Contents and Time Sensitive Document Ranking of Scientific Literature

Han Xu Eric Martin Ashesh Mahidadia

University of New South Wales, Australia
{hanx,emartin,ashesh}@cse.unsw.edu.au

Technical Report
UNSW-CSE-TR-201408
March 2014

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

A new link-based document ranking framework is devised with at its heart, a contents and time sensitive random literature explorer designed to more accurately model the behaviour of readers of scientific documents. In particular, our ranking framework dynamically adjusts its random walk parameters according to both contents and age of encountered documents, thus incorporating the diversity of topics and how they evolve over time into the score of a scientific publication. Our random walk framework results in a ranking of scientific documents which is shown to be more effective in facilitating literature exploration than PageRank measured against a proxy of ground truth based on papers' potential usefulness in facilitating later research. One of its many strengths lies in its practical value in reliably retrieving and placing promisingly useful papers at the top of its ranking.

1 Introduction and Motivation

The explosive growth of the Internet and the overabundance of data fuel the creation and development of information networks, which constantly poses new challenges for information retrieval. As the searched domains expand, even queries targeted at some niche field retrieve a large volume of potentially relevant information that far exceeds human processing capabilities. Ranking addresses the challenge of information overload by identifying material of the highest "quality" among all "relevant" material; it has become an integral part of virtually any information retrieval system.

Scientific citation networks are a specific type of information network: they consist of academic publications connected by citations that embody the cumulative research endeavours in scientific domains. Researchers are better enabled to make scientific breakthroughs by taking advantage of current knowledge—borrowing from insights of past studies and availing themselves of data gathered and systems developed—, making exploring citation networks crucial in conducting research. However, citation networks are large in scale and dynamic in nature with high concentration of information and intricate interactions among academic entities (e.g., authors, papers, concepts), making them particularly challenging to navigate. For those reasons, the effective exploration of a citation network requires that high-quality work be identified through ranking. Specifically, recognising publications that have the potential to facilitate later research, or publications with *high scientific utility*[1], is of special interest as they form the most fruitful part of a scientific field and serve as solid starting points to further explore new possibilities. Intuitively, the scientific utility of a paper is not a constant measure but relative to its contents and age: subsequent research is more likely to benefit from a relatively recent work in a highly relevant topical area whose scientific merits have not yet been fully exploited. In this paper, we present a new link-based ranking framework aimed at helping researchers in locating work that contains useful information for them to make progress in their own studies. More particularly, we design our ranking framework to account for both the contents and age of a paper, producing a ranking of papers that better reflects their potential scientific utility.

The rest of the paper is organised as follows. Section 2 briefly surveys the field of scientific document ranking. In Section 3, we identify the gaps in the literature on scientific document ranking. We fill those gaps in Section 4, where we present our link-based ranking framework, specifically designed to model human literature explorers more accurately by taking both document contents and age into consideration. Section 5 discusses experimental results. In Section 6, we conclude and point to future directions.

2 Scientific Document Ranking

Scientific document ranking is a challenging task whose core problem is to quantify the importance of academic publications. Citation count based metrics have a long lineage, tracing back to the pioneering work done by Garfield on citation analysis in the 1970's [10, 11], and they are still widely used today. However, citation count has been challenged for being a quantitative measure of the popularity of a scientific document that fails to properly capture qualitative aspects, such as potential scientific impact [26, 25, 18, 20, 24]. The unique challenge of measuring the scientific value of academic publications has fuelled ever-increasing research efforts.

Link-based ranking approaches such as PageRank [22] have been remarkably successful in ranking webpages. By recognising hyperlinks (a form of citation) from one page to another as an implicit conveyance of authority, PageRank calculates the prominence of a page using a less democratic vote, taking the quality of the citing pages into account. Inspired by the success of PageRank in Web search and similarities in problem formulation, a plethora of research has been carried out to rank scientific documents thanks to PageRank or variants of it. [7] directly used PageRank to assess the relative importance of publications in Physical Review journals. By identifying outliers with a moderate number of citations but a high PageRank score, exceptional papers are promoted that would otherwise be disregarded in a citation count based ranking. [18] also directly applied PageRank to citation analysis and compared their results with the traditionally used citation count metrics. Both studies concluded that, at least empirically, PageRank yields qualitative rankings that resonate better with human judgements. More ambitious studies aim at adapting PageRank to avail themselves of the complex interactions among various academic entities (e.g., publications and authors) in citation networks to incorporate more diverse sources of information. Inspired by the mutual reinforcement between authors and papers—high quality papers are written by renowned researchers and prestigious researchers write high-impact articles—, [30] proposed the Co-Ranking framework that couples two random walks on the heterogeneous network of authors and papers to produce rankings for both entity types. With similar rationale, [15] proposed a simple alternative where basic PageRank was used to generate entity type sensitive rankings in a heterogeneous network of authors, papers, venues, institutions, and terms. Another line of research focusses on generating relative rankings of papers in the "domain frontier". PageRank in its original formulation fails this task as it is negatively biased against young papers that have not yet been given enough exposure to attract citations. To address this issue, some studies proposed to adapt PageRank and take time into account to promote the ranking of recent papers. [25] designed CiteRank as a random surfer visiting papers to which the assigned probabilities are exponentially discounted as a function of their age. In a similar work, [17] employed both an exponential time decay factor and a trend factor calculated using recent citation time series of a paper to elevate the rankings of new papers. [24] incorporated both a temporal penalty and heterogeneous entities ranking (in this case, authors and papers) into their FutureRank

[1] From here onwards we use the terms utility and usefulness interchangeably.

system. They found that time decay plays a much more important role in producing good rankings than the mutual reinforcement between authors and papers.

3 Problem Statement

We aim at ranking documents in a citation network to help researchers identify papers of high scientific utility in their field, a paper's usefulness being acknowledged in the kind of incoming citations it receives from later work. A scientific citation network has the same abstract structure as any other directed network, but it is distinctively static in nature: the contents of a document and the references it includes are frozen at the time of publication, imposing a strict temporal constraint on the link structure of a citation network. At any stage, papers only refer to the literature at the time of writing; no pointer to subsequent work can ever be additionally provided. This feature accounts for the fact that a citation network has a strong *ageing characteristic*: first, the temporal constraint causes directed links to point towards progressively older nodes; second, the static nature of a citation network constrains it to evolve with less plasticity than the Web which is constantly updated both in contents and structure. This strong ageing characteristic makes ranking metrics based on citation count unsuitable for the task, although they are traditionally used, as new publications have not been given enough exposure to accumulate citations. Furthermore, it limits the appropriateness of more sophisticated approaches, such as PageRank, which strongly biases its rankings towards older papers. This undesirable bias has the profound implication that rankings produced by PageRank pertain more to "historical importance" and are not reflective of the scientific landscape of a literature at query time, thus severely limiting their usefulness to a researcher exploring current literature for inspiration. We see this as a fundamental issue in scientific document ranking that has not yet been properly addressed. A few studies (e.g., [25, 17, 24]) have attempted to, to some extent, counter the bias towards older documents, but they have the narrower aim of ranking new papers. Furthermore, existing studies in scientific document ranking completely disregarded documents' contents when modelling a random reader, not reflecting the fact that the way a researcher explores the scientific literature is highly contents-aware. Not taking document contents into account is inadequate: a ranking is likely to be more satisfactory if it integrates the diversity of topics in a given domain and their dynamics over time [19, 14]. Those dimensions are arguably very important and intuition suggests that they should play a key role in generating utility-based rankings of publications in a constantly evolving body of research.

3.1 Data

The data used in our experiments is the paper citation network of the ACL Anthology Network (AAN) [23] 2009 release[1]. The AAN contains 14,912 papers published in various ACL venues covering topics in the field of Natural Language Processing (NLP). 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 (converted from the PDF file) along with metadata including authors names, paper title and year of publication, is also provided with the corpus.

4 Our Approach

Given the recent relative success of applying PageRank to scientific document ranking [25, 17, 24, 15], we propose to use PageRank as a starting point for our link-based ranking framework. But as argued in Section 3, some modifications are required to adapt PageRank and properly deal with the ageing factor and the intricate topic dynamics in a citation network in order to make the resulting ranking more useful. PageRank can be thought of as a model of Web users' collective behaviour in terms of a "random surfer" who randomly follows hyperlinks with a fixed probability α at each webpage to stop link following and "teleport" to a random page [5]. The teleportation rate α is usually set to 0.15 for ranking webpages, which is roughly equivalent to randomly following 5 to 6 $(1/\alpha - 1)$ hyperlinks on average before teleportation takes place [7, 20]. Former research (e.g., [2, 4]) has suggested that PageRank's behaviour is governed by α. The success of PageRank is thus partly determined by how closely its random surfer model captures the browsing behaviour of real users in the form of α [27, 12].

In order to adapt PageRank and rank scientific documents in terms of their potential usefulness, the underlying random surfer model should be tailored to better model a "random literature explorer", who we argue is more likely to stop link following earlier than a random Web surfer, based on the couple of observations that follow. Firstly, the topical focus of a literature explorer (researcher) is all the more narrower that her exploration of a paper citation network is more targeted at only those publications that have a high degree of relevance. When she encounters a reference to a paper on topics which are not very relevant to her own fields of study, she is more likely to stop her inquiry and start anew from another paper that is potentially more useful to her. We suspect that the high intellectual cost of reading a scientific document crucially contributes to this phenomenon. Secondly, compared to Web surfers, scientific literature explorers are equally, if not more, interested in the latest information. Due to the strong ageing characteristic of a scientific citation network, reference following will quickly lead a researcher to older literature where

[1]From here onwards and unless stated otherwise, we refer to the AAN 2009 release simply as AAN.

the information is likely to be outdated and thus is less useful, resulting in earlier link following termination. In other words, *information freshness decay* is likely to be another major contributing factor for when reference following terminates. It is crucial to note that our notion of "information freshness" is defined along both the temporal and the topical dimensions of documents so as to capture a subtle feature of literature development, namely, that different topical areas evolve at vastly varying paces. We will discuss the definition of "information freshness" in detail in Section 4.3.

In the following subsections, we seek potential empirical evidence that confirms the previous observations and we discuss how to make our ranking framework contents and time sensitive.

4.1 Topic Modelling

To model topics in the AAN, we used the Latent Dirichlet Allocation [1], a highly successful unsupervised topic modelling approach. Essentially, LDA is a generative probabilistic model that represents each document in a corpus as a random mixture of latent topics, each topic being characterised as a multinomial distribution over words. The generative process posits that each document is generated by first choosing a distribution over topics and then choosing each word from a topic selected according to this distribution. After discovering the optimal set of latent topics that are most likely to have generated the corpus, each document can be represented as a mixture of topics according to the proportion of its contents (words) assigned to each topic. Intuitively, the proportion of contents covered by a given constituent topic of a given document captures how likely the document is about the corresponding topic. In later sections, we refer to this likelihood as the *topic probability mass*.

Following [14] and [15], we ran LDA with 100 latent topics over the full texts of papers in the AAN. A manual inspection on the generated topics was then carried out to assign them names. The assignment of topic name was based on inspecting the most frequent topical words in each generated topic, as well as examining the titles and/or abstracts of papers whose top topic (the topic with the highest topic probability mass in a paper's topic mixture) is the target topic. For the sake of simplicity, we did not follow [14] and invoke a second run of LDA using topics derived in the first pass as a Dirichlet prior. Albeit simple, we found that our method induced a similar set of topics as those reported in [14]. We found 40 out of the 100 topics to be relevant and excluded the irrelevant ones from subsequent analysis. The top ten most frequent words for several of the relevant topics found by LDA are shown in Table 4.1, along with the names assigned through manual inspection.

Each latent topic discovered by LDA using the full texts of papers in the AAN can be seen as a subfield of NLP. Though there can be definitional discrepancies between automatically discovered topics and "traditionally" defined subfields, it can be argued that subfields automatically 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 [13, 19]. In this section, we use the term topic and subfield interchangeably.

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 Extraction	patterns terms term semantic relation extraction relations information pattern based
Information Retrieval	word words query web retrieval based pages using terms corpus
Language Resource	language text information knowledge data research systems resources linguistic database
Lexical Grammar	lexical feature default type lexicon value hpsg head inheritance information
Lexical Semantics	semantic object case noun verb example relations meaning knowledge relation
Statistical MT	alignment model translation word phrase english based source models words
Meta Learning	learning data set words using based used training word number
Metaphor	concept transfer metaphor language concepts knowledge metaphors meaning target source
Morphology	morphological stem forms morphology form word rules words root stems
Phonology	word words language speech vowel phoneme phonetic languages syllable stress
Plan-based Dialogue	dialogue user plan information utterance action act speaker task model
PP Attachment	verb lexical noun preposition corpus syntactic verbs pp attachment prepositions
Probability Models	model models word language probability words data probabilities used speech
Question Answering	question answer questions trec answers answering query qa retrieval information
Semantic Roles	semantic task argument role predicate features data information arguments entity
Sentiment	annotation negative positive opinion annotations sentiment polarity subjective annotators annotator
Speech Recognition	speech data dialog recognition spoken user words utterances corpus utterance
Summarisation	speech data dialog recognition spoken user words utterances corpus using
Tagging/Chunking	word words tag pos corpus tagging tags tagger based training
Tutoring Systems	students student null human al tutor learning et tutoring feedback
Word Sense Disambiguation	sense word senses words disambiguation chinese wsd lexical corpus collocation
WordNet/Ontology	wordnet synset sense words corpus lexical relations senses semantic word

Table 4.1: Top 10 words for several topics induced using LDA

Empirically, subfields of NLP are closely related due to the fact that many NLP applications consist of a pipeline with higher level techniques built on top of lower level processes. For example, Part-of-Speech Tagging is usually

required as a preprocessing step for many higher level NLP tasks such as Parsing and Machine Translation. As a result, we expect papers in the AAN to be more topically diverse than papers in a domain with fewer interactions among its subfields. To quantify the *topical diversity* of papers in the AAN, we propose to use the *topic entropy* of the topic distribution of papers, a measure adapted from [14]. Let d be a paper with a topic mixture T consisting of k constituent topics $t_1, \ldots, t_k$, and for all strictly positive $i \leq k$, let $P(t_i|d)$ denote the topic probability mass of t_i. Then the topic entropy of d is defined as:

$$H(T|d) = -\sum_{i=1}^{k} P(t_i|d) \cdot \log\big(P(t_i|d)\big) \tag{4.1}$$

Intuitively, papers with a more evenly distributed topic mixture (e.g., interdisciplinary papers) have a higher topic entropy, while a lower topic entropy is expected for papers that are more focused on a small number of topics. Figure 4.1 shows the topic entropy distribution of papers in the AAN, which confirms the intuition that the great majority of papers are topically diverse, with a mean topic entropy of 2.34. There are only a small number of papers that are very topically focused (they form the first peak in the kernel density estimation curve). At the other extreme, there are also papers representing a diverse topic mixture reaching a topic entropy over 4. Table 4.2 shows some examples of highly topically focused papers, as well as extremely topically diverse papers in the AAN.

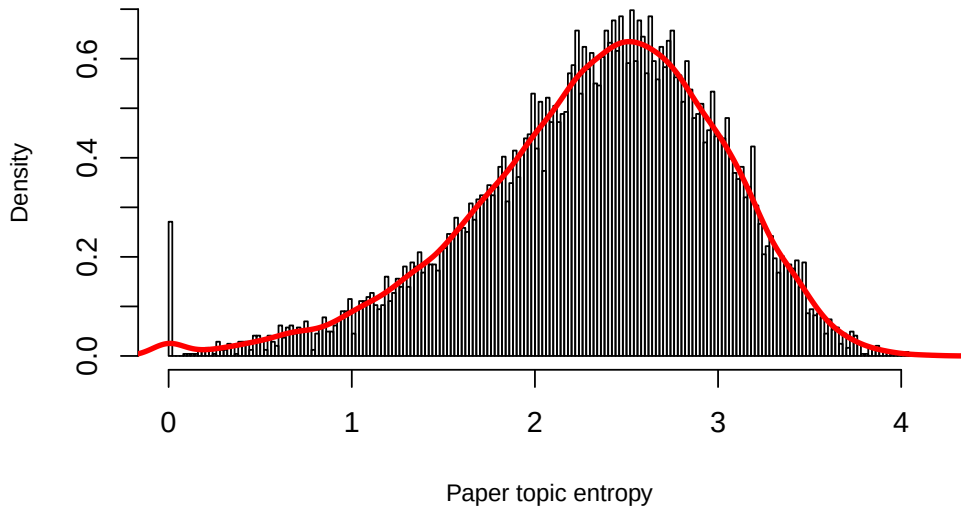

Figure 4.1: Histogram of topic entropy distribution of papers in the AAN and kernel density estimation

Paper ID	Title	Topic entropy
P04-3013	Exploiting Unannotated Corpora For Tagging And Chunking	0.01
J03-1002	A Systematic Comparison of Various Statistical Alignment Models	0.03
L08-1543	A Hybrid Morphology-Based POS Tagger for Persian	0.01
P06-1138	Identifying Foreign Person Names in Chinese Text	0.12
H94-1049	A Report Of Recent Progress In Transformation-Based Error-Driven Learning	0.18
W04-2801	Robustness versus Fidelity in Natural Language Understanding	4.15
W06-0802	Hybrid Systems For Information Extraction And Question Answering	4.04
W90-0106	Natural Discourse Hypothesis Engine	3.95
W07-1705	Automatic Processing of Diabetic Patients Hospital Documentation	3.89
C04-1124	Detecting Multiword Verbs in the English Sublanguage of MEDLINE Abstracts	3.87

Table 4.2: Examples of highly topically focused/diverse papers

Expectedly, we found that many papers of low topic entropy focus on low-level NLP tasks such as tagging and chunking, while papers with high topic entropy tend to be very application-oriented, describing high-level NLP appli-

cations. Interestingly, we found that papers published in the Language Resources and Evaluation Conference (LREC) are the most topically focused, while papers published in various workshops are the most topically diverse.

Given the high topical diversity of papers in the AAN, we were interested to further investigate and find out how many top topics are enough to, on average, collectively capture most of the contents of papers in the AAN. We rank topics in a paper's topic mixture according to their probability mass in descending order. Figure 4.2 shows the distribution of topic probability mass over the top 4 topics of papers using box and whiskers plot (means are marked out using red dots). It can be seen that on an average, the top topic of a paper captures a large percentage of 38% of the paper's topical contents, while the coverage drops significantly to below 20% for the secondary topic. The average topical coverage further drops to 10% and 7% for topics ranked 3rd and 4th, respectively. It should be noted that outliers with a much higher than average topical coverage can be found in each rank. Figure 4.3 shows the cumulative topic coverage when progressively including lower ranked topics. A clear diminishing gain is observed with the top 4 topics collectively reaching a coverage of 80% on average for papers in the AAN. These observations aid to decide what should be the primary topics of a paper; we will elaborate on this in Section 4.3.

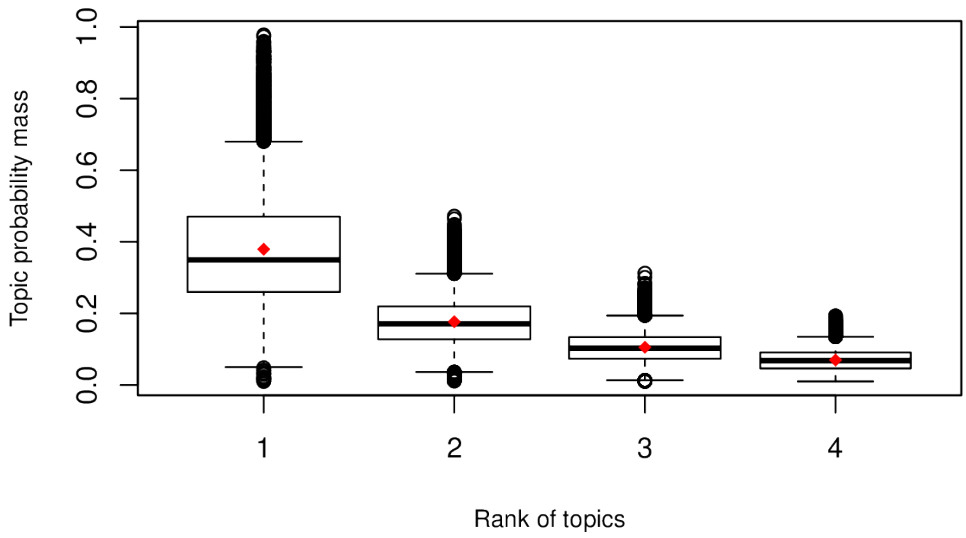

Figure 4.2: Topic coverage of the top 4 topics of papers in the AAN

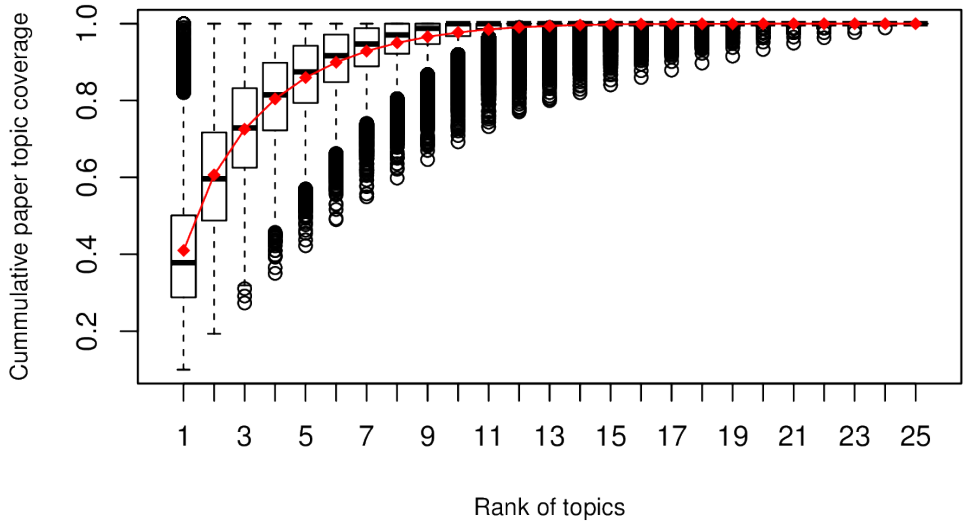

Figure 4.3: Cumulative topic coverage of the top ranked topics of papers in the AAN

4.2 Topical Drifts and Damping Factor Selection

As mentioned at the beginning of Section 4, we feel that topic drifts along citation paths play a main role in reference following termination for human readers. Empirically, we observe that the topical relevance of referenced papers to the original citing paper drops rapidly along the citation path, and we are interested in estimating the rate at which topical relevance drops for papers in the AAN. To measure topical drifts along reference paths, we first calculate the shortest paths between all pairs of papers in the directed paper citation network using breadth first search. We then calculate the *topical similarities* as the cosine of the angle between the topic vectors of papers in the latent topic space. The distribution of topical similarity scores between papers of up to 4 citations away from each other is shown in Figure 4.4.

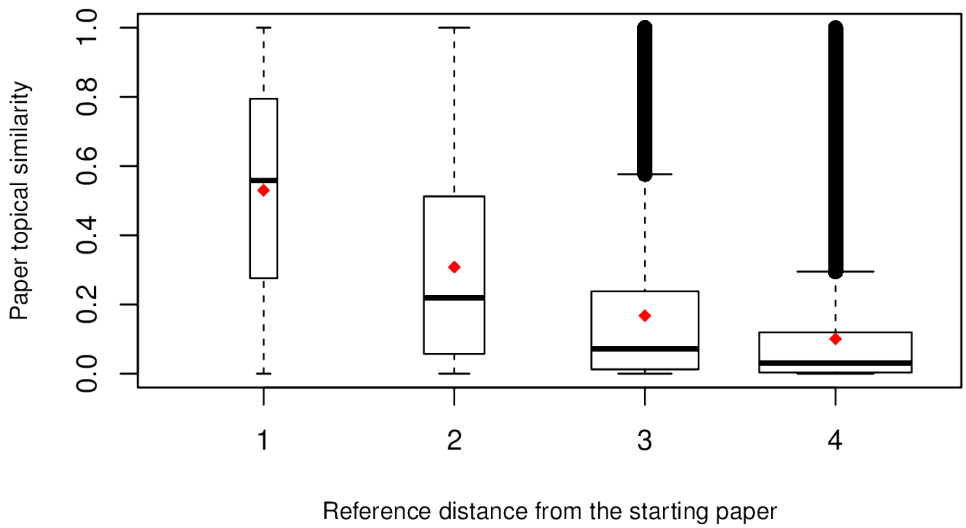

Figure 4.4: Paper topical similarity along reference paths

Figure 4.4 shows that overall, the topical similarity between the starting paper and papers reached via reference following drops as the length of the reference path being followed increases. This suggests that generally, papers cited along a reference path become increasingly less topically relevant to the starting paper. It can be seen that the topical similarity drops fairly fast along reference paths. After following only 3 references from a given starting paper, the mean topical similarity between the paper that has been reached and the starting paper drops to below 20%, and the median drops even further to below 10%. Based on our insights into the domain, we feel that this topical drift is practically significant enough not to let a literature explorer read papers which are more than 2 citations away from the starting paper. In other words, the average reference following length d is estimated to be 2 before teleportation takes place. Admittedly, this general topical relevance threshold does not necessarily apply to other domains, or even to individual readers in NLP. Nevertheless, the threshold obtained from our case study for the AAN corresponds to a reasonable scenario. Based on this rationale, we hypothesise that a teleportation probability of $1/3$ (in general, $1/(1+d)$) at each point in a chain of citations should better reflect the link following behaviour of human literature explorers. This translates to a damping factor of 0.667. In Section 5, we demonstrate that this value of the damping factor, empirically derived from topical drifts in the analysed data, indeed yields better ranking results than the more traditionally used damping factor of 0.85.

It should be pointed out that we chose the damping factor based on our hypothesis that topical drift is a main cause for reference following termination in a content-driven manner. This presents a new angle for damping factor selection alongside expensive browser log tracking (e.g., [12]) and exhaustive optimisation using some application-dependent objective function (e.g., [25]). The value chosen here does not necessarily yield the best performance, but it is intuitively justified. We leave the fine tuning of this parameter in a fully data-driven context for future work.

4.3 Topical Longevity and Information Freshness Decay

We mentioned at the beginning of Section 4 that inexorably visiting older literature via reference following, that is, information freshness decay, is another main cause why scientific document readers terminate link following. We aim at making our random literature explorer model sensitive to the information freshness decay to more accurately reflect the behaviour of human readers. One challenge is that no proper measure exists to gauge the freshness of the information contained in a publication. Many studies (e.g., [29, 25, 17]) have attempted to use a single decay

factor based on the assumption that the information freshness of papers published in the same year drops equally. Consequently, a simple exponential temporal penalty calculated from the age of papers and an empirically estimated characteristic decay has been used to penalise older papers. We argue that using a single characteristic decay factor is overly simple and ignores the fact that various subfields of a domain exhibit various paces of progress. Intuitively, a paper covering slow-moving subfields should be penalised less harshly than a paper published in the same year on some fast advancing topics. This calls for a *contents sensitive, per paper* decay factor, designed to more accurately measure the freshness of information contained in each publication.

In this section, we propose a novel metrics of information freshness based on the *topical longevity* of a paper's *primary topics* in its topic mixture. Firstly in Section 4.3, we describe how papers are assigned to their primary topics as a preliminary step for subsequent discussion. Section 4.3 defines the topical longevity for a subfield and its calculation. Finally in Section 4.3 we present our metrics of information freshness of a paper and discuss how it is modelled as a decay penalty to the damping factor estimated based on topic drift.

Deciding the Primary Topics of a Paper

In order to calculate the topical longevity of a subfield and further measure papers' information freshness, we assign each paper in the AAN to one or more subfields based on their topic mixture found by the LDA. Some former studies (e.g., [19]) employed a simple strategy where a paper is assigned to all topics in its topic mixture that have a topical probability mass above some threshold (10% was used in [19]). While this simple method may work well for papers with a relatively balanced mixture of topics, it may be overly nondiscriminatory for papers with strong topical foci. For example, consider a paper with a topic mixture of topic 1: 77%, topic 2: 12%, topic 3: 11%. Arguably, topic 1 should be chosen as the primary topic, while topics 2 and 3 are comparatively "secondary". It can be seen that simple threshold-based primary topics selection methods disregard imbalanced topic mixtures. To address this issue, we take into consideration the "shape" of topic mixture using the following 2-step approach.

Step 1 : By default, the top 4 topics of a paper are chosen as its primary topics.[1] This decision is supported by the findings presented in Section 4.1 that on average, the top 4 primary topics ensure a reasonable topical contents coverage of over 80% for papers in the AAN.

Step 2 : Eliminate "secondary" topics from a paper's set of primary topics if the paper is "sufficiently topically focused" or "overly topically diverse" that no primary topics can be decided. From the paper's top topic, we check whether its probability mass is a mild outlier in the distribution of content coverage for topics of that rank in the AAN. As we consider mild outliers only, the upper outlier limit is not computed as it is traditionally, *i.e.*, $upper\_quartile + 1.5 * interquartile\_range$ (Figures 4.3 and 4.4), but as $upper\_quartile + interquartile\_range$ (Figure 4.2). If yes, then all lower ranking topics are eliminated from the set of primary topics for the paper. If no, then topics ranked 2nd and 3rd are checked iteratively in the same fashion.

We found that our simple outlier-based heuristic works well to find primary topics discriminatively, based on the extent to which the paper topic mixture is balanced. The average number of primary topics per paper in the AAN is 2.3.

Topical Longevity of a subfield

We use the *cited half-life* [8], a measure of the median citation age within a collection of publications assigned to a given subfield of NLP, as an indicator of a paper's topical longevity. Following [19], we calculate the half-life of a subfield using Brooke's estimator as originally proposed in [6]. This involves estimating a *constant ageing rate*, based on citation counts to papers in the subfield. First, a benchmark number of years i (6 in our calculation) is chosen. Then the number k of citations to papers in the chosen subfield t published at least i years ago (we use the end of our AAN data collection, 2009 as the census year), is calculated. Let ℓ be the number of citations to papers in subfield t published less than i years ago. The constant ageing rate a is calculated as follows:[2]

$$a = \sqrt[i]{\frac{k}{k+\ell}} \tag{4.2}$$

The topical longevity of subfield t is then:

$$\mathrm{TL}(t) = -\frac{\log(2)}{\log(a)} \tag{4.3}$$

We refer the reader to [9] for details on the derivation. Intuitively, a subfield with a short topical longevity is likely to evolve at a fast pace, with new papers quickly superseding older publications. On the other hand, a subfield with a long topical longevity is expected to be all the more static that old papers are still highly relevant and are frequently referred to.

[1] If a paper has fewer than 4 constituent topics, then all topics are selected initially.

[2] Additive smoothing has been carried out for both k and ℓ to handle potentially very new ($k = 0$) and dead ($\ell = 0$) subfields.

We have found a large span of topical longevity scores for subfields in the AAN with a median of 17.2 years and an inflated mean of 32.1 years due to several severe outliers. In Table 4.3, we list some exemplar subfields in NLP of both short and long topical longevities. It is worth pointing out that NLP subfields and their topical longevities are calculated based on the AAN data only. Results are inevitably confounded by corpus specific topical patterns, thus do not necessarily depict an accurate landscape of the entire NLP domain.

subfield	**Topical Longevity**
Dependency Parsing	3.5
Sentiment Analysis	5.3
Machine Translation	7.7
Text Summarisation	8.4
Information Retrieval	8.5
Computational Semantics	86.6
Lexical Grammar	85.2
Discourse Relations	54.2
Lexical Semantics	31.4
Tagging/Chunking	22.3

Table 4.3: Examples of NLP subfields discovered in the AAN of short/long topical longevities

Manual inspection shows that topical longevity scores correlate well with empirical recent research trends in the NLP field. A trendy subfield such as Dependency Parsing has a very short topical longevity of 3.5 years, meaning that 50% of all citations to papers in this subfield are made towards papers published within 3.5 years prior to the census year of 2009. Subfields that are more theoretical, such as Computational Semantics, have a much longer topical longevity. We have also identified some severe outliers with extremely large scores. For example, a topical longevity of 176.4 years has been calculated for Message Understanding. A careful inspection reveals that this topic is a tightly knitted niche field attracting much fewer citations after the MUC conference ceased to exist in 1997, which explains its unrealistic topical longevity. The fact that the AAN excludes papers published outside its indexed venues further aggravates the data sparsity issue. Indeed, the topical longevity scores are only a rough estimation under the constant ageing rate assumption based on partial data, which might deviate significantly from reality. We still feel that the estimated topical longevity is a reasonable relative measure of the progressing pace of a subfield within the AAN, which can be further used to measure information freshness of publications.

Per Document Information Freshness Decay Factor

From the considerations developed in the previous sections, it is clear that the information freshness measure of a publication should be dependent on its contents. More specifically, papers covering slow-moving subfields age more slowly than papers on fast-moving topics. We propose to calculate the *information freshness decay* of a paper d using its set T of primary topics as follows where for any primary topic t, $P(t)$ is the topic probability mass of t:

$$\mathrm{IFD}(d) = \sum_{t \in T} \mathrm{TL}(t) \cdot \frac{P(t)}{\sum_{t' \in T} P(t')} \tag{4.4}$$

The second factor in the equation above normalises a paper's topic vector to have unit ℓ_1-norm. In other words, we define the information freshness decay factor of a paper as the weighted sum of topical longevity scores of its primary topics according to its normalised topic mixture. We use a paper's primary topics only for this calculation, based on the assumption that papers are predominantly cited for the contents on their primary topics. Similar to [25] and [17], we use the following exponential *information freshness decay penalty* for a paper d:

$$\mathrm{IFDP}(d) = \beta^{-\mathrm{age}(d)/\mathrm{IFD}(d)} \tag{4.5}$$

where β is a corpus dependent parameter that could be optimised against an objective function. In this paper however, we set $\beta = \mathrm{e}$ as a pilot study.

We feel that topical relevance drift, discussed in Section 4.2, plays a more important role in reference following termination than information freshness decay, because a reader will notice the drop in topical focus relatively easily, regardless of her level of expertise in her fields of study. Information freshness judgement, however, is more dependent on the level of knowledge of the literature explorer. For example, papers that experts perceive as containing nothing but "standard" field knowledge can still be of interest to beginners. We thus propose to estimate a *global* damping factor based on topical drift, and to encapsulate information freshness decay as a weight to the global damping factor and turn it into a *local* damping factor. Consequently, we have the following *per document damping factor*

that combines topical drift and information freshness decay in modelling a random literature surfer's link following termination:

$$\mathrm{DF}(d) = \mathit{global\_df} * \mathrm{IFDP}(d) = 2/3 * e^{-age(d)/\mathrm{IFD}(d)} \tag{4.6}$$

Consequently, we calculate each node's score in a power iteration as follows to account for local damping factors:

$$Pr(d_i) = \sum_{j=1}^{n} \Big(pt(d_i) \cdot \big(1 - \mathrm{DF}(d_j)\big) + A_{ji} \cdot \mathrm{DF}(d_j) \Big) Pr(d_j) \tag{4.7}$$

where A is the transition matrix derived from the citation network. With $O(d_j)$ denoting the out-degree of paper d_j in the citation network, cell A_{ji} of matrix A is defined as:

$$A_{ji} = \begin{cases} 1/O(d_j) & \text{if } d_j \text{ cites } d_i \\ 0 & \text{otherwise} \end{cases}$$

The term $pt(d_i)$ in (4.7) is the personalised teleportation probability for document d_i. In this paper, we use a simple uniform teleportation vector, making our random literature explorer teleport equally likely to all papers in the AAN (i.e., for all i in $\{1, \ldots, n\}$, $pt(d_i) = 1/n$).

To summarise, our ranking framework adapted from PageRank models a random literature surfer having the following characteristics: starting from some random paper in a citation network, it follows a reference path of average length 2. Upon encountering a paper (either through reference following or teleportation), it decides whether it should keep on reading the paper's references taking into consideration both contents and age of the paper. If the paper is young and/or it primarily covers fairly static subfields (indicative of the paper's usefulness), the random literature explorer is more likely to keep following its references. On the other hand, if the paper was published a long time ago and/or it is mainly on fast-moving topics (suggesting that the paper is of less utility), the random reader is prone to cease reference following and restart its literature exploration from another randomly selected paper. Perceived as a document ranker, our ranking framework propagates rank mass discriminatively in a contents and time sensitive manner. A paper's contents is assumed to better reflect the current lines of research if it contains fresh information and its implicit endorsement of utility conveyed to referenced work is recognised. On the other hand, a paper's implicit endorsement of usefulness towards its reference work is discounted if it is likely to contain outdated information. Finally, viewed from a global perspective, our system effectively achieves a discriminative soft pruning of a citation network that constrains the otherwise unchecked flow of rank mass into outdated portions of a scientific literature towards achieving the goal of making the ranking more pertaining to documents' practical scientific utility rather than their historical importance. We name our ranking framework RALEX for **RA**ndom **L**iterature **EX**plorer.

5 Results and Discussion

5.1 A Proxy of Ground Truth

Without a notion of ground truth, evaluating ranking results on intrinsic measures such as the potential scientific utility of academic publications is challenging [25, 15]. [15] further showed that the high level of subjectiveness in human-based evaluations leads to an inter-judge agreement so low that it disqualifies such evaluations as a reasonable surrogate. It is clear that some metrics, preferably derived directly from the data in an objective manner, are more suitable to be utilised to generate a *proxy gold standard* of relative usefulness of academic publications.

Qualitative citation analysis strives to measure the quality of scientific papers by not only considering the quantity, but also the *functions* (e.g., comparing results, acknowledging sources of claims) of citations they receive [21]. For example, a citation to a paper expressing criticism would carry lower (or even negative) weight in the qualitative measure of the cited paper [3]. Though an agreement has not yet been reached on a standard set of citation functions, it has been found that including references to papers with little intellectual contribution to the citing work in a "perfunctory" manner is prevalent in scientific documents [28]. Compared with non-perfunctory, or "functional" citations, perfunctory citations serve a less significant role in relaying crucial information in a scientific literature as the omission of a perfunctorily cited work would not hinder the understanding of the citing document. Intuitively, the endorsement of usefulness carried in a perfunctory citation towards the cited work should thus be discounted. More specifically, we feel that the *functional citation count* of a paper reveals how it has been collectively perceived as a useful and often *indispensable* piece of work in a scientific field by its peers, and thus is an accurate measure of its scientific utility. Based on this rationale, we propose to generate a proxy gold standard based on paper's functional citation counts.

We first use our automatic citation function classifier described in [28] to categorise all citations in the AAN into either functional or perfunctory and each paper's functional citation count is calculated. Figure 5.1 shows the distribution. As expected, the great majority of papers have very few functional citations. A maximum likelihood power-law fitting reveals that this distribution exhibits a power-law (xmin = 6, $\alpha = 2.48$), which is shown in Figure 5.2. Unsurprisingly, the functional citation counts and raw citation counts of papers in the AAN are positively correlated with a high Spearman's ρ ranking correlation coefficient of 0.83 ($p < 0.0001$). However, a less strong Kendall's τ

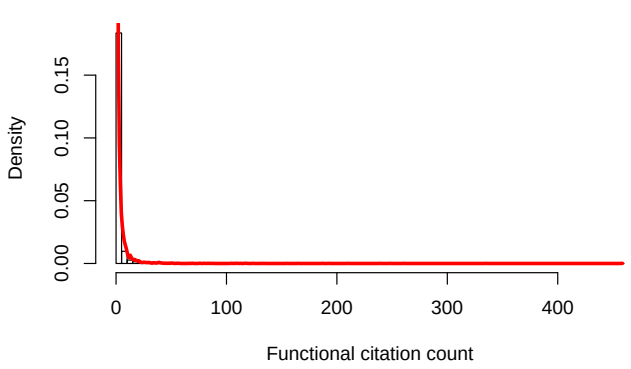

Figure 5.1: Kernel density estimation of number of functional citation counts

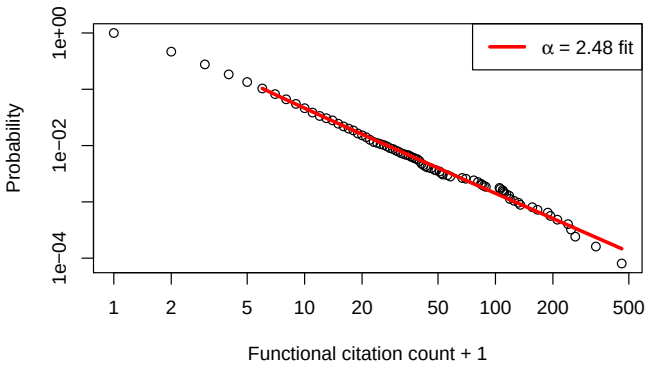

Figure 5.2: Power-law fitted using maximum likelihood estimation

ranking correlation coefficient $(0.75, p < 0.0001)$ shows that there are many discordant pairs in the rankings (i.e., pairs of papers whose relative ranking orders are changed from one ranking to the other), highlighting the differences between a quantitative and a usefulness-based ranking[1]. A proxy gold standard ranking of papers is subsequently generated according to their functional citation counts in reverse order. The first 3 columns in Table 5.1 show the ids, titles and functional citation counts along with their raw citation counts (in parenthesis) of the top 10 papers, respectively, in the proxy gold standard.

It should be pointed out that there is no existing proxy ground truth of papers' scientific utility; it is thus necessary to generate one to properly evaluate RALEX's performance. Admittedly, the classification of citation links into functional and perfunctory is not perfect (F1 score 0.9), especially for references with missing citation contexts (F1 score 0.74)[2]. However, one has to rely on some automatic classifier to make large scale qualitative citation analysis possible. We feel that our classifier's reasonable performance warrants the validity of our proxy gold standard, especially in lack of better options. The performance of citation classification (and thus the quality of the proxy gold standard) can surely be further improved but it is not the focus of the current work.

ID	**Title**	**Functional (Raw) citation count**	**RALEX**	**PR** (df = 0.667)	**PR** (df=0.85)
J93-2004	Building A Large Annotated Corpus Of English: The Penn Treebank	458 (637)	1	5	6
J93-2003	The Mathematics Of Statistical Machine Translation: Parameter Estimation	335 (490)	4	7	8
P02-1040	Bleu: A Method For Automatic Evaluation Of Machine Translation	261 (353)	7	21	47
J03-1002	A Systematic Comparison Of Various Statistical Alignment Models	248 (304)	16	58	112
P03-1021	Minimum Error Rate Training In Statistical Machine Translation	239 (292)	22	59	108
J86-3001	Attention, Intentions, And The Structure Of Discourse	210 (332)	3	6	7
J96-1002	A Maximum Entropy Approach To Natural Language Processing	193 (295)	6	8	16
A00-2018	A Maximum-Entropy-Inspired Parser	187 (290)	10	29	51
N03-1017	Statistical Phrase-Based Translation	165 (291)	25	67	118
W96-0213	A Maximum Entropy Model For Part-Of-Speech Tagging	155 (214)	9	15	29

Table 5.1: Top 10 papers ranked using functional citation count and their corresponding ranks using PageRank-based methods

[1]The two non-parametric correlation tests had been chosen over Pearson's r due to the fact that both quantities involved in the correlation test demonstrate power-law's, which severely violates normality.

[2]Roughly 1/3 of all citations in the AAN have missing citation context (i.e., the citing sentence). Our classification system had to rely on a set of non-textual, suboptimal features to classify such citations. We refer the reader to [28] for details.

5.2 Results

We generated a ranking of papers in the AAN using RALEX and a PageRank baseline ranking using a damping factor of 0.85. To inspect the effectiveness of using the contents and time sensitive weight factor (the second component in Equation (4.7)), we generated another baseline ranking using PageRank with a damping factor of 0.667 estimated based on topical drifts. We used the same convergence criteria of 10^{-8} ℓ_1-error for all 3 rankers. As mentioned in Section 1, rankers usually function as an integral part of an information retrieval system in a ranked retrieval context. Aiming at obtaining the most comprehensive picture possible, we evaluated the ranking results from both a ranked retrieval perspective and a pure ranking perspective. It is worth pointing out that both evaluations differ subtly in their main aim: the former focuses on evaluating a ranker's ability to rank the best papers (according to the proxy gold standard) at the top of its ranking list, while the latter is more concerned on how closely the produced ranking list mirrors that of the proxy gold standard. We separately provide two sets of evaluation results in the following subsections.

Evaluation as Ranked Retrieval

Following [24], we used the top 50 ranked papers in the proxy gold standard as the true result set and evaluated RALEX's performance in a ranked retrieval context. The Precision-Recall curves of RALEX and the 2 baseline rankers are shown in Figure 5.3. It can be seen that RALEX consistently outperformed the 2 PageRank baselines in its ability to retrieve the best papers according to the proxy gold standard. Notably, RALEX's superior performance is especially pronounced at the top of the produced ranking lists. To further inspect the rankers' relative performance at the top of their rankings, we selected the top 50 papers in the produced 3 rankings as the corresponding ranker's return set while still using the top 50 papers in the proxy gold standard as the target retrieval set to generate the Prevision-Recall curves in Figure 5.4. With this more detailed visualisation, we can clearly see that RALEX beat the 2 baseline rankers by a considerable margin in its ranked retrieval performance. These results suggest that RALEX is likely to perform better practically and return a more useful top list of papers to users.

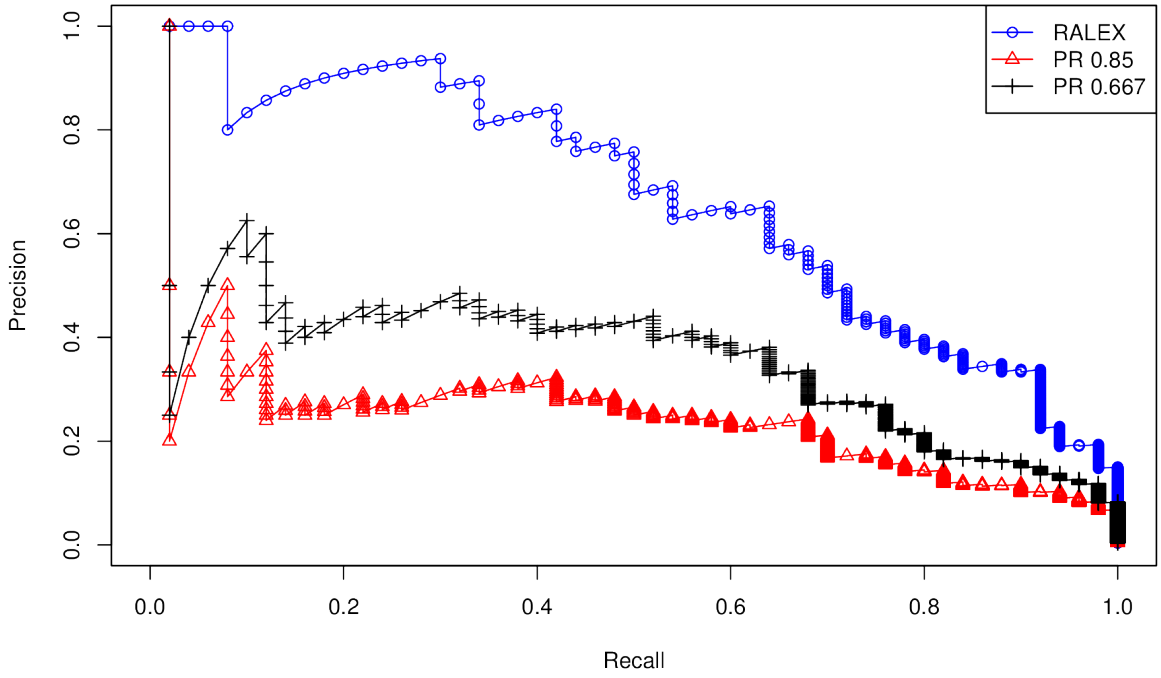

Figure 5.3: Precision-Recall curves using top 50 ranked papers in the proxy gold standard as the target retrieval set

To provide a more readily interpretable numerical summarisation of the Precision-Recall curves, we also report the Average Precisions (AP) at different top-k cutoff points achieved by the 3 rankers. The AP is a popular metric of ranked retrieval results that considers the order in which the returned documents are presented. Let k be the length of the top portion of a ranked list of documents deemed as the retrieval set. For all nonzero $i \leq k$, let $P(i)$ denote the precision at cutoff i in the ranked list of documents, and let $\Delta R(i)$ be the change in recall between cutoff $i-1$ and cutoff i. The AP at top k of a ranking is defined as:

$$AP(k) = \sum_{i=1}^{k} P(i) \cdot \Delta R(i) \tag{5.1}$$

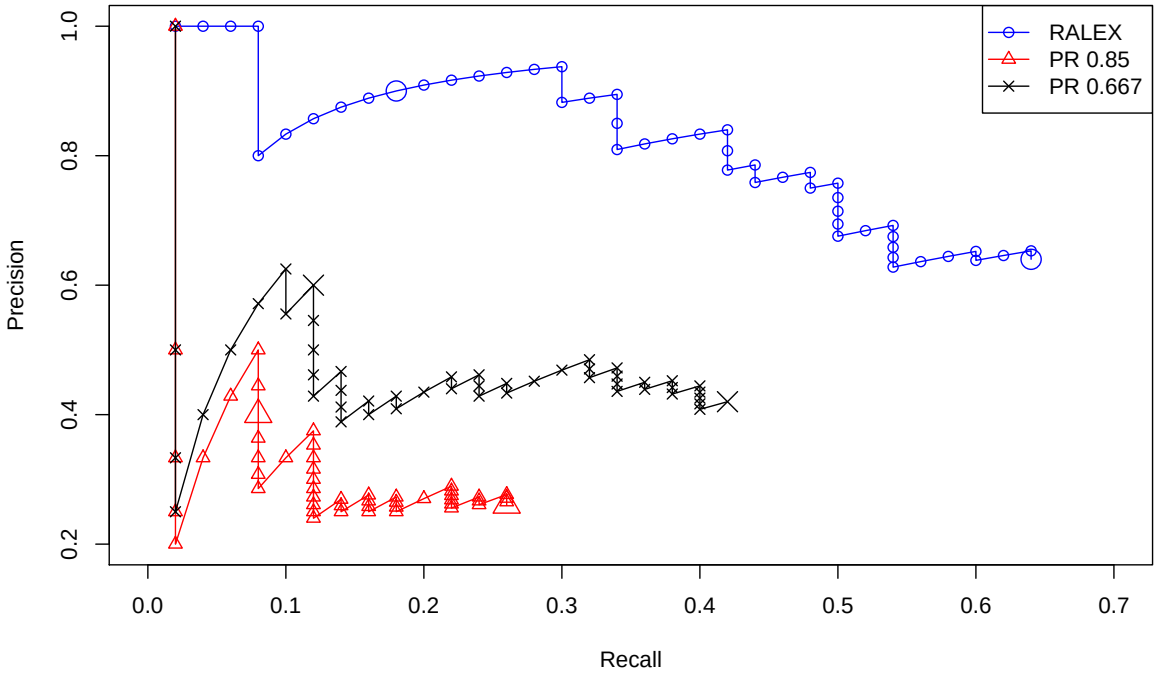

Figure 5.4: Partial Precision-Recall curves using top 50 ranked papers in the proxy gold standard as the target retrieval set. Performance of the 3 rankers at top 10 and top 50 of their respective rankings are highlighted using larger symbols

Intuitively, the AP metric incurs a penalty to errors made towards the top of a ranking and it is effectively a close approximation of the area under the Precision-Recall curve. Table 5.2 summarises the AP of the 3 rankers at various top-k cutoff point in their rankings. Again, the top 50 ranked papers in the proxy gold standard were used as the target retrieval set.

k	RALEX	PR (df=0.667)	PR (df=0.85)
10	0.84	0.34	0.21
50	0.53	0.21	0.10
100	0.60	0.31	0.17
200	0.65	0.34	0.22
500	0.67	0.37	0.25
1000	0.67	0.37	0.25
2000	0.67	0.37	0.25

Table 5.2: Average Precision at top-k ranking using the top 50 ranked papers in the proxy gold standard as the target retrieval set

The results so far have shown that RALEX compares favourably to the 2 PageRank baseline rankers in retrieving the most useful papers especially at the top of its ranking. To further test RALEX's ability in producing a ranking list similar to the proxy gold standard, we perform more evaluations in a pure ranking context in the following subsection.

Evaluation as Ranking

To measure the similarity between two rankings, we need a suitable distance metric. Following [17], we used the normalised Spearman's footrule distance [16] to measure the ranking distance of the top-k elements. Let R_1 and R_2 be two rankings of n papers and let $R_1(d_i)$, $R_2(d_i)$ be the ranking positions of paper d_i in R_1 and R_2, respectively. The top-k ranking distance between R_1 and R_2 is calculated using the following formula:

$$D(R_1, R_2) = \frac{\sum_{i=1}^{k} |R_1(d_i) - R_2(d_i)|}{k * n} \tag{5.2}$$

The top-k ($k \leq 1000$) ranking distances between each of the 3 rankings to the proxy gold standard is shown in

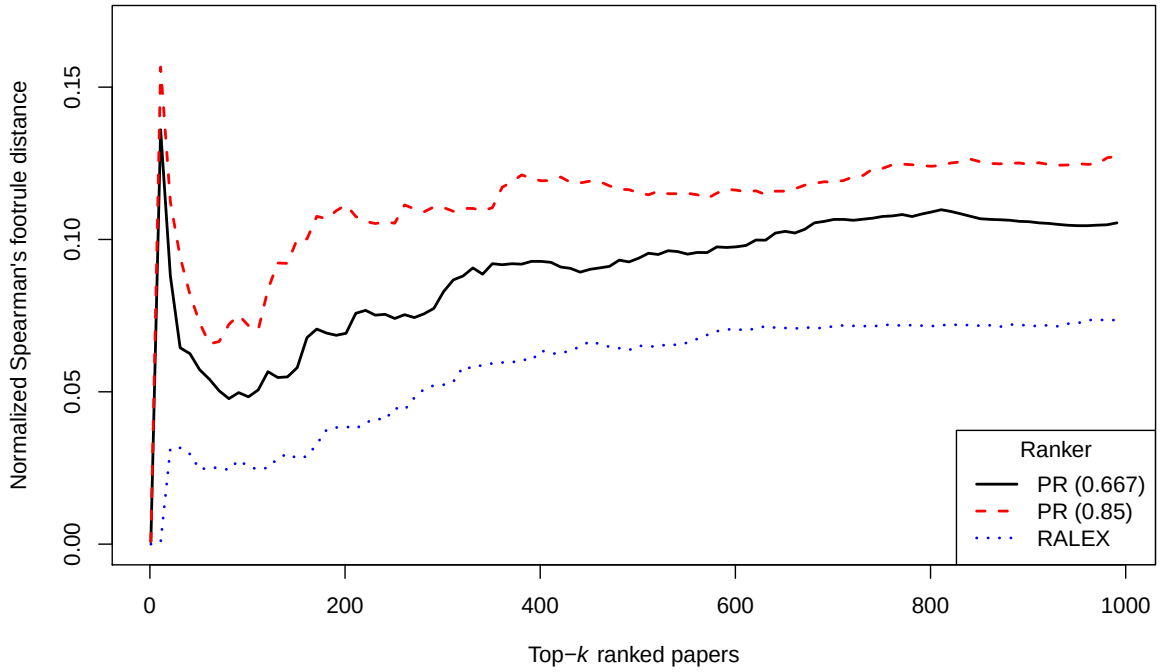

Figure 5.5: Spearman's footrule top-k ranking distance of papers in the AAN using the functional citation count proxy gold standard

Figure 5.5. It can be seen that RALEX consistently outperformed the 2 PageRank baselines in the top-k ($k \leq 1000$)[3] ranking results evaluated using the proxy gold standard, confirming RALEX's ability in producing better rankings that are more reflective of scientific utility. Consistent with the ranked retrieval evaluation results, the top section of the ranking list RALEX produced is much closer to the proxy gold standard than those generated using the 2 baselines, which again indicates RALEX's practical advantage.

Discussion

In the previous subsections, we have evaluated RALEX's ranking results both as ranked retrieval and pure ranking using appropriate metrics. All results confirmed RALEX's overall effectiveness in producing better rankings that are more reflective of papers' scientific utility. It is also interesting to note that the PageRank baseline using a damping factor of 0.667 consistently produced better rankings than that using a damping factor of 0.85, suggesting that the new base damping factor estimated using topical drifts works better than the traditionally used value of 0.85 at generating usefulness-based rankings of scientific papers. Furthermore, the fact that RALEX can dynamically adjust its roaming parameters with regards to contents and age of documents allows it to consistently outperform the PageRank baseline using a damping factor of 0.667. This confirms the benefit of making the ranker contents and time sensitive. However, due to the fact that contents and time have been organically integrated into the information freshness decay factor, it is hard to separate out their individual contributions. We consistently observed in the two sets of evaluations that RALEX's superior performance is especially pronounced at the top of the produced ranking lists. This suggests that RALEX will perform better in practice to return a more useful top list of papers to users. Table 5.1 shows the top 10 papers in the proxy gold standard and their ranks produced by RALEX and the 2 PageRank baselines. It serves to provide a sneak peak of rankers' relative performance in this arguably most vital, albeit tiny, section of the whole ranking. It is evident that RALEX can rank the best papers at the top in its ranking with a shorter distance to the proxy gold standard than the PageRank baselines.

It is worth pointing out that our RALEX framework is very generic and highly customisable to work on different datasets to facilitate researchers' literature search through the following tuneable parameters:

1. global damping factor
2. base (β) of the information freshness decay penalty
3. personalisation to introduce selective teleportation

[3]As a matter of fact, RALEX consistently outperformed the baselines for all values of k. Results are not shown due to space limitations.

Rather than striving for the best possible performance on a specific corpus against a proxy of ground truth, the primary objective of the research described in this paper was to propose a generic ranking framework that takes both document contents and age into consideration to yield better rankings of academic papers that are more reflective of their potential scientific utility. Consequently, we decided to use the default values for the parameters mentioned above without fine-tuning. This thus leaves a potentially large room for future improvements. The main purpose of this paper was to hypothesise and confirm that making a random-walk based document ranker sensitive to both document contents and age is a step in the right direction to produce usefulness-based rankings of scientific documents.

There are some trivial similarities between RALEX and a few existing systems that belong to the broad category of time-sensitive random walk-based document rankers (e.g., [25, 17, 24]). However, it is important to highlight the fundamental differences in the problems they strive to address as well as the distinctions in their design philosophies. Time-sensitive scientific document rankers specifically aim at obtaining a relative ranking of new papers according to some metric of "future popularity". The design of such systems is strongly directed towards promoting young papers' ranks, thus usually involves distributing random surfers exponentially with age in favour of more recent publications through personalised teleportation. System parameters are also optimised to achieve the best results possible in an application-driven manner. RALEX on the other hand proposes to address the much broader and more fundamental problem of generating rankings of an existing body of research that facilitate further related research. Consequently, we intentionally steered ourselves away from an application-oriented design path and focused on adapting PageRank to produce usefulness-based rankings of scientific papers. We set out to design a random surfing model that more accurately captures human scientific literature explorers' relevance-driven reference following the way they search for potentially useful work. This early divergence in system purposes has led to very different system evaluations. Proxy gold standards based on *near* future citation counts and PageRank scores from some census year have been used in evaluating system performance in [25] and [24], respectively. While tenable in their usage for evaluation of rankings of new papers, such proxy gold standards are not suitable to evaluate RALEX's performance as momentary popularity does not necessarily warrant that a paper can maintain its high utility over time and eventually have a profound impact. As a result, no sensible performance comparisons can be drawn between RALEX and several seemingly similar systems such as those mentioned above. More broadly, though closer in their goals, virtually none of the systems reviewed in Section 2 can be directly compared with RALEX due to either a lack of qualitative evaluation results (e.g., [18]) or to the fact that the proxy gold standard used for the evaluation has not been published (e.g., [15]).

6 Conclusion and Future Work

In this paper we described RALEX, a random-walk based document ranking framework for scientific literature to help researchers identify publications with a high potential for scientific utility in their domains of interest, which is all the more challenging that those domains are at the frontier of science and technology. To the best of our knowledge, this work represents the first attempt to rank scientific papers in decreasing order of potential usefulness taking both topical contents and age of documents into consideration. More specifically, we demonstrated that using corpus specific damping factors estimated on the basis of topical drift along reference paths, better rankings can be produced than the PageRank baseline with a default damping factor of 0.85. We further showed that by making RALEX sensitive to information freshness, taking both document contents and age into consideration, we could further enhance its performance. As a pilot study, we focused on presenting the generic framework without fine-tuning various parameters of RALEX, leaving potentially large room for future improvements. More specifically, we implicitly assumed that the random literature explorer always starts its exploration from papers that are exactly in line with its topics of interest. While somewhat untenable in the context of global ranking, this assumption will be better founded under a personalised setting where the initial rank mass distribution as well as the destination probabilities for random teleportation are adjusted according to some predefined topics of interest. Furthermore, by tracking reading history to estimate a user's topics of interest, RALEX could potentially be used as a *literature guide* to recommend papers and produce personalised rankings. Finally, due to fundamental differences in their aims, no sensible comparison between RALEX and other time-sensitive random walk-based document rankers (e.g., [25, 17, 24]) can be drawn at this stage. However, we acknowledge that recognising high quality recent publications is of great research interest, as they represent the domain frontier and are likely to contribute to future research directions. We are currently working on a version of RALEX with a strong focus on ranking new papers with personalised teleportation as a function of papers' information freshness. We are very interested to explore the combined effects of constraining the flow of rank mass towards outdated literature and infusing rank mass into domain frontiers and to further compare our results to the above mentioned systems directly.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and constructive suggestions.

Bibliography

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

[2] P. Boldi, M. Santini, and S. Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 557–566. ACM, 2005.

[3] S. Bonzi. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, 33(4):208–216, 1982.

[4] M. Bressan and E. Peserico. Choose the damping, choose the ranking? *Journal of Discrete Algorithms*, 8(2):199 – 213, 2010. Selected papers from the 3rd Algorithms and Complexity in Durham Workshop {ACiD} 2007.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and Isdn Systems*, 30(1):107–117, 1998.

[6] B. Brooke. Optimum p% library of scientific periodicals. *Nature*, 232:458–461, 1971.

[7] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, Jan 2007.

[8] V. P. Diodato. *Dictionary of Bibliometrics*. Haworth Press, Inc., 1994. ISBN-1-56024-852-1.

[9] L. Egghe and R. Rousseau. *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Elsevier, 1990.

[10] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, November 1972.

[11] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Information sciences series. Isi Press, 1979.

[12] D. F. Gleich, P. G. Constantine, A. D. Flaxman, and A. Gunawardana. Tracking the random surfer: Empirically measured teleportation parameters in pagerank. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 381–390. ACM, 2010.

[13] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[14] D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371. Association for Computational Linguistics, 2008.

[15] B. King, R. Jha, and D. R. Radev. Heterogeneous Networks and Their Applications: Scientometrics, Name Disambiguation, and Topic Modeling. *Transactions of the Association for Computational Linguistics*, 2013.

[16] R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 571–580. ACM, 2010.

[17] X. Li, B. Liu, and P. Yu. Time Sensitive Ranking with Application to Publication Search. In *2008 Eighth IEEE International Conference on Data Mining (ICDM)*, pages 893–898. IEEE, 2008.

[18] N. Ma, J. Guan, and Y. Zhao. Bringing PageRank to the citation analysis. *Information Processing & Management*, 44(2):800–810, March 2008.

[19] G. S. Mann, D. Mimno, and A. McCallum. Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, pages 65–74. ACM, 2006.

[20] S. Maslov and S. Redner. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *Journal of Neuroscience*, 28(44):11103–11105, October 2008.

[21] M. J. Moravcsik and P. Murugesan. Some Results on the Function and Quality of Citations. *Social Studies of Science*, 5(1):86–92, 1975.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[23] D. R. Radev, P. Muthukrishnan, and V. Qazvinian. The acl anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 54–61. Association for Computational Linguistics, 2009.

[24] H. Sayyadi and L. Getoor. Futurerank: Ranking scientific articles by predicting their future pagerank. In *In Proc. of the 9th SIAM International Conference on Data Mining*, pages 533–544, 2009.

[25] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010–P06010, June 2007.

[26] P. Weingart. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1):117–131, 2005.

[27] A. Wissner-Gross. Preparation of topical reading lists from the link structure of wikipedia. In *Advanced Learning Technologies, 2006. Sixth International Conference on*, pages 825–829, 2006.

[28] H. Xu, E. Martin, and A. Mahidadia. Using heterogeneous features for scientific citation classification. In *Proceedings of the 13th Conference of the Pacific Association for Computational Linguistics*, September 2013.

[29] P. Yu, X. Li, and B. Liu. Adding the temporal dimension to search - a case study in publication search. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 543–549, 2005.

[30] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles. Co-ranking authors and documents in a heterogeneous network. In *International Conference on Data Mining*, pages 739–744. IEEE Computer Society, 2007.