

Which Trend Metrics Predict Emerging Trends Better?

Yuen-Hsien Tseng*

National Taiwan Normal University, No.162, Sec. 1, Heping East Road., Taipei, Taiwan.
E-mail: samtseng@ntnu.edu.tw

Yu-I Lin

Taipei Municipal Univ. of Education, Taiwan, 100. E-mail jg141@mail.tpc.edu.tw

Yi-Yang Lee, Wen-Chi Hung, Chun-Hsiang Lee

Science & Technology Policy Research and Information Center, Taiwan, 106.
E-mail {yylee, wchung, chlee}@mail.stpi.org.tw

* Corresponding author

Theme: Theme 2 S&T indicators for the identification of emerging fields;

Keywords: Clustering, Trend Index, Eigen-Trend, Linear Regression, Evaluation.

1 Background

In scientometrics for trend analysis, different year spans may be used to create the time sequence and different indices were chosen for trend observation. However, most encountered the lack of expert feedback to the results [Noyons & Raan, 1998]. In addition, effectiveness of such prediction is rarely known, quantitatively.

2 Problem/Application

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. If a good trend index is used, the inspection in the order sorted by the index should be efficient. In this work, we compare some of these trend prediction indices and options to know which one performs best.

3 Methodology

For each topic detected in a collection, the time series (simple trend) is created and an index to predict its tendency is calculated. To take authority differences among individual sources into consideration, [Chi, Tseng, & Tatemura, 2006] proposed the use of eigen-trends. Specifically, if there are m sources (journals or countries) and n intervals, the number of documents contributed by each source in each interval can be denoted as D , which can be recast into $D=USV^T$ or as follows:

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

The following formula clarifies the concepts of various trends:

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

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

Three kinds of indices are then compared. (1) 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 (STFC, 2004) and in (Noyons & van Raan, 1998) when $n=2$.

(2) The slope of the linear regression line (denoted as *slp*):

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

Another variation is to convert all the numbers in the sequence into their z scores and calculate its *slp*. This is denoted as *slp_z*.

(3) The third kind includes the slope of the eigen-trend of the original series when breaking down by countries, denoted as *slp_c*, and the similar one when breaking down by journals, denoted as *slp_j*.

The identified structures are sorted by a trend index in decreasing order and evaluated by the two metrics: (i) NAP 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, and (ii) Pre@ R (Precision rate at Recall position) which 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.

4 Data:

Two data sets were used for trend index evaluation. A set of 72500 records regarding safety agriculture ranging from 1996 to 2005 were downloaded from the ISI's Web of Science database. Co-word analyses based on the fields of SC (Source Category) and DE (Descriptor) were used for topic detection. The resultant clusters were then presented to 9 experts for trend type labelling as a benchmark for evaluating the proposed trend indices. Table 1 shows the judgment result, where '++', '+', '=', '-', '--', and '?' denote *sharp increase*, *increase*, *fluctuation*, *decrease*, *sharp decrease*, and *inconclusive*, respectively.

For the second data set, we use the data by (Smeaton, et al, 2003) for the first 25 years of 853 SIGIR papers. Table 2 lists the clusters and ordering shown in their paper, with an ID column inserted to rank the ordering. From this table, 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**". Based on their expectation, we listed three possible sets of relevant topics in Table 3 for evaluation.

5 Outcome/Findings/Results

Table 4 shows that *slp* performs best, the two eigen-trends are the second, and *api* the worst when the year span is 1, which is somewhat surprising to know. Figure 1 shows the performance 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.

Table 5 shows that the best predictor is again the *slp*. The ordering of Smeaton et al is only good enough for judgment set J5, but worse than the best index for the other judgement sets.

6 Conclusions

Our work provides reflection for those mentioned in the above. Another contribution is that the proposed method suggests a rigorous procedure to evaluate and 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.

Table 1. Statistics of experts' judgment.

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

Table 2. Clustering and ordering of SIGIR papers by topics made by Smeaton et 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	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	1	3	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		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	2	1	4	3	1	5	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	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	2	3	1	3	3	1		1	1	2	1				38
Visualisation	9									1	1	1		1			1		2	1	1	2			1			12
Signature files	8						1	1	1	1	2	2		1	1													9
Distributed IR	7					1	2	1	2	1	2		1		1				3	1	1	3	4	2	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

Table 3. Three set of hot topics for SIGIR 2003 based on the of ideal paper title expected by Smeaton et al.

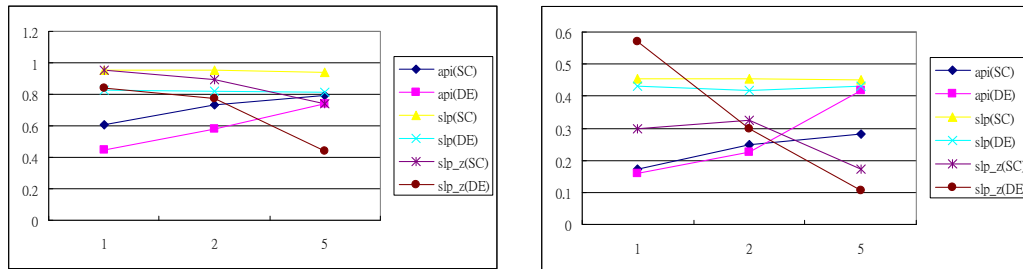
J5	J8	J10
26: Question answering 14: Probabilistic & Language models 6: Evaluation 2: Document summarisation 1: Cross lingual	26: Question answering 14: Probabilistic & Language models 6: Evaluation 2: Document summarisation 1: Cross lingual 22: Clustering 17: Filtering 3: Text categorisation	26: Question answering 14: Probabilistic & Language models 6 :Evaluation 2: Document summarisation 1: Cross lingual 22: Clustering 17: Filtering 3: Text categorisation 18: Message understanding & TDT 5: Topic distillation & Linkage retrieval

Table 4. Performance of different trend indices.

The Avg rows are the averages of the values in the SC and DE rows.

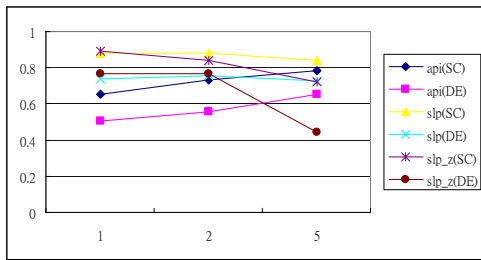
Yearly Index	Judgment Field	+ or ++		++	
		NAP	Pre@R	NAP	Pre@R
<i>api</i>	SC	0.6093	0.6533	0.1733	0.1111
	DE	0.4454	0.5062	0.1587	0.1500
	Avg	0.5273	0.5798	0.1660	0.1306
<i>slp</i>	SC	0.9521	0.8800	0.4552	0.3333
	DE	0.8254	0.7407	0.4293	0.5500
	Avg	0.8887	0.8104	0.4423	0.4417
<i>slp_z</i>	SC	0.9524	0.8933	0.2992	0.2778
	DE	0.8424	0.7654	0.5689	0.5500
	Avg	0.8974	0.8294	0.4340	0.4139
<i>slp_c</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

Figure 1. Prediction effectiveness when year span varies from 1, 2, to 5.

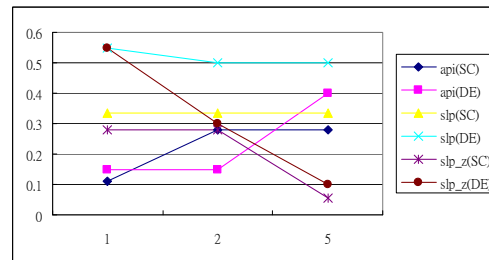


(a) NAP for + or ++.

(a) NAP for ++.



(c) Pre@R for + or ++.



(d) Pre@R for ++.

Table 5. Performance of different trend indices on the data of Smeaton et al in SIGIR 2003.

Judgment	J5		J8		J10	
Index	NAP	Pre@R	NAP	NAP	Pre@R	NAP
<i>api</i>	0.2527	0.4000	0.4824	0.3750	0.5724	0.5000
<i>slp</i>	0.5852	0.4000	0.6584	0.7500	0.8234	0.8000
<i>slp_z</i>	0.3958	0.4000	0.5597	0.6250	0.7340	0.8000
Smeaton	0.5956	0.4000	0.6253	0.5000	0.6712	0.5000

References

- 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).
- 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.
- Smeaton, A. F., Keogh, G., Gurrin, C., McDonald, K., & Soding, 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, Japano. Document Number)