Topical PageRank: A Model of Scientific Expertise for Bibliographic Search

James Jardine **Simone Teufel**
Natural Language and Information Processing Group
Computer Laboratory
Cambridge University, CB3 0FD, UK
{jgj29,sht25}@cam.ac.uk

Abstract

We model scientific expertise as a mixture of topics and authority. Authority is calculated based on the network properties of each topic network. ThemedPageRank, our combination of LDA-derived topics with PageRank differs from previous models in that topics influence both the bias and transition probabilities of PageRank. It also incorporates the age of documents. Our model is general in that it can be applied to all tasks which require an estimate of document–document, document–query, document–topic and topic–query similarities. We present two evaluations, one on the task of restoring the reference lists of 10,000 articles, the other on the task of automatically creating reading lists that mimic reading lists created by experts. In both evaluations, our system beats state-of-the-art, as well as Google Scholar and Google Search indexed againt the corpus. Our experiments also allow us to quantify the beneficial effect of our two proposed modifications to PageRank.

1 Introduction

For search, the presence of links in a document collection adds valuable information over that contained in the text of the documents alone. Each act of linking can be interpreted as a latent judgement of authority or trust which is bestowed onto the linked documents (Kleinberg, 1998). This makes authority an objective measure of how important that paper is to a community who confer that authority. The citation count is the simplest of these, which has been used successfully for decades for bibliometrics (Garfield, 1972) and for mapping out scientific fields via bibliometric coupling (Kessler, 1963) and co-citations (Small, 1978). More recently, citation counts have been shown to improve effectiveness of ad-hoc retrieval (Meij and De Rijke, 2007; Fujii, 2007).

In science, the peer review process ensures that the right to cite is hard-earned, but on the web, hyperlinking is infinitely cheap. This means that that the authority of webpages cannot simply be approximated as the number of incoming links. Algorithmically more complex authority such as the random-surfer model PageRank (Brin and Page, 1998) or the authorities/hub based algorithm HITS (Kleinberg, 1998)) have spectacularly improved search results in comparison to standard IR models relying on similarity calculations based on the words in the text and other text-internal informatioh.

Much recent work in bibliographic search has been driven by the intuition that what works for the web should also work for science, even though citations are more comparable to each other in weight than hyperlinks. Case studies comparing PageRank-based authority measures against citation counts alone report some cases where PageRank is superior (Chen et al., 2007; Ma et al., 2008), but experimental proof of standard PageRank outperforming citation counts in a large-scale bibliographic search experiment is still outstanding. In at least one such experiment, PageRank performed worse than citation count (Bethard and Jurafsky, 2010).

Straightforward PageRank calculations, when applied to the scientific literature, are hampered by two factors: on the one hand, the progression of time imposes a directional structure on the citation network. Therefore, PageRank values of older papers are systematically inflated as PageRank can only ever flow from newer to older papers (Walker et al., 2007).

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 501–510,
Gothenburg, Sweden, April 26-30 2014.

Secondly, and more interestingly, researchers earn their expertise in particular, well-defined scientific fields. We propose that this requires a more fine-grained notion of specific – not global – expertise.

Our solution is to use LDA-derived topics (Blei et al., 2003) as approximations for scientific fields, and to model the importance of a paper as a mixture of its relative expertise in each of the topics it covers. The second aspect of our solution, somewhat more mundane but still necessary to adapt PageRank successfully to model scientific expertise, is to age-taper the resultant estimation.

In this paper, we present ThemedPageRank (TPR), our model of topic-specific scientific expertise, which incorporates the two modifications, and provide evidence that both are necessary for the adequate application of PageRank-style authority calculations to the scientific literature. In two evaluations, our model beats standard PageRank and citation counts by a large margin. Previous models exist which combine the idea of personalising PageRank by topics, but our manipulation of both PageRank's bias and transition probabilities differs from these. Our experiments also support the claim of our system's superiority over these models.

We use two tasks to evaluate the system's performance. The first is the reintroduction of an article's reference items that have been artificially removed. The assumption here is that a good model of document–document similarity should be able to guess which articles any given paper would have cited. The second task is the automatic creation of reading lists, of the kind that an expert might prepare for their students. We asked experts to create a gold standard of such reading lists, and compare our system against the current *de facto* state-of-the-art in such tasks, Google Scholar, and again find that our system beats it comfortably.

This article is structured as follows: the next section describes our model, which section 3 contrasts to related work. The evaluations are described in sections 4 and 5. Section 6 concludes.

2 Authority Model

Our model first determines an LDA space (Blei et al., 2003) representing the entire document collection, which results in a set of topics describing the entirety of the field. It then calculates an author-

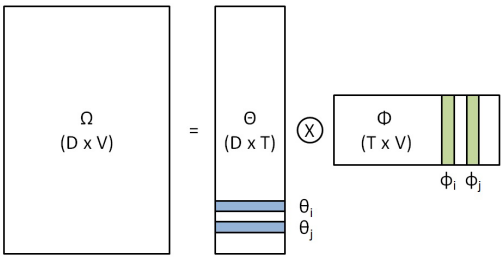

Figure 1: A High-level view of LDA.

ity model for each topic based on a modification of Personalised PageRank (Page et al., 1998). Depending on the search need, the input (one or more keyword(s) or paper(s)) is converted into a topic distribution, which we then use to linearly combine the multiple topic-specific expertise scores of our model into a unique authority score representing the fit between search need and document.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a Bayesian generative probabilistic model for collections of discrete data, which has become popular for the modelling of scientific text corpora (Wei and Croft, 2006; He et al., 2009; Blei and Lafferty, 2006). In LDA, a document in the corpus is modelled and explicitly represented as a finite mixture over an underlying set of topics, while each topic is modelled as an infinite mixture over the underlying set of words in the corpus. We use LDA predominantly to produce the latent topics that form a foundation for the relationships between papers and technical terms in a corpus.

Technical terms act as the terms in our model (rather than words), because technical terms are important artefacts for formulating knowledge from scientific texts (Ananiadou, 1994; Justeson and Katz, 1995), because descriptions of topics are better understandable using technical terms rather than words (Wallach, 2006; Wang et al., 2007); and to make our model more scalable to large corpora. The method we use to find technical terms is light-weight and requires little infrastructure, but does not represent state-of-the-art in terminology detection (Lopez and Romary, 2010; Wang et al., 2007). We collect all n-grams of words which appear in 2 or more titles of all documents in the corpus, filter out all unigrams appearing in the Scrabble TWL98 word list, then all n-grams starting or ending in stop words. To de-

cide whether a subsumed term should be removed if the subsuming term exists ("statistical machine translation" subsumes both "statistical machine" and "machine translation"), we remove those n-grams whose frequency is lower than 25% of their subsuming terms. Finally, only the most frequent 25% of the remaining unigrams and bigrams are retained.

We then build a $D \times V$ matrix Ω, which contains the counts of V technical-terms (the columns) in each of the D documents (the rows) in Fig. 1. Our own implementation of LDA (with LDA parameters $\alpha = \beta = 0.01$) is used to collapse matrix Ω into two denser, smaller matrices Θ (containing the distribution of documents over topics), and Φ (containing the distribution of topics over technical-terms).

To model topic-specific expertise in science, we modify the original PageRank calculation of Page at al. (1998) by adding a topic dimension to the score of both the bias and transition probabilities:

$$TPR(t, d, k+1) = \alpha B(t, d)$$

$$+(1-\alpha) \sum_{d' \in l_i(d)} T(t, d, d') TPR(t, d', k)$$

where $TPR(t, d, k)$ is the topic-specific PageRank of topic t for paper d at iteration k; $B(t, d)$ is the probability that paper d is chosen at random from the corpus, given topic t, and $T(t, d, d')$ is the transition probability of reaching page d from page d', given topic t. In our formula, the transition probability $T(t, d, d')$ takes into account the probabilities of topic t not only in documents d and d', but also in the other documents d'' referenced by document d':

$$B(t, d) = \frac{P(t|d)}{\sum_{d^* \in D} P(t|d^*)}$$

$$T^*(t, d, d') = \sqrt{\frac{P(t|d')}{\sum_{d^* \in D} P(t|d^*)} \frac{P(t|d)}{\sum_{d'' \in l_o(d')} P(t|d'')}}$$

$$T(t, d, d') = \frac{T^*(t, d, d')}{\sum_{d^* \in l_i(d)} T^*(t, d, d^*)}$$

Here d is a document whose TPR is being calculated, d' is a document that refers to document d and whose TPR score is being distributed during this iteration of the algorithm, and d'' is a document that is referred to by document d'. The first term in the transition function ensures that TPR scores are propagated only from citing documents that are highly relevant to topic t. The second term ensures that a larger proportion of a documents TPR score is propagated to cited documents that are highly relevant to topic t. The value $P(t|d)$ can be read directly from matrix Θ in Fig. 1.

In a final step, we age-taper TPR by dividing TPR values by the age of the citation concerned in years. Experimentally, this achieved the best model in comparison to more complex dampening methods (e.g., exponential).

3 Related Work

Others before us have observed that time effects bias PageRank if applied unmodified to the scientific literature (Walker et al., 2007). Walker et al.'s CiteRank algorithm modifies the bias probabilities of PageRank exponentially with age, favouring more recent publications.

We are also not the first to have combined a notion of topic-specification with Personalised PageRank. The idea goes back to the original PageRank paper by Page et al. (1998), who discuss the personalization of PageRank by introducing a bias towards only a set of trusted web sites W. Page et al. alter only the bias probability B, while leaving the transition probabilities T unchanged from global PageRank:

$$B(t, d) = \begin{cases} \frac{1}{|W|} & \text{if } d \in W \\ 0 & \text{if } d \notin W \end{cases}$$

$$T(t, d, d') = \frac{1}{|l_o(d')|}$$

Richardson and Domingos (2002) first used PageRank personalisation for specialisation at search time. For query q with corresponding topic $t = q$, they use the relevance of document d to query q as a bias. Haveliwalla (2003) calculates a Personalised PageRank for each of a set of 16 manually created topics t comprised of several documents by altering only the Bias term B, using Page et al.'s formula above. This solution avoids the computational scalability problem with Richardson and Domingos' approach, but is limited in applicability by requiring predefined topics. Several researchers followed Brin and Page and Haveliwala in altering only the bias

probabilities, including Wu et al. (2006) and Gori and Pucci (2006).

In contrast, Narayan et al. (2003) and Pal and Narayan (2005) propose a model of personalisation that alters the transition probabilities instead of the bias probabilities. Under their model, the transition probability $T(t, d)$ is proportional to the number of words in document d that are strongly present in the documents contained in topic t. Nie et al. (2006) produce a more computationally scalable version of the ideas presented in Pal and Narayan (2005) by associating a context vector with each document, with a fixed set of topics (12 in their case), for which they learn these context vectors using a naive Bayes classifier. They then provide the possibility to alter both the bias and transition probabilities of each webpage as follows:

$$B(t, d) = \frac{1}{D} C_t(d)$$

$$T(t, d, d') = \gamma \frac{1}{|l_o(d')|} + (1 - \gamma) \sum_{t' \neq t} \frac{C_{t'}(d')}{l_o(d')}$$

where $C_t(d)$ is the context vector score for topic t associated with document d; the first term in $T(t, d, d')$ corresponds to the probability of arriving at page d from other pages in the *same* topic context; the second term is the probability of arriving at page d from other pages in a different context; and γ is a factor that weights the influence of same-topic jumps over other-topic jumps. Their results suggest that γ should be close to 1, indicating that distributing PageRank within topics generates better Personalised PageRank scores.

Other than the fact that they treat bias and transition probabilities differently to how we treat them, all personalisation methods discussed up to now have the disadvantage that they rely on a fixed list of manually selected topics, whereas our method offers adaptive specialisation to corpus or domain.

The previous work closest to ours is Yang et al. (2009), who were the first to use LDA to automatically discover abstract topic distributions in a corpus of scientific articles, and to combine them with Pagerank by – in principle – altering both the bias and transition probabilities according to the following model:

$$B(t, d) = \frac{1}{D} P(t|d)$$

$$T(t, d, d') = \gamma T_{s\_t}(t, d, d') + (1 - \gamma) T_{o\_t}(t, d, d')$$

$$T_{s\_t}(t, d, d') = P(d|d', t) \cong \frac{1}{|l_o(d')|}$$

where T is the number of LDA topics, $P(t|d)$ is a probability of topic t given document d, which can be read directly from the generated LDA probabilities, T_{s_t} is the probability of arriving at page d from other pages in the same topic context, whereas T_{o_t} treats the case of arriving at a different topic. Like Nie et al., they achieve best results with $\gamma = 1$, so they ultimately only use bias probabilities, like the models discussed above. Crucially, their decision that $P(d|d', t)$ does not to involve any of the LDA topic distributions is surprising. Under their model, as in ours, when the reader randomly jumps to a new paper, they will tend to favour papers that are closely associated with the topic. However, when they follow a citation in Yang et. al's model, one is picked with equal probability. In contrast, our model implements the obvious intuition that if one follows citations, one should also favour those that are closely associated with the topic.

Let us now turn to the task of reference list reintroduction (RLR), i.e., the prediction of which papers a target papers originally cited, given only some information about the paper which stands in as a search need – either its abstract, author names and other bibliometric information, and/or the full text of a paper (with citation information redacted). Evaluation of a search model by RLR is cheap because of the readily available gold standard, and it thus allows for experiments with large data sets.

State-of-the-art solutions to RLR combine lexical similarity (often via topic models), measures of authority over a citation graph, and information about social constructs and historic patterns of citation behaviour. Strohman et al. (2007) perform RLR with the paper text as a query to their recommendation system, using text similarity, citation counts, citation coupling, author information, and the citation graph. Their model achieves a mean-average precision of 0.102 against a corpus from the Rexa10 database. Bethard and Jurafsky (2010) improve on Strohman et al. by the use of a SVM with 19 features from 6 broad categories: similar terms; cited by others; recency; cited using similar terms; similar topics; and social habits. They achieve a MAP of

0.279 against the ACL Anthology Reference Corpus (Bird et al., 2008), with the following features performing best: publication age, citation counts, the terms in citation sentences, and the LDA topics of the citing documents. They also use (unchanged) PageRank authority counts as one of the features, but find that it provides little discriminative power to the SVM. A drawback of their method is the large amount of information that has to be provided to create their SVM features, and the expensive training routine, which is based on pairwise paper–paper comparisons in the corpus.

Variations of the RLR tasks exist, which additionally determine the position in the text of a paper where each recommended citation should occur (Tang and Zhang, 2009; He et al., 2011; Lu, 2011), a task which is typically solved by comparing a moving window in the query paper against millions of previously located citation contexts with. The drawback of this technique in contrast to ours is the fact that new papers, which have not collected sufficient contexts in the literature, are severely disadvantaged and will never be recommended.

We first create topics and then apply PageRank to find expertise within topical networks. It is however also possible to simultaneously model citations and terms (Cohn and Hofmann, 2001; Mann et al., 2006). Such models are not normally directly comparable to ours; for instance Bharat and Henzinger's (1998) model, a modified version of HITS (Kleinberg, 1998), is query-specific.

There are numerous extensions to LDA that incorporate external information in addition to the lexical information inside the documents in a corpus, via author-topic models and models of publication venues (Steyvers and Griffiths, 2007; Rosen-Zvi et al., 2010; Tang et al., 2008). Erosheva et al. (2004) model a corpus using a multinomial distribution simultaneously over the citations and terms in each document. Topics (which they call aspects) are associated with a list of the most likely words (interpretable as topics) and citations (interpretable as authorities) in that aspect. Extensions of the model exist (Nallapati and Cohen, 2008; Gruber et al., 2007; Chang and Blei, 2010; Kataria et al., 2010; Dietz et al., 2007).

We avoid the tight coupling of topic discovery and citation modeling that the above-mentioned works follow for several reasons. Firstly, such models only work for papers and citations that were present during the learning stage, and there is no mechanism for predicting influential citations for topics in general, or for combinations of topics. The tight coupling might also result in overlooking some authorities, namely those that are authoritative across several topics, which will be penalised via low joint distribution probabilities in combined methods because of the division of the probabilities across several topics. Secondly, and more disturbingly, such models will not locate topics that lack an authority because the authority component of the joint distribution will be near-zero. This rules out niches in a corpus where papers are equally relevant to each other, or where the niches are so young that they do not yet have an established citation network. There is also a scalability issue with joint models of topics and citations. The evaluation data used in coupled models is generally small, with the number of papers ranging under around 2,000, the number of citations ranging under 10,000, and the number of topics in their models ranging from eight to twenty. But LDA has been shown to scale to corpora of millions of terms (Newman et al., 2006), and PageRank to billions (Page et al., 1998) of documents. Our model, which advocates a pipelined approach, benefits from the fact that separate topic modelling is computationally tractable using LDA, and the fact that citation graph modelling is cheap using Personalised PageRank.

4 Evaluation 1: RLR

We evaluate our authority-based search model using the 2010 ACL Anthology Network (Radev et al., 2009). We removed from it corrupted documents, i.e., those of less than 100 characters or containing only control characters. The ACL Anthology Network provides external meta-data about the articles, which was manually curated. We do not use this meta-data because we wanted to build as system that can be applied to any large collection of articles, where external meta-data would not normally exist. We therefore build an approximate citation graph from the paper text itself, as a one-off task when constructing the LDA space. We extract titles, dates and full-text from every article and perform a search of each articles title in the full-text of all other

Model	MAP
800 test papers, as in B&J (2010)	
B&J; best model	0.287
TPR-NoDB	0.264
TPR-NoAge	0.267
TPR	**0.302**
10,000 test papers	
A: NFIDF Cosine	0.062
B: NFIDF + citation count	0.092
C: NFIDF + global PageRank	0.099
D: NFIDF LDA (KL divergence)	0.115
E: TPR-NoDB	0.233
F: TPR-NoAge	0.242
G: **TPR**	**0.268**

Figure 2: RLR results

articles (i.e., under the assumption that the reference list is the (only) place where we will find such titles).

Our system generates the RLR output (the recommended articles) for an article d by extracting technical terms as described in section 2, examining the topic distribution for that article $\theta_{d,t}$ (i.e. a θ_i in Fig. 1). We use the topic distribution of article d in place to generate the unique age-adjusted TPR tailored to the article, $TPR(d, d')$. The 100 articles d' with the highest ThemedPageRanks are recommend as citations for article d. Results are reported as mean average precision (MAP) of these 100 documents against the actual citations in the article.

We first compare our model to the state-of-the-art (Bethard and Jurafsky, 2010). We emulate their experimental setup by including only the pre-2004 articles in the corpus and testing only on the roughly 800 2005/6 articles with more than 5 intra-corpus citations in their reference list, for which we have per-paper average precision scores. The top part of Fig. 2 shows that our model (MAP=0.302) outperforms their best model (MAP=0.287; difference at 5% confidence with Wilcoxon Ranked Squares test), despite our model being a general, light-weight IR system, which relies on LDA and PageRank alone, and theirs is a specialised state-of-the art system, which relies on heavy-weight machine learning and on additional sociological features.

The lower part of Fig. 2 compares the influence of citation count, global PageRank, topic similarity, and combinations of topic similarity with citation counts or global PageRank, and our model (TPR). For these tests, we use the entire corpus of 10,000 papers with more than 5 citations. Over the baseline (**A**), n-gram-frequency-inverse-document-frequency (NFIDF), both citation counts (**B**) and global PageRank (**C**) make a small improvement. Global LDA similarity scores (**D**) fare little better.

As the performance of the full model (**G**; MAP=0.268) shows, the inclusion of topic models lead to a large improvement over any of the above. This is, as far as we are aware, the first time that a large-scale evaluation that finds significant improvements of a PageRank implementation over citation counts in scientific search.

We next consider our two modifications, age-adjusting **(E)** and double-biasing **(F)**, in isolation. We use two versions of our system where we switched off age-tapering and double-biasing (ie., we only work with a change in the bias probabilities, as do Nie etal. (2006), Havaliwala (2003) (although their models do not include automatically generated topics) and Yang et al. (2009)). Our model comfortably outperforms TPR-NoDB in both the 800 and 10,000 paper experiment. Similarly, the effect of age-tapering alone can be seen from the performance of TPR-NoAge (our model without age-adjusting), in the difference between 0.267 and 0.302 and that between 0.242 and 0.268 (significant at 99%). This confirms our claim that a topic-specific age-tapered PageRank is superior to global PageRank in scientific citation networks.

5 Evaluation 2: Reading Lists

The aim of the second experiment is to test our model against a much cleaner, albeit smaller gold standard: on the task of reconstructing the material of expert-created reading lists. We compare our system's performance to three standard, commonly used search engines: Lucene TFIDF, the Google-indexed ACL Anthology, and Google Scholar. We chose Google-index and Google Scholar because they represent commonly used state-of-the-art commercial search engines, and the Google-index is what is currently offered as the standard ACL Anthology search tool. In contrast, Lucene TFIDF was chosen to represent an easy-to-interpret, reproducible, out-of-the-box baseline implementing the simplest kind of lexical similarity search without any notion of authority. Of the three search engines,

we would predict Google Scholar to be the toughest competitor to TPR, because it uses citation information directly and it is reasonable to expect that the Google Scholar algorithm employs some domain adaptation to the scientific domain.

We created gold standard expert-written reading lists using the following protocol. Eight experts were recruited from the computational linguistics groups of two universities (3 from one, 5 from the other). All experts had a PhD in computational linguistics and several years of research experience. They were asked to choose a subject for an (imaginary or existing) reading list for an MPhil student, concerning an area in which they know the literature well. We purposefully did not give them guidance as to the size of the reading list as we wanted to observe how experts create reading lists. During the interview, the experimenter documented the final list chosen by the expert and made sure all papers chosen were present in the 2010 version of the ACL Anthology Network.

This procedure resulted in reading lists of the following topics and sizes: statistical parsing (22 papers); parser evaluation (4); distributional semantics (14); domain adaptation for parsing (11); information extraction (9); lexical semantics (14); statistical machine translation models (5); and concept-to-text generation (16).

In our retrieval model, which topic distribution is chosen for a query depends on whether the query is an exact match to one of the technical terms found by our model. If it is, then the topic distribution of the technical term is used directly as the query topic distribution $\theta q, t$ (i.e. a transposed renormalized ψ in Fig. 1). If not, we perform a keyword-based search (using Lucene TFIDF), and use the average topic distribution of the top 20 documents returned as the query topic distribution (i.e. several θi in Fig. 1). The query topic distribution is then used to linearly combine the topic-specific TPRs into a unique TPR tailored to the query. The 20 documents with the highest TPR are recommended.

The three baselines are used as follows in the experiment: The experiment is performed by issuing the topic of the reading list (exactly as given to us by the experts) as a key-word based query to each system and recording the top 20 resulting papers answers. For Lucene TFIDF, we downloaded Lucene.NET v2.9.2 and indexed our 2010 snapshot of the ACL Anthology using standard Lucene parameters for the TFIDF model. For the Google-indexed ACL Anthology (AAN), we use the interface provided on the ACL Anthology website. In order to provide an identical search ground, we automatically exclude from the return lists papers added after the creation of the AAN snapshot. For Google Scholar (GS), we use the interface provided at `scholar.google.com`, and parse returns to exclude non-AAN material semi-automatically. In the case of Google Scholar, we restrict the search ground to the ACL Anthology by filtering the top 200 return sets (which may lead to fewer than 20 papers returned).

We report FCSC, RCSC and F-score for each algorithm. FCSC and RCSC are new metrics which address the problem that F-score, being binary, does not support the notion of a "close hit", combined with the fact that we require a fine-grained comparison of the quality of different systems retrieved lists despite the small size of our gold standard. Citation Substitution Coefficient (FCSC), a new metric for RLR, gives higher scores to papers closely related to the target papers by citation distance. The FCSC of each expert paper is the inverse of the number of nodes in the minimal citation graph connecting each expert paper to any system-retrieved paper (thus ranging between 0 and 1; non-connected expert papers receive a zero score). We also introduce Reverse Citation Substitution Coefficient (RCSC), which measures the inverse of the number of nodes in the minimal citation graph connecting each system-retrieved paper to any expert paper. RCSC makes sure that systems cannot simply increase their FCSC values by returning many irrelevant papers. RCSC thus corresponds to precision, while FCSC corresponds to recall. The system RCSC and FCSC scores we report are the average scores of all the system-retrieved and expert papers, respectively. Reporting both scores gives a good overall picture of system performance, particularly when read together with the F-score.

Fig. 3 shows that our model comfortably beats the competitor systems according to all metrics. In particular, our model $>$ GS/AAN $>$ Lucene TFIDF[1].

[1]For FCSC, the differences are statistically significant at

	FCSC	RCSC	F-score
AAN/Google	0.527	0.317	0.117
GS	0.519	0.364	0.112
Lucene TFIDF	0.412	0.330	0.040
TPR	**0.563**	**0.456**	**0.128**

Figure 3: Reading List Creation: Results.

	FCSC	RCSC
TF-CC	0.419	0.359
TF-CC-A	0.491	0.442
TF-PR	0.450	0.360
TF-PR-A	0.512	0.407
TPR-NoDB	0.541	0.440
TPR-NoAge	0.526	0.436

Figure 4: Citation counts and PageRank variants.

Concerning simpler methods of estimating authority, Fig. 4 shows that a multiplication of TFIDF by citation count (as Fujii (2007) does) results in a FCSC/RCSC of 0.419/0.359 (reported as TF-CC), and age-tapering of citation-count by dividing the citation count by the age of the paper in years (reported as TF-CC-A) results in FCSC/RCSC of 0.491/0.442. We again compare different versions of PageRank. Global PageRank can be built into the system by simple multiplication of PR scores as above, with and without age-tapering (reported as TF-PR and TF-PR-A, respectively). We observe a similar effect to the one reported by Bethard and Jurafsky and seen in experiment 1, namely that global PageRank only performs similar to citation counts (0.450/0.360 vs 0.419/0.359). With respect to double-biasing and age-tapering we see the same effect as in experiment 2[2]. In fact, we can see from these results that global PageRank barely improves over standard TFIDF, while age-tapering even without topics already brings quite some improvement. Overall, these results confirms our claim of the superiority of a topic-specific PageRank over global PageRank in scientific citation networks.

6 Conclusions

We present here the first experiments that pinpoint which modifications to PageRank are necessary to adequately cater for the highly specialised situation we encounter in science. The modification we suggest are to use LDA-derived topics (Blei et al., 2003) as approximations for scientific fields, to calculate authority in a topic-specific way, and to age-taper the authority scores. We present formulae where topics personalise both the bias and the transition probabilities. This results in a general IR model for science incorporating a robust notion of authority. Our implementation requires only minimal resources and relies only on LDA and PageRank calculation, which means that it is efficient during training, retraining and at search time.

We perform two evaluations. In both, our model significantly outperforms not only state-of-the-art, but also standard PageRank, non-age-tapered (but topical) PageRank, and non-topical (but age-tapered) PageRank. Our model achieves its competitive performance by using only the raw text and citation links. It requires no external information, neither explicit sociological information such as past collaborations between authors, nor the expertise and cooperation of like-minded readers, as collaborative models do. While successful applications of collaborative filtering to bibliometric search are rife (Goldberg et al., 2001; Agarwal et al., 2005; McNee et al., 2006; Torres et al., 2004), including to reading list generation (Ekstrand et al., 2010), we wanted an entirely independent authority-based IR model similarity. CF also suffers from a cold-start phenomenon, where recommendations are generally poor where data is sparse, and has to wait for papers to be rated by a large number of authors (rather than cited) before it can rank them.

Should the reader wish to evaluate the performance of TPR on their own PDF papers, it has been incorporated into the Qiqqa reference management software [3].

99% confidence via a two-tailed Wilcoxon Signed Ranks test, except that between GS and AAN (for which the confidence interval is only 96%) and that between Lucene and AAN, where it is 98%. Non-parametric paired tests such as the Wilcoxon Signed Ranks test can be used on FCSC, but not on RCSC, as there are different sets of underlying system-retrieved papers in each case. For RCSC, differences between our model and all others at 99% confidence interval, between GS and AAN/Lucene TFIDF at the 95% interval. F-score is reported for completeness.

[2]Wilcoxon Signed Rank test found all differences significant at the 99% level, except that between TF-PR and Lucene TFIDF (significant only at the 90% level), and the following equivalences: Lucene TFIDF = TF-CC; TF-PR = TF-CC; TF-CC-A = TF-PR-A; TF-CC-A = TF-PR.

[3]Available at http://www.qiqqa.com

References

N. Agarwal, E. Haque, H. Liu, and L. Parsons. 2005. Research paper recommender systems: A subspace clustering approach. *Advances in Web-Age Information Management*.

S. Ananiadou. 1994. A methodology for automatic term recognition. In *Proceedings of COLING*.

S.K. Pal B. Narayan, C. Murthy. 2003. Topic continuity for web document categorization and ranking. In *IEEE/WIC International Conference on Web Intelligence*.

B.D. Davison B. Wu, V. Goel. 2006. Topical trustrank: Using topicality to combat web spam. In *Proceedings of the 15th international conference on World Wide Web*.

S. Bethard and D. Jurafsky. 2010. Who should i cite: learning literature search models from citation behavior. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*.

K. Bharat and M.R. Henzinger. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR*.

S. Bird, R. Dale, B.J. Dorr, B. Gibson, M.T. Joseph, M.Y. Kan, D. Lee, B. Powley, D.R. Radev, and Y.F. Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. of LREC08*.

D.M. Blei and J.D. Lafferty. 2006. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, page 147. Citeseer.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

J. Boyd-Graber, D. Blei, and X. Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of EMNLP-CoNLL*, pages 1024–1033.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*.

J. Chang and D.M. Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.

P. Chen, H. Xie, S. Maslov, and S. Redner. 2007. Finding scientific gems with google's pagerank algorithm. *Journal of Infometrics*, 1(1):8–15.

D. Cohn and T. Hofmann. 2001. The missing link-a probabilistic model of document content and hypertext connectivity. *Advances in neural information processing systems*, pages 430–436.

L. Dietz, S. Bickel, and T. Scheffer. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, page 240. ACM.

M.D. Ekstrand, P. Kannan, J.A. Stemper, J.T. Butler, J.A. Konstan, and J.T. Riedl. 2010. Automatically building research reading lists. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 159–166. ACM.

E. Erosheva, S. Fienberg, and J. Lafferty. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220.

A. Fujii. 2007. Enhancing patent retrieval by citation analysis. In *Proceedings of SIGIR*.

E. Garfield. 1972. Citation analysis as a tool in journal evaluation. *American Association for the Advancement of Science*.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151.

M. Gori and A. Pucci. 2006. Research paper recommender systems: A random-walk based approach. *IEEE Computer Society*.

A. Gruber, M. Rosen-Zvi, and Y. Weiss. 2007. Hidden topic markov models. In *Proceedings of AISTATS*.

T.H. Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, pages 784–796.

Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles. 2009. Detecting topic evolution in scientific literature: how can citations help? In *Proceeding of the 18th ACM conference on Information and knowledge management*.

Q. He, D. Kifer, J. Pei, P. Mitra, and C.L. Giles. 2011. Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining*.

J.S. Justeson and S.M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(01):9–27.

S. Kataria, P. Mitra, and S. Bhatia. 2010. Utilizing Context in Generative Bayesian Models for Linked Corpus. In *Proceedings of AAAI*.

M.M. Kessler. 1963. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25.

J. Kleinberg. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*. Also available from `http://www.cs.cornell.edu/home/kleinber/`.

P. Lopez and L. Romary. 2010. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. In *SemEval 2010 Workshop*.

Y. et al. Lu. 2011. Recommending citations with translation model. In *Proceedings of the 20th ACM international conference on Information and knowledge management*.

N. Ma, J. Guan, and Y. Zhao. 2008. Bringing pagerank to the citation analysis. *Information Processing and Management*, 44(2):800–810.

G.S. Mann, D. Mimno, and A. McCallum. 2006. Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*.

S.M. McNee, J. Riedl, and J.A. Konstan. 2006. Making recommendations better: an analytic model for human-recommender interaction. In *CHI'06 extended abstracts on Human factors in computing systems*.

E. Meij and M. De Rijke. 2007. Using prior information derived from citations in literature search. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*.

R. Nallapati and W. Cohen. 2008. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *International Conference for Weblogs and Social Media*.

D. Newman, P. Smyth, and M. Steyvers. 2006. Scalable Parallel Topic Models. *Journal of Intelligence Community Research and Development*.

L. Nie, B.D. Davison, and X. Qi. 2006. Topical link analysis for web search. In *Proceedings of SIGIR*.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. *Stanford Digital Library Technologies Project*.

D.R. Radev, P. Muthukrishnan, and V. Qazvinian. 2009. The ACL Anthology Network Corpus. In *Proceedings, ACL Workshop on NLP and IR for Digital Libraries*, Singapore.

M. Richardson and P Domingos. 2002. The intelligent surfer: Probabilistic combination of link and content information in pagerank. *Advances in neural information processing systems*, 14:1441–1448.

M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. 2010. Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)*, 28(1):1–38.

B. Narayan S.K. Pal. 2005. A web surfer model incorporating topic continuity. *IEEE Transactions on Knowledge and Data Engineering*, 17:726729.

H.G. Small. 1978. Cited documents as concept symbols. *Social Studies of Science*, 8:327–340.

M. Steyvers and T. Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of latent semantic analysis*, page 427. Erlbaum, Hillsdale, NJ.

T. Strohman, W.B. Croft, and D. Jensen. 2007. Recommending citations for academic papers. In *Proceedings of SIGIR*.

J. Tang and J. Zhang. 2009. A discriminative approach to Topic-Based citation recommendation. *Advances in Knowledge Discovery and Data Mining*.

J. Tang, R. Jin, and J. Zhang. 2008. A topic modeling approach and its integration into the random walk framework for academic search. In *Eighth IEEE International Conference on Data Mining*.

R. Torres, S.M. McNee, M. Abel, J.A. Konstan, and J. Riedl. 2004. Enhancing digital libraries with TechLens+. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*.

D. Walker, H. Xie, K.K. Yan, and S. Maslov. 2007. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007:P06010.

H.M. Wallach. 2006. Topic modeling: beyond bag-of-words (powerpoint). In *Proceedings of the 23rd international conference on Machine learning*.

X. Wang, A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE international conference on data mining*.

X. Wei and W.B. Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of SIGIR*.

W. Wong, W. Liu, and M. Bennamoun. 2009. A probabilistic framework for automatic term recognition. *Intelligent Data Analysis*, 13(4):499–539.

Z. Yang, J. Tang, J. Zhang, J. Li, and B. Gao. 2009. Topic-level random walk through probabilistic model. *Advances in Data and Web Management*.

D. Zhou, S. Zhu, K. Yu, X. Song, B.L. Tseng, H. Zha, and C.L. Giles. 2008. Learning multiple graphs for document recommendations. In *Proceeding of the 17th international conference on World Wide Web*.