Topical Establishment Leveraging Literature Evolution

Han Xu
UNSW Sydney, Australia
hanx@cse.unsw.edu.au

Eric Martin
UNSW Sydney, Australia
emartin@cse.unsw.edu.au

Ashesh Mahidadia
UNSW Sydney, Australia
ashesh@cse.unsw.edu.au

ABSTRACT

From an evolutionary perspective, a body of research is an evolving ecosystem, consisting of research topics subjected to a form of natural selection as topics come into existence, and thrive more or less over a variable period of time. Identifying the form of establishment of a given topic in a scientific domain, in terms of its *momentum* at the time of inquiry, can provide useful insights into where this topic is heading, and can facilitate effective literature research. Here we propose to identify three forms of establishment of topics, emerging from a comparison between two different methodologies in ranking papers, taking advantage of the mutual relationship between recognition of papers and recognition of topics. More specifically, by analysing the correlation between the rankings obtained by applying both methodologies, we discover thee clusters of topics, each of which is associated with a particular momentum of establishment.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

Keywords

Topical ranking, Topical establishment, PageRank, RALEX

1. INTRODUCTION AND MOTIVATION

The way a topic establishes itself through recognition in a scientific community, depends on the evolution of the hosting domain: in the scientific ecosystem, some topics wither, others manage to survive, while the most successful ones thrive. Being able to distinguish how the various topics in a scientific domain have been establishing themselves, in terms of their momentums at the time a literature survey is conducted, allows one to gain insights on the current state of a field of research, and helps predict where it is heading. We distinguish more specifically three momentums of topic establishment that are essential to someone surveying a scientific literature: *persistent*, or lasting, establishment characterises topics that lie at the core of a field which covers essential domain knowledge; *withering* establishment characterises topics that have failed in that respect and have become out of trend; *booming* establishment characterises topics whose establishment is very recent, and that are promising of a scientific breakthrough. It is challenging to identify the qualitative characteristics of a scientific topic, including its form of establishment, and it is even more so to cluster topics according to various forms of establishment. In this paper, we propose an approach where we compare the rankings of topics obtained after different methodologies have been applied, that all take advantage of the mutual reinforcement between the way a paper gains recognition and the way a topic does. More specifically, we consider two specific algorithms for ranking papers, that score papers according to their *cumulative recognition* and their *sustained recognition*, respectively. We then show how we can further distinguish between the three types of establishment of topics defined above by comparing the rankings of topics derived from the scorings of papers given by both algorithms.

The rest of the paper is organised as follows. Section 2 states our objectives and describes our generic measure of topical recognition, as well as the data we use. In Section 3, we present our simple yet intuitive approach for distinguishing between the three forms of topic establishment by comparing the rankings of topics obtained from two different methods. In Section 4, we discuss experimental results. We conclude and point to future directions in Section 5.

2. PROBLEM STATEMENT

We aim at discovering topics with a persistent, withering, or booming form of establishment in an evolving body of research to facilitate surveying the literature, particularly in the most challenging domains where complex interactions take place among many topics. More specifically, we mine rich information on how topics evolve, extracted from a network of scientific citations between publications, that embodies the cumulative research endeavours in a discipline. We then discover clusters of topics exhibiting different momentums in the way they get established. Most research efforts towards identifying topic evolution have been tracking a complete history of trends over time (e.g., [3]). We argue that a concise summarisation of global patterns on how topics get established as the literature evolves over a long period of time, as opposed to a full history of transient and local trends, is of more interest to researchers seeking quick insights into a domain at query time.

2.1 A Generic Measure of Recognition of Topic

From a corpus linguistics point of view, topics can be discovered from a collection of documents as distributions over some fixed vocabulary [1]. As topics are not hard defined, measuring their recognition is nontrivial. A closer look at the relationship between topics and papers reveals that both are tightly knitted: topics emerge from papers, and papers are mixtures of topics. Furthermore, paper recognition and topic recognition mutually reinforce each other: a highly recognised paper promotes the visibility of a given topic, helping attract more research efforts to develop the area.

978-1-4799-5569-5/14/$31.00 ©2014 IEEE.

At one extreme, a seminal work may seed a new topic with subsequent studies conforming to its terminology. At the same time, an established scientific topic draws more attention from a research community and spurs the production of papers that will be highly recognised. Intuitively, the recognition of a topic should be reflected in both its prevalence in a collection of papers and in the level of recognition of papers relevant to that topic, calculated from the underlying citation network. On the basis of the previous considerations, we propose the following relationship between the measure of recognition $I(t)$ of a a scientific topic t and the measure of recognition $d(i)$ of the ith paper in a collection:

$$I(t) = \sum_{i=1}^{N} d(i) \cdot P_t(i) \qquad (1)$$

where N is the number of papers in the collection, and $P_t(i)$ denotes the proportion of topic t in the normalised distribution of primary topics in the ith paper (Section 2.3 will provide more details on how this distribution is determined). Essentially, the recognition of a scientific topic is derived from the recognition of its instantiations: papers. Equation (1) measures the recognition of a scientific topic as the aggregated recognition of papers across the underlying collection, weighted to reflect the topic's proportion in each paper's normalised primary topic mixture. The term $d(i)$ is determined by the choice of ranking algorithm. When $d(i)$ is set to 1 for all is, $I(t)$ decays into the raw topical prevalence in the collection (assuming that papers are of equal length). The proposed Equation (1) can hence be seen as capturing *topical prevalence weighted by paper recognition.* It is generic thanks to the flexibility in the terms $d(i)$.

2.2 Data

The data we use is the paper citation network of the ACL Anthology Network (AAN) [7] (2009 release). The AAN contains 14,912 papers published in various ACL venues covering topics in the field of Natural Language Processing (NLP) and 61,171 citation links among them. Papers published in other venues, though involved in citation relationships with papers in the AAN, are not included. Each paper is an individual node in the network uniquely identified by a paper id; citations between papers are represented as directed edges. The full text of each paper along with metadata including authors names, paper title and year of publication, are also provided with the corpus.

2.3 Topic Modelling

To model topics in the AAN, we used the Latent Dirichlet Allocation [1], a highly successful unsupervised approach. Essentially, LDA is a generative probabilistic model that represents each document in a corpus as a random mixture of latent topics with each topic characterised as a multinomial distribution over words. The Bayesian generative probabilistic process posits that each document is generated by first sampling a multinomial distribution d over topics from a Dirichlet distribution and then probabilistically drawing from d a topic for each word slot in the document. Each word is in turn generated though random sampling from a topic-specific multinomial distribution over words. The optimal set of latent parameters of the generative model is obtained through approximate posterior inference.

Following [3], we ran LDA with 100 latent topics over the full texts of papers in the AAN. A manual inspection on the generated topics found 66 out of the 100 topics to be relevant and excluded the irrelevant ones from subsequent analysis. The top 10 most frequent words for several of the topics found by LDA are shown in Table 1, along with their manually assigned names. Each latent topic discovered by LDA using the full texts of papers in the AAN can be seen as a corpus-bound subfield of NLP. Though there could be definitional discrepancies between automatically discovered topics and "traditionally" perceived subfields, it can be argued that subfields discovered from texts better capture the dynamics of a field, and thus more accurately represent the topical divisions of a body of research that evolves into finer degrees of granularity [5].

Empirically, subfields of NLP are tightly knitted due to the fact that NLP applications typically consist of a pipeline with higher level techniques built on top of lower level ones, and many core techniques find their applications across topical boundaries. We have shown in [8] that papers in the AAN are indeed highly diverse in their topical compositions. As a post-processing to distill primary topics of papers in the AAN, we used an outlier-based heuristic detailed in Section 4.3.1 of [8] that considers the balance of a paper's topic mixture. The average number of primary topics per paper in the AAN is 2.3. As a final step, we renormalised each paper's primary topic mixture to have unit ℓ_1-norm.

3. OUR APPROACH

Following the introduction of our generic measure of recognition of topics in Section 2.1, we discuss in Section 3.1 two link-based methods that calculate the cumulative and sustained recognition of papers in a citation network and how they can be used to generate topical rankings pertaining to cumulative and sustained recognition of topics, respectively. In Section 3.2, we discuss how we cluster topics with momentums characterising one of the three forms of establishment discussed in Section 1, by analysing ranking discrepancies between cumulative and sustained recognition of topics.

3.1 Cumulative/Sustained Paper Recognition

Judging the recognition of scientific papers has been of interest to multiple research communities. Link-based ranking approaches such as PageRank [6] have been remarkably successful. By acknowledging citations from one paper to another as an implicit form of recognition, PageRank calculates the recognition of a paper recursively taking the recognition of its citing papers into account. However, PageRank scores nodes only from the link structure of the network; it thus has a *recency bias* against recently added nodes that have not received enough exposure to accumulate links [2]. This bias is especially pronounced in paper ranking due to the *strong ageing characteristic* exhibited in a citation network resulting from citation temporal constraint and static paper content. The lack of plasticity in link structure updates in a citation network inevitably causes PageRank to be relatively insensitive to literature evolution. Paper recognition calculated by PageRank thus pertains more to cumulative recognition; it does not reflect how well the recognition is sustained in current lines of research at query time [8].

To measure papers' sustained recognition in current literature, the intricate topic dynamics over time has to be modelled into the scoring process of paper recognition. In [8], we devised such a framework for the ranking of scientific documents, RALEX (**RA**ndom **L**iterature **EX**plorer), that scores papers with consideration to literature evolution. More specifically, RALEX implements a discriminative rank mass distribution characterised by a dynamic link following strategy that is sensitive to both topical relevance and information freshness (a measure we devised based on age and topical longevity of papers). RALEX recognises endorsement conveyed in citations initiated from current literature while

Table 1: Top 10 words for a selected subset of topics induced using LDA

Dependency Parsing	dependency parsing parser head czech based treebank projective dependencies sentence
French Function	la french le des et les en du pour est
Information Retrieval	word words query web retrieval based pages using terms corpus
Language Knowledge Base	language text information knowledge data research systems resources linguistic database
Lexical Semantics	semantic object case noun verb example relations meaning knowledge relation
Machine Translation	alignment model translation word phrase english based source models words
Probability Models	model models word language probability words data probabilities used speech
Sentiment Analysis	annotation negative positive opinion annotations sentiment polarity subjective annotators annotator
Summarisation	document documents sentence topic text sentences summary evaluation summarization information
Tagging/Chunking	word words tag pos corpus tagging tags tagger based training

discounting those carried via references originated from outdated literature; this strategy favours work that contributes more to the development of recent domains. In contrast to PageRank, scores for paper recognition calculated using RALEX better reflect their sustained recognition from a constantly growing scientific literature. We refer the reader to Section 4 of [8] for details on the design of RALEX.

3.2 Cumulative/Sustained Topic Recognition

From the previous considerations, we argue that the cumulative recognition of topics in a research domain can be derived collectively from the cumulative recognition of papers. In the same fashion, the sustained recognition of papers collectively reflects the sustained recognition of topics. We can thus calculate topics' cumulative and sustained recognition by instantiating $d(i)$ in Equation (1) with papers' PageRank scores and RALEX scores, respectively. Two topical rankings can subsequently be generated by ordering topics using their scores in reverse order. Due to the fact that both rankings of topics are generated on the same citation network, discrepancies between both rankings can be naturally attributed to literature evolution captured in RALEX's calculation of paper recognition. Intuitively, a momentum classification of the form of establishment of topics can be revealed via ranking correlation analysis.

4. RESULTS AND DISCUSSION

Arguably, topical patterns present in the most important part of a literature (i.e., top ranked papers) provide crucial insight, and also more trustworthy insight, due to the fact that papers ranked at the top are less prone to transient surges in topic evolution. We first generated two recognition rankings of papers in the AAN using PageRank and RALEX and took the top 500 papers from each ranking. We found that 366 papers are common to both sets (but with shifts in their relative ranks) and the remaining 134 papers are unique to each set. We refer to the 134 papers unique to the PageRank ranking as the "phased-out" set and the 134 papers unique to the RALEX ranking as the "included-in" set. The min, max, and average age for the phased-out set are 7, 36, 22.6, respectively, and the values for papers in the included-in set are 1, 23, 9.2, respectively. Table 2 shows the top 10 most common topics (normalised using their raw topical prevalence in both sets – recall the terminology in reference to setting $d(i)$ to 1 in Equation 1) in both sets of papers mentioned above. Topics in column 2 of Table 2 have a withering form of establishment in NLP, while those in column 3 have a booming form of establishment. By zooming into the discrepancies in raw topical prevalence among top ranked papers, this result serves to provide a microscopic view on topic evolution that may be of limited long-term significance but is still revealing. Admittedly, the use of 500 as a cutoff is somewhat arbitrary and can be improved to a decision more statistically tenable in future work.

We subsequently generated two rankings of recognition of topics as discussed in Section 3.2 based on the union of both sets of the top 500 ranked papers. Figure 1 shows the rank correlation between cumulative and sustained recognitions. Clearly, both rankings are positively correlated (the red line is the fitted linear regression), with changes in relative ranks that cluster topics into 3 momentum regions. The region above the diagonal (dashed line) marks out topics with higher ranks in sustained recognition ranking than in cumulative recognition ranking; those are topics with a booming establishment, suggesting that they have become more prominent in recent literature. In contrast, the area below the diagonal contains topics whose form of establishment has been withering in recent years. Topics whose ranks have changed dramatically[1] are highlighted with their names shown in the figure as exemplars of the above mentioned two clusters of topics. Most noticeable among topics with booming establishment, Information Retrieval stands out as a relatively young but trendy topic, whose rapid development has presumably been driven by the ever-increasing demand for IR applications in recent years. In contrast, the well-established subfield of Language Knowledge Base seems to receive a declining recent attention and is withering from the research frontier of NLP. A third cluster of topics ranked consistently high in both rankings captures a persistent form of establishment. The topics in this cluster are marked out using solid dots in Figure 1. This cluster recognises topics that have been successfully attracting a sustained attention and research interests over a long period of time and have secured their position in the core of NLP. Some exemplar topics in this cluster are Probabilistic Models and Machine Translation. Note that though it is still among the most popular topics, Lexical Semantics is demonstrating a withering establishment. It can be expected to get phased out as a popular topic of NLP if this trend persists.

Rank	Phased-out	Included-in
1	Biomedical NLP	Sub-language Processing
2	Idiom Detection	Wordnet/Ontology
3	Metaphor	Shallow Asian Language Processing
4	Lexical Semantics	Information Retrieval
5	Language Knowledge Base	Verb Classification
6	NLP Frameworks/Systems	Summarization
7	Dialogue Systems	Dependency Parsing
8	Morphology	Domain Adaptation
9	NLG evaluation	CCG Parsing
10	Phonology	Machine Translation

Table 2: Top-10 most common topics found in papers phased out from/included into RALEX's top 500 ranks

In the previous section, we have demonstrated how clusters of topics exhibiting different momentums in topical establishment can be obtained by comparing two rankings of topics. To further study the relationship among topical metrics of prevalence, cumulative recognition and sustained recognition, we plot out in Figure 2 the three measurements for each topic calculated using the entire AAN.

[1] The linear regression is with Gaussian residuals. We deem topics with regression residual more than 1 standard deviation away from the mean residual as changed dramatically.

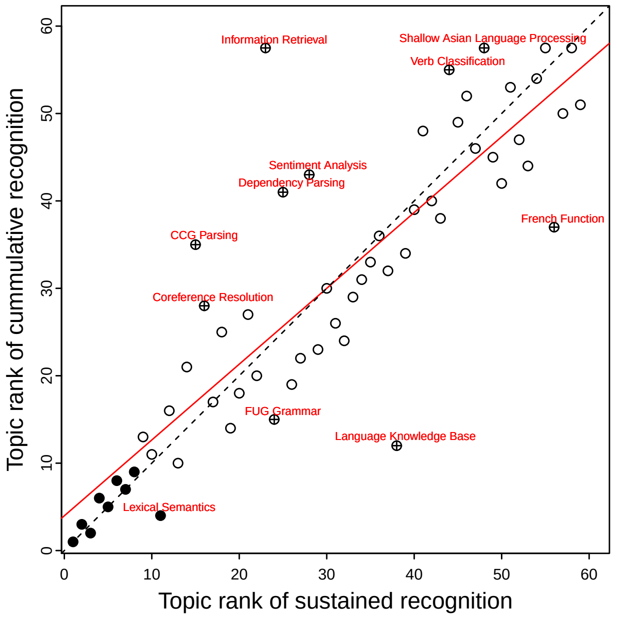

Figure 1: Cumulative/sustained recognition correlation

Topics are ordered according to their prevalence (calculated using Equation (1) with $d(i)$ set to 1) in the AAN on the x-axis. To facilitate comparison, we normalised the 3 scores using their respective range, resulting in a unit-less score scale. An interesting observation from Figure 2 is that sustained recognition is a tradeoff between cumulative recognition and prevalence: its score scale largely falls in-between those of cumulative recognition and prevalence. The fact that topics' sustained recognition agrees better with their prevalence shows that sustained recognition indeed heeds more closely to literature development. But cumulative recognition takes longer to build up and fade away, demonstrating a lag in reflecting the current landscape of a domain. In particular, the discrepancies among scores are most noticeable for prevalent topics, while they are less pronounced for less prevalent topics, and are nearly nonexistent for niche topics (e.g., Chinese NER and Biomedical NLP). This phenomenon may be explained by the topic-paper dynamics discussed in Section 2.1, that small subfields attract less research interests, thus develop more slowly than prevalent topics, making the three scores converge. To sum up, the measure of sustained recognition, being more of a synergy between topical prevalence and cumulative recognition, provides a more balanced view of the current landscape of a discipline.

5. CONCLUSION AND FUTURE WORK

In this paper, we described a simple comparison based approach of different rankings of topics to discover topics with persistent, withering and booming establishment in a scientific field. We first proposed a simple, yet intuitive generic measure of recognition of topics that implements a mutual reinforcement between paper recognition and topic recognition. We then demonstrated how rankings of cumulative and sustained recognition of topics can be obtained from rankings of cumulative and sustained recognition of papers, respectively. Finally, through correlation analysis of the rankings of topics, we found topics becoming more established, loosing their appeal, and maintaining their significance in the core of NLP by clustering them using momentum of topical establishment as demonstrated in the AAN. Quantitative analysis posited topics' sustained recognition to be a tradeoff between topical prevalence and cumulative recognition, revealing its potential in depicting a more balanced view of the landscape of a scientific domain at query time.

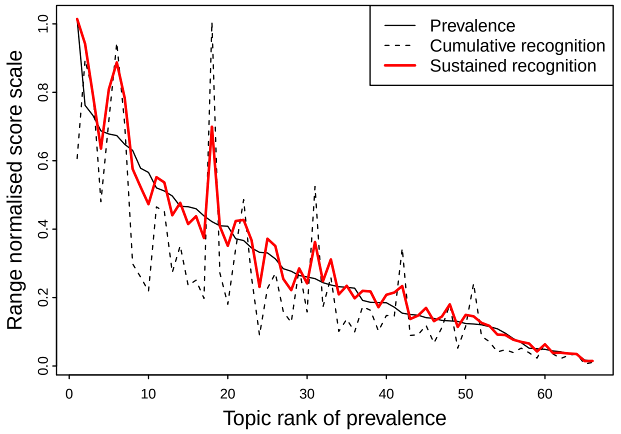

Figure 2: Prevalence, cumulative/sustained recognition

Our approach is simpler than other approaches on topic trends mining– for lack of space, we only nominate [4] as an exemplar related work. This is a result of our design philosophy. Instead of tracking topic trends longitudinally, our method provides a concise summarisation of only the lasting and most significant trends in topic evolution of a scientific domain at query time. Since it omits transient and local trends of topics as the literature evolves, our method is arguably more useful to those who seek essential insights and concise representations. In future work, we plan to experiment with more paper recognition metrics (e.g., citation counts) and we are very interested to see how the resulting rankings of topics correlate with other qualitative measures such as the ones discussed in this paper. We also intend to apply our proposed approach to other scientific fields.

6. REFERENCES

[1] D. M. Blei *et al.*, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[2] R. Ghosh *et al.*, "Time-aware ranking in dynamic citation networks," in *Proc. 11th IEEE International Conference on Data Mining Workshops*, Vancouver, 2011, pp. 373–380.

[3] D. Hall *et al.*, "Studying the history of ideas using topic models," in *Proc. Conference on Empirical Methods in Natural Language Processing*, Waikiki, HI, 2008, pp. 363–371.

[4] Y. Jo *et al.*, "The web of topics: discovering the topology of topic evolution in a corpus," in *Proc. 20th International Conference on World Wide Web*, Lyon, 2011, pp. 257–266.

[5] G. S. Mann *et al.*, "Bibliometric impact measures leveraging topic analysis," in *Proc. 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, Chapel Hill, NC, 2006, pp. 65–74.

[6] L. Page *et al.*, "The pagerank citation ranking: Bringing order to the web," Tech. Rep. 1999-66, InfoLab, Stanford Univ., Nov. 1999.

[7] D. R. Radev *et al.*, "The acl anthology network corpus," in *Proc. Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, Singapore, 2009, pp. 54–61.

[8] H. Xu *et al.*, "Contents and time sensitive document ranking of scientific literature," *J. Informetrics*, vol. 8, no. 3, pp. 546–561, July, 2014.