comparatively. Based on sorting research topics in decreasing order of interest predicted by a trend index and then evaluating this ordering by way of information retrieval measures, we have compared a number of metrics (percentage of increase vs. regression slope), trend formulations (simple trend vs. eigen-trend), and options (various year spans and durations for prediction) in different domains (safety agriculture and information retrieval) with different collection scales (72500 papers vs. 853 papers) for different topic sources (SC vs. DE terms). Our work provides clues to better interpret the results when a certain choice was made. For example, when api was chosen for trend prediction, it should be noted that it is only effective for large year span or short time series. Also when the collection is split into only two periods, both

⁷ Note that because the relevant topics become more (e.g., in the Pre@R metric, R=14 in Table 7, which is larger than R=5, 8, or 10 in Table 6), the performance values become smaller even though the ordering is the same.

slp and *api* can be used. Thus the work of Noyons et al. that often used *api* to analyze two periods of publications remained effective. In any cases, *slp* is recommended since it performs consistently well under the various conditions that we evaluated.

The validity of the above insights is based on the validity of the relevance judgments that were used in our evaluations. The fact that the judges were allowed to use the trends in the data to make their judgement does not jeopardize their validity. As the case of the SIGIR data (Table 5) shows, even when the data were viewed visually, it does not prevent their authors to select a low ranking topic (e.g., “Question Answering” at the 26th position) as a genuine hot topic and vice versa (“Distributed IR” at the 7th position was not regarded as a hot topic). Furthermore, despite the different relevance judgments exist as shown in Table 9, high performing index still performs high in general. This phenomenon is often observed in other similar tasks involving human judgment (such as those in the work of TSENG & TEAHAN [2004]). Thus the inevitable discrepancy in judgments among individual experts can be safely ignored.

Our goal is to explore the best way to predict upward trends in an environment where a large number of topics are to be monitored. Our work is important to know which index is the best under a certain condition. If a good trend index is used, the inspection in the order sorted by the index should be efficient, which could relieve the burden of analysts on monitoring any upward trends from a large stream of scientific publications.

Despite the various conditions been experimented in this work, more cases using data from other domains are worth of study. The information retrieval based method for evaluating the trend index performance suggests a relatively objective and repeatable procedure to indentify better indices and to gather evidence to support (or invalidate) our current results. For example, whether high impact (authority) journals should be given more weight in trend analysis can be further verified with data from more domains, although our experiment in safety agriculture does not confirm this viewpoint.

In this work, we only examine topic trends based on the document frequencies over time. More internal or external knowledge structures have not been used. For example, sophisticated evolutionary models studied in past scientometrics literature [BRUCKNER, & AL., 1990] or citation information among identified topics [NOYONS & AL., 1999] are of this kind. Application of these models to predict future trends may result in better performance. Our future work may explore this direction.

*

This work is supported in part by NSC 96-2221-E-003-017-, NSC 96-2524-S-003-001- and NSC 97-2631-S-003-003-.

References

- BRAAM, R. R., MOED, H. F., VAN RAAN, A. F. J. (1989), *Comparison and Combination of Co-citation and Co-word Clustering*. Leiden: DSWO Press, University of Leiden.
- BRUCKNER, E., EBELING, W., SCHARNHORST, A. (1990), The application of evolution models in scientometrics. *Scientometrics* 18 (1–2) : 21–41.
- BUCKLEY, C. *trec_eval IR evaluation package*. http://trec.nist.gov/trec_eval/, accessed on 2007/02/10.
- CALLON, M., COURTIAL, J. P., TURNER, W. A., BAUIN, S. (1983), From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22 : 191–235.
- CHEN, C. (2006), CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57 (3) : 359–377.
- CHI, Y., TSENG, B. L., TATEMURA, J. (2006), Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. Paper presented at the *Proceedings of the 15th ACM international Conference on Information and Knowledge Management (CIKM)*.
- LATHAUWER, L. D., MOOR, B. D., VANDEWALLE, J. (2000), A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21 (4).
- MENDENHALL, W., SINCICH, T. L. (2003), *A Second Course in Statistics: Regression Analysis* (Sixth ed.): Prentice-Hall.
- NOYONS, E. C. M., MOED, H. F., VAN RAAN, A. F. J. (1999), Integrating research performance analysis and science mapping. *Scientometrics*, 46(3) : 591–604.
- NOYONS, E. C. M., VAN RAAN, A. F. J. (1998), *Mapping Scientometrics, Informetrics, and Bibliometrics*. Retrieved November 23, 2006, from <http://www.cwts.nl/ed/sib/home.html>
- NOYONS, E. C. M., VAN RAAN, A. F. J. (1998), Monitoring science developments from dynamic perspective: self-organized structuring to map neural network research. *Journal of the American Society for Information Science and Technology*, 49 (1) : 68–81.
- RIP, A. I., COURTIAL, J. (1984), Co-word maps of biotechnology: an example of cognitive scientometrics. *Scientometrics*, 6 : 381–400.
- SMEATON, A. F., KEOGH, G., GURRIN, C., MCDONALD, K., SODFING, T. (2003), Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century? *ACM SIGIR Forum*, 37 (1) : 49–53.
- STFC (2004), *The 8th Science and Technology Foresight Survey – Study on Rapidly-Developing Research Areas – Interim Report*. National Institute of Science & Technology Policy, Japan.
- TSENG, Y. H. (2002), Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology*, 53 (13) : 9.
- TSENG, Y. H., LIN, C. J., CHEN, H. H., LIN, Y. I. (2006), Toward generic title generation for clustered documents. Paper presented at the *Proceedings of Asia Information Retrieval Symposium, Singapore*, 2006, Oct. 16–18.
- TSENG, Y. H., LIN, C. J., LIN, Y. I. (2007), Text mining techniques for patent analysis. *Information Processing and Management*, 43 (5) : 1216–1247.
- TSENG, Y. H., TEAHAN, W. J. (2004, July 25–29), Verifying a Chinese collection for text categorization. Paper presented at the *27th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '04*, Sheffield, U.K.