Using Social Annotations for Trend Discovery in Scientific Publications

Meiqun Hu
meiqun.hu@gmail.com

Ee-Peng Lim
eplim@smu.edu.sg

Jing Jiang
jingjiang@smu.edu.sg

School of Information Systems
Singapore Management University
80 Stamford Road, Singapore 178902

ABSTRACT

Social tags and citing documents are two forms of social annotations to scientific publications. These social annotations provide useful contextual and temporal information for the annotated work, which encapsulates the attention and interest of the annotators. In this work, we explore the use of social annotations for discovering trends in scientific publications. We propose a trend discovery process that employs *trend estimation* and *trend selection and ranking* for analyzing the emerging trends shown in the social annotation profiles. The proposed sigmoid trend estimator allows us to characterize and compare *how much*, *when* and *how fast* the trends emerge. To perform topic-specific trend analysis, we further adopt *topic modeling* on the annotation content to decapsulate the multitude of impact created by the annotated work.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

social annotations, temporal profiles, emerging trends

1. INTRODUCTION

Social annotations are auxiliary information users create for resources on the Web. Specifically for the scientific literature, both social tags and citing documents are social annotations to the published work. When there is an increasing attention given to a topic or an individual work, it shows up in these social annotations. In this work, we propose the task of trend discovery using social annotations, focusing on scientific publications.

Discovering and analyzing trends using social annotations for scientific publications has several useful applications. In library science and information studies, profiling the publications to support better search and reference is an important task. While the content of a publication becomes immutable once it is published, the impact it has on subsequent work can be observed over some period of time. Such impact can be shown in its social annotations, since these annotations provide temporal and topical relevance from the perspectives of the annotators. For information seekers, especially junior researchers who often conduct survey on unfamiliar research areas, selecting interesting publications among a large collection is a challenging task. Given a publication, one may want to ask: *How much interest did people have on this work*? *When did such interest emerge*? *How fast was the emergence*? One may further pinpoint a particular topic of research and ask: *When did the interest on this work emerge from wireless networks research*?

Traditional approach to determining the impact of the published work mainly relies on citation indexes, known as *bibliometrics*. However, most citation indexes provide only a snapshot view of the citation database, and they do not use the annotation content. In this work, we make use of the temporal information in social annotations to construct *social annotation profiles* for the annotated work. Based on each social annotation profile, we derive the corresponding time series, on which *trend estimation* can be performed to discover *emerging trends*. Furthermore, we analyze the annotation content through topic modeling to decapsulate the multitude of impact shown in the social annotation profiles.

In this research, we seek to answer the following questions:

1. How to find emerging trends from social annotations?
2. How to use emerging trends to answer questions that are useful to researchers and information seekers?

2. A TREND DISCOVERY PROCESS

An overview of our proposed trend discovery process is depicted in Figure 1. In order to perform trend analysis tasks that address publication-specific and topic-specific questions, we decompose the trend discovery process into three main modules, namely *topic modeling*, *trend estimation* and *trend selection and ranking*.

The *topic modeling* module performs content analysis on the social annotations. Social annotations for the same annotated work may come from different topics of interest. By analyzing the annotation content, we are able to decapsulate the multitude of interest. This allows us to perform trend analysis tasks that address topic-specific questions, such as *How much interest does the wireless networks research community have on the annotated work*?

The *trend estimation* module finds and parameterizes the emerging trends shown in the social annotations. To perform trend estimation, we first construct temporal profiles

Copyright is held by the author/owner(s).
HCIR '11, October 20, 2011, Mountain View, California, USA

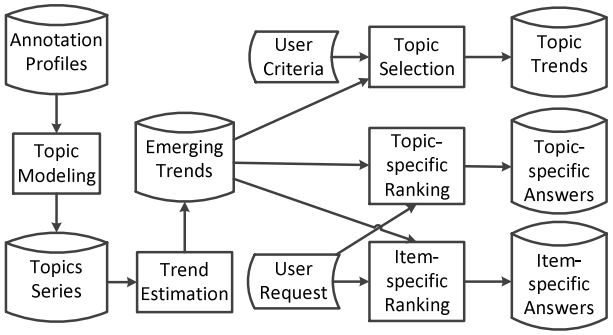


Figure 1: An Overview of Trend Discovery Process

using the social annotations, and then derive time series corresponding to the temporal profiles. Given each time series, we perform function fitting to estimate the trend. The trend estimator should allow us to capture characteristics such as *how much*, *when* and *how fast* the trend emerges.

The *trend selection and ranking* module identifies interesting and significant emerging trends using the estimated trend parameters. To demonstrate the usefulness of the emerging trends found, we perform various topic-specific and publication-specific trend analysis tasks.

In what follows, we focus on discussing the *trend estimation* and *trend selection and ranking* modules. We leave out the details about topic modeling in this paper. Interested readers may refer to [1, 2].

2.1 Constructing Social Annotation Profiles

A *social annotation profile* consists of a stream of *annotation documents*. We consider two types of social annotation documents for scientific publications, where each type is based on the contributions from the corresponding social annotation community. From the social tagging community, each annotation document corresponds to one bookmark, which contains a set of tags assigned to the annotated work and a timestamp. From the scientific research community, each annotation document corresponds to one citing document, which contains content words and a timestamp, *i.e.* the publication year. By aligning a collection of annotation documents with their corresponding timestamps, we construct a stream of annotation documents, which we call the *social annotation profile*.

We now define some terms and notations for formally representing publications and their social annotation profiles. We use the term *item*, denoted as i, to refer to a publication being annotated. We use the term *topic*, denoted as k, to refer to a research community specializing in an area of interest, *i.e.* latent topic. We use the symbol $\mathbb{D}$ to denote a social annotation profile. In this study, we focus on the following three types of social annotation profiles.

- *Item-wise document profile*, denoted as $\mathbb{D}_i$, consists of the stream of annotation documents that are used to annotate item i.
- *Topic-wise document profile*, denoted as $\mathbb{D}^k$, consists of the stream of annotation documents that are associated with topic k.
- *Item-wise topic profile*, denoted as $\mathbb{D}_i^k$, consists of the stream annotation documents that are associated with topic k and are used to annotate item i.

Our definition for topics follows Blei *et al.* [1]. Given a corpus consisting of a set of annotation documents, we assume that there are K topics in the corpus, *i.e.* $k \in [1, K]$. We learn the association of each document with topics by performing topic modeling on the social annotation corpus.

For each social annotation profile $\mathbb{D}$, we construct the corresponding time series $\mathbb{Q} = \{(t, q_t)\}$, where t denotes a time window and q_t denotes the number of annotation documents during time window t in the social annotation profile $\mathbb{D}$. We use calendar months and publication years as time windows for social tags and citing documents respectively. Note that, we have $\mathbb{Q}_i$ for $\mathbb{D}_i$, $\mathbb{Q}^k$ for $\mathbb{D}^k$, and $\mathbb{Q}_i^k$ for $\mathbb{D}_i^k$. Without causing any confusion, we omit their superscripts and subscripts in the following discussion.

To define $\mathbb{D}$ and $\mathbb{Q}$, we use d to denote an annotation document, which consists of its *annotation content* (denoted by $\vec{w}_d$) and a *timestamp* (denoted by s_d), and s_t to denote the starting timestamp of the time window t. Formally,

$$\mathbb{D} = \left\{d_n : n \in \mathbb{N}, s_{d_n} \leq s_{d_{n+1}}\right\}$$
$$\mathbb{Q} = \left\{(t, q_t) : 1 \leq t \leq T, q_t = \textstyle\sum_{d \in \mathbb{D}} I\left(s_t \leq s_d < s_{t+1}\right)\right\}$$

where $I(*)$ is the indicator function that returns 1 if the condition $*$ is true, and 0 otherwise.

2.2 Estimating Trend from Time Series

For each time series derived from a social annotation profile, we apply function fitting to obtain its estimated *trend*, denoted as $\hat{Q}(t)$. Given a time series, we are interested in *how much*, *when*, and *how fast* a trend emerges, if there is any. Based on these three requirements, we choose the sigmoid function as our trend estimator. It is defined with three parameters in Equation 1.

$$\hat{Q}(t) = \frac{\lambda}{1 + e^{-\sigma(t-\tau)}} \quad (1)$$

Parameter λ represents the asymptotic amplitude of the curve. Parameter τ indicates the time at which the series reaches half of the asymptotic amplitude, *i.e.* $\hat{Q}(\tau) = \frac{\lambda}{2}$. It is also the time at which the curve has its largest gradient. Parameter σ controls how fast the curve approaches its asymptote. The higher σ is, the faster it approaches the asymptote.

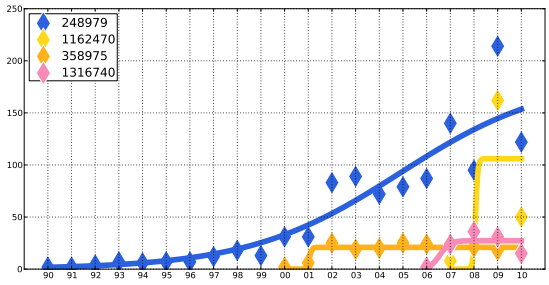


Figure 2: Sigmoid Functions and Fitting Examples

The choice of sigmoid function also matches our observation from the data at hand. When plotting the $\mathbb{Q}_i$ time series for items and $\mathbb{Q}^k$ time series for topics, we see a vivid S shape, where there is a phase with low values, followed by a transition phase from low to high values, and lastly a phase of plateau, in which values remain high and do not drop much. Figure 2 shows four examples of $\mathbb{Q}_i$ time series, which correspond to the item-wise document profiles (citation) for four publications in ACM Digital Library. It also plots the estimated sigmoid functions fitted to these time

series. We observe that these time series exhibit different amplitudes, emergence times and gradients. The fitted sigmoid functions capture all these characteristics.

Although there exist other candidate functions exhibiting an S shape, we choose sigmoid function based on empirical explorations, for it captures the three key characteristics of emerging trends, yet it makes the most general assumption about the particular shape of the curve.

Not all time series have emerging trends. We observe the following three cases where the corresponding time series cannot find emerging trend.

1. The series does not fit any sigmoid curve. This happens when the function optimizer cannot find a suitable set of parameters, *i.e.* goodness of fit is too low.
2. The series fits a downward sigmoid curve, *i.e.* the estimated σ is negative. The proposed sigmoid estimator is capable of capturing such downward trend. However, since downward trends are of less interest than upward trends, we omit them in this work.
3. The series fits a sigmoid curve, but the emergence is not visible. This happens when the estimated τ falls beyond the time range of the series.

By excluding the above three cases, we define a data series as having an *emerging trend* if it has fitted an upward sigmoid curve with the upward transition shown within its time range. In other words, a trend is *emerging* if its fitted curve satisfies both $\tau \in [1, T]$ and $\sigma > 0$.

2.3 Interpreting Emerging Trend Parameters

Given a time series with an estimated sigmoid curve satisfying an emerging trend, we interpret the three parameters defining the sigmoid curve as follows.

We interpret parameter λ as the *amplitude* of the emerging trend. It characterizes *how much* the trend emerges. We interpret parameter τ as the *emergence time* of the emerging trend. It characterizes *when* the trend emerges. We interpret the gradient Δ at $t = \tau$ as the *ruling gradient* of the emerging trend. It is derived as $\Delta_{t=\tau} = \frac{\lambda\sigma}{4}$. It characterizes *how fast* the trend emerges.

3. EXPERIMENTS

In this section, we evaluate our proposed trend discovery process. We show how the process can be employed in the following trend analysis tasks, which can potentially help the researchers and information seekers explore the different research specialties as well as the emerging publications.

1. Discovering emerging topic trends (Section 3.2). For this task, we use the corpus-wise topic profiles to find emerging topic trends.
2. Selecting important publications for a given topic (Section 3.3) and understanding the topical impact of a given publication (Section 3.4). For these topic-specific and item-specific tasks, we examine the emerging trends discovered from the item-wise topic profiles.

We conducted the experiments for task 1 on both tagging and citation datasets. Due to data sparseness in the tagging dataset, task 2 was performed on the citation dataset only.

3.1 Data Preparation

Our two data sources are CiteULike[1] (for tagging annotations) and ACM Digital Library[2] (for citation annotations). Our data dump from CiteULike is dated on May 19, 2010. It contains bookmark records to 2,419,452 items, by 49,509 users with 10,577,486 tag assignments. The bookmarks were posted between 2004 and 2010. Our data dump from ACM DL is dated on November 14, 2010. It contains 1,634,599 records, covering 14 types of publications. The earliest record was published in 1956, and the latest in 2010.

Our task 1 is concerned with publications having both tagging and citation annotations. However, the publication collections covered by CiteULike and the ACM DL are not identical. Fortunately, CiteULike provides linkout data from items in CiteULike to other digital libraries. The linkout data we obtained, dated on December 9, 2010, contains 66,388 items linked to ACM DL. Since multiple CiteULike items may be linked to the same ACM DL record, we resolved co-references and identified 64,066 distinct ACM DL records. Having extracted the citing documents for these records in ACM DL, we identified 44,123 distinct publications with both social tags in CiteULike and citation annotations in ACM DL.

We compiled a topic learning corpus consisting of the document content for all items in the joint set and all publications citing these items. Specifically, for 44,123 items in the joint set, 327,857 ACM DL documents are included for topic learning. For each document in the corpus, we concatenate the title and the abstract to form the document content. Stopwords and words appearing in less than 5 documents are removed. Documents with no more than 5 valid word tokens are also removed. As a result, 313,268 documents containing 68,725 distinct words are used for topic learning. The resulting topic model is also used as priors to learn topic assignments for social tags.

We adopt the GibbsLDA++[3] software for learning topic model from the corpus. Following [3], we set $K = 200$. Given the topic learning results, we associate a document to a topic if more than 10% word tokens in the document are assigned to the topic [3]. The choice of 10% is to filter out minor topics assigned to word tokens by chance. As a result, each document is associated to 2.03 topics on average.

3.2 Topic Trends for Annotation Corpora

In this section, we seek to compare the emerging topic trends in the social tagging community and the scientific research community by answering the following questions:

- *What are the topics that emerge mostly in each annotation community?*
- *What are the topics that emerge fastest in each annotation community?*
- *What are the topics that emerge most (or least) recently in each annotation community?*

To answer these questions, we compare the emergence amplitude (λ^k), ruling gradient (Δ^k) and emergence time (τ^k) estimated for the corpus-wise topic profiles $\mathbb{D}^k$. Due to space limitation, we show only the comparison using Δ^k.

3.2.1 Topic Trends in the Citation Community

The ruling gradient Δ^k indicates *how fast* the topic trends emerge in the annotation community. Notable topics in Table 1 include: topic 155 on *channel capacity*, topics 145 and 073 related to *wireless sensor networks*, topics 160 and 135 related to *computer vision* and topic 157 on *social community*.

[1] www.citeulike.org
[2] portal.acm.org
[3] gibbslda.sourceforge.net

Table 1: Top Topics in Citation Community

Δ^k	k	Top Keywords
6474.8	155	channel channels capacity interference spectrum
262.4	145	sensor networks nodes network wireless node
223.0	166	number asynchronous show strong consensus
172.6	073	wireless networks access network throughput
168.1	184	medical diagnosis health patients clinical care
152.7	160	image images segmentation color regions region
143.1	135	face recognition fusion facial using expressions
136.4	157	social community online communities users email
123.5	189	routing networks ad hoc network nodes multicast
122.3	130	security attacks attack secure malicious

3.2.2 Topic Trends in the Tagging Community

We note that the top topic trends in Table 2 are mostly related to web and text mining. These topics include topic 157 on *social community*, topic 089 on *recommender systems*, topic 027 on *information retrieval* and topic 104 on *tagging*. This observation suggests that the annotation community of CiteULike have been actively annotating publications in web and text mining related research. In contrast, users from other research specialty have lower surge of activies in using CiteULike.

Table 2: Top Topics in Tagging Community

Δ^k	k	Top Keywords
35.1	122	2006 2007 2005 2008 2004 thesis 2009 acm vldb
29.6	027	ir retrieval relevancefeedback relevance queryexpan
24.3	148	p2p network networks peertopeer dht overlay
23.9	068	web hypertext www hypermedia pagerank
20.4	189	routing manet adhoc sensornetworks multicast dtn
20.4	082	collaboration cscw collaborative awareness
19.1	007	mobile ubicomp pervasive ubiquitous mobility
18.9	157	social community wiki email socialnetwork blogs
18.6	089	recommender collaborativefiltering personalization
16.3	104	tagging folksonomy tag tags folksonomies 519 flickr

Note that the top keywords for topics have changed after learning on the tags. Many abbreviations now have higher probabilities of being generated by the topics.

3.3 Influential Items for Topics

Given a topic, which are the influential publications? To answer this question, we examine the topic trends estimated from the item-wise topic profiles $\mathbb{D}_i^k$. In particular, for a given topic k, we are interested in items with the largest emergence amplitude λ_i^k. As noted, λ_i^k indicates *how much* interest is found in the annotated item for topic k. We select topic 155 noted in the previous section as a case study.

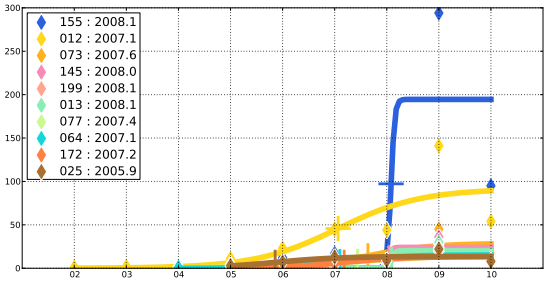
In Table 3, the top 5 items by λ_i^k are shown together with their corresponding τ_i^k, Δ_i^k, and cc_i (citation counts). It shows that, the ranking by emergence amplitude is different from that by citation counts, and the item-wise topic trends emerge around the same time as the corpus-wise topic trend for topic 155.

Table 3: Top Items for Topic 155

k	τ^k	Top Keywords		
155	2008.0	channel channels capacity interference spectrum		
λ_i^k	τ_i^k	Δ_i^k	cc_i	Title
200.0	2008.1	1018.6	2410	Elements of information theory
194.5	2008.1	1378.9	1239	Convex Optimization
146.5	2008.1	782.5	487	On Limits of Wireless Communications in a Fading Environment when UsingMultiple Antennas
93.0	2008.1	562.5	242	NeXt generation/dynamic spectrum access/cognitive radio wireless networks
43.5	2008.1	224.2	1121	Matrix computations (3rd ed.)

3.4 Emerging Topics for Items

Given a publication, which topics are mostly impacted by this work? To answer this question, we examine the item-wise topic profiles $\mathbb{D}_i^k$ for a given item. Figure 3 plots the



155 : channel channels capacity interference spectrum power
012 : optimization problem linear function optimal formulation
073 : wireless networks access network throughput protocol
145 : sensor networks nodes network wireless node sensors energy
199 : noise signal filter filtering signals filters proposed frequency

Figure 3: Emerging Trend for *Convex optimization*

top emerging topics for the book *Convex optimization* by Boyd and Vandenberghe. Two notable topics citing this book are topic 155 on *channel capacity* and topic 012 on *optimization*. By comparison, topic 012 refers to optimization theory, and topic 155 refers to an application domain of the theory. While the topic related to optimization theory shows a steady growth over the years, the topic on the application shows sharp and intense emergence. This observation suggests that much attention on the book comes from the applications, such as the specialty on channel capacity for network coding.

In our extended studies, we also observe similar patterns in other items, *e.g.* the book *Elements of Information Theory* by Cover and Thomas, as shown in Table 3. In general, for emerging trends found in citing the same theory-oriented work, topics on fundamental theories show steady growth, while topics on applications may show intense emergence.

4. SUMMARY

In this research, we proposed to use social annotations to profile scientific publications for trend discovery. We proposed a trend discovery process (shown in Figure 1) and a trend estimation method (the sigmoid estimator) for the task at hand. With the discovered trends from the social annotations, we were able to perform analysis tasks for understanding, comparing and selecting the scientific publications, helping users navigate the information space built by the social annotation communities.

5. ACKNOWLEDGEMENTS

This work is supported by Singapore's National Research Foundation under research grant NRF2008IDM-IDM004-036. We also wish to thank ACM for providing the ACM DL data for this research.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *In JMLR'03*, 3:993–1022.

[2] T. L. Griffiths and M. Steyvers. Finding scientific topics. *In PNAS'04*, 101(Supp 1):5228–5235.

[3] G. S. Mann, D. Mimno, and A. McCallum. Bibliometric impact measures leveraging topic analysis. *In JCDL'06*, pages 65–74.