Beyond Keyword Search: Discovering Relevant Scientific Literature

Khalid El-Arini
Computer Science Department
Carnegie Mellon University
kbe@cs.cmu.edu

Carlos Guestrin
Machine Learning Department
Carnegie Mellon University
guestrin@cs.cmu.edu

ABSTRACT

In scientific research, it is often difficult to express information needs as simple keyword queries. We present a more natural way of searching for relevant scientific literature. Rather than a string of keywords, we define a query as a small set of papers deemed relevant to the research task at hand. By optimizing an objective function based on a fine-grained notion of influence between documents, our approach efficiently selects a set of highly relevant articles. Moreover, as scientists trust some authors more than others, results are personalized to individual preferences. In a user study, researchers found the papers recommended by our method to be more useful, trustworthy and diverse than those selected by popular alternatives, such as Google Scholar and a state-of-the-art topic modeling approach.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*query formulation,relevance feedback,retrieval models*

General Terms

Algorithms,Experimentation

Keywords

personalization, citation analysis

1. INTRODUCTION

For generations, scientists have built upon the published work of their predecessors and contemporaries in order to make new discoveries. However, as the number of publications has grown, it has become increasingly difficult for scientists to find relevant prior work for their particular research. In fact, as early as 1755, the French philosopher Denis Diderot presciently forewarned that there would come a day when "it will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes," [13]. Today, we can quantify this "immense multitude" to include tens of millions of articles published in tens of thousands of journals and conferences [43].

Currently, researchers primarily rely on keyword search of online indices such as Google Scholar and PubMed to help them combat this overload of information. While these tools are indispensable, there are many instances where a researcher's information need cannot be easily specified as a simple string of keywords. Often, such a keyword query is either overly broad, returning many articles that are at best loosely related to the researcher's specific need, or too narrow, potentially returning no articles at all. In these occasions, it may be more natural for the scientist to specify his query as a small set of papers rather than as a set of words. In particular, having already read some articles that are related to the specific task at hand, the scientist can ask, "given that these papers represent my immediate research focus, what else should I read?".

Here, we present an algorithm for discovering relevant scientific literature by responding to queries of this form. More formally, given a small set of papers $\mathcal{Q}$ that we refer to as the *query set*, we seek to return a set $\mathcal{A}$ of additional papers that are related to the concept defined by the query. Intuitively, a paper that cites all of the articles in $\mathcal{Q}$ is likely to represent related research. Likewise, a paper that is cited by every article in $\mathcal{Q}$ might contain relevant background information. However, it is restrictive to require the papers in $\mathcal{A}$ to have a direct citation to or from every article in the query set, as such papers are not guaranteed to exist. Instead, we wish to select a set $\mathcal{A}$ that maximizes a more general notion of *influence* to and from the papers in $\mathcal{Q}$.

2. MODELING SCIENTIFIC INFLUENCE

To define a notion of influence in scientific literature, we observe that the content of a publication is an amalgam of several sources, combining cited prior work with the authors' novel insights and background experience. For a given collection of articles, ideas travel *from cited papers to citing papers*, and from earlier to subsequent papers by the same author (Figure 1A). Our notion of influence should capture this transfer of ideas, modeling both the extent to which ideas travel between documents, as well as their topical matter. To achieve such fine-grained detail, we define influence with respect to the *individual concepts* found in a document collection, which could be, e.g., technical terms or informa-

KDD'11, August 21–24, 2011, San Diego, California, USA.
Copyright 2011 ACM 978-1-4503-0813-7/11/08 ...$10.00.

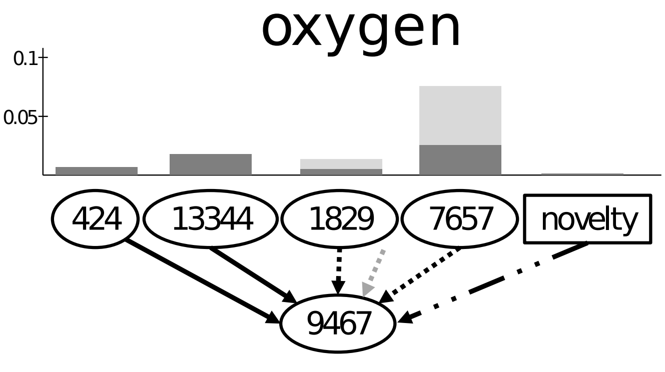

Figure 2: An example from the PNAS data set, illustrating the edge weight computation for a node in G_{oxygen}. Solid black edges indicate citations, while dotted black edges indicate common authorship. The dotted gray edge refers to a paper sharing an author with 9467, but not containing the concept "oxygen." Edge weights are assigned proportional to the bar chart, indicating the prevalence of "oxygen" in each parent node. The bars over 1829 and 7657 are shortened to one third of their original height (indicated in light gray), such that the contribution due to common authorship is equivalent to that of a single paper. The novelty node is only used to normalize the edge weights, and in this case is dominated in influence by the other articles.

tive phrases.[1] For example, we might say that the ideas transferred from one paper to another involve the concepts "energy" or "nitric oxide."

For each concept c in our vocabulary of concepts $\mathcal{C}$, we define a directed, acyclic graph G_c, where the nodes represent papers that contain c and the edges represent citations and common authorship. Figures 1B and 1C show two such graphs for a subset of articles from the Proceedings of the National Academy of Sciences (PNAS), for the concepts "plant" and "stress." While a path between two nodes in such a graph may indicate influence with respect to a particular concept, mere existence of a path does little to express the *degree* to which this influence occurs. To capture the degree of influence, we define a weight $\theta^{(c)}_{x\to y}$ for each edge (x, y) in graph G_c, representing the probability of *direct* influence from paper x to paper y with respect to concept c. We can then use these edge weights to define a probabilistic, concept-specific notion of influence between any two papers in the document collection.

2.1 Defining edge weights

Figure 2 shows an example from the PNAS data set illustrating how we define the weight $\theta^{(c)}_{x\to y}$ on each edge. Here, article 9467 cites two articles containing the concept "oxygen," $\{424, 13344\}$, indicated by the solid black edges. The dotted black edges indicate that two other articles, $\{1829, 7657\}$, contain the concept "oxygen" and share authors with 9467. (The dotted gray edge indicates that there is a third article sharing authors with 9467 that *does not* contain "oxygen.") We assume that every occurrence of the concept "oxygen" in 9467 is either a novel idea or is directly inspired by one of these sources. Thus, we view the weight $\theta^{(c)}_{x\to y}$ as the probability a random instance of concept c in paper y was directly inspired by paper x.

The bar graph over the nodes in Figure 2 illustrates the proportion of the content of each paper consisting of the "oxygen" concept. For instance, the height of the first bar on the left is $n^{(oxygen)}_{424}/N_{424}$, where $n^{(c)}_x$ is the frequency of concept c in document x, and $N_x = \sum_{c\in\mathcal{C}} n^{(c)}_x$ is the total length of document x. Additionally, the bars over 1829 and 7657 are shortened to one third of their original height (indicated in light gray), representing the intuition that an explicit citation is a more informative relationship than common authorship. The authors of 9467 have three prior publications in this example, and thus by dividing by three, the effective total contribution of these papers is that of a single paper. Finally, we represent the novelty distribution for a particular paper y as the average distribution over concepts for all papers published in the same year as y. In this case, the novelty contribution for "oxygen" is dominated by the four papers. (We note that there are no actual novelty nodes in the graph, as the associated distribution is only used for normalization.)

Here, $\theta^{(oxygen)}_{x\to 9467}$ is proportional to the height of the corresponding bar in the plot. More generally, if a paper y cites papers $\{r_1, \ldots, r_k\}$, and the authors have previously written papers $\{b_1, \ldots, b_l\}$, then the edge weights are defined as follows:

$$\begin{aligned}\theta^{(c)}_{r_i\to y} &= \frac{1}{Z}\frac{n^{(c)}_{r_i}}{N_{r_i}},\\ \theta^{(c)}_{b_i\to y} &= \frac{1}{Z}\frac{n^{(c)}_{b_i}}{l\cdot N_{b_i}},\end{aligned}$$

with normalization constant,

$$Z = \sum_{j=1}^{k}\frac{n^{(c)}_{r_j}}{N_{r_j}} + \frac{1}{l}\sum_{j=1}^{l}\frac{n^{(c)}_{b_j}}{N_{b_j}} + novel^{(c)}_y,$$

where $novel^{(c)}_y$ is the average proportion of concept c across all papers published in the same year as y.

2.2 Calculating influence

Given a concept-specific weight for each edge in the citation graph, representing the *direct* influence between two neighboring nodes, we can now define the influence between any two papers in our collection. In particular, if we say that each edge $x \to y$ in G_c is *active* with some probability $\theta^{(c)}_{x\to y}$, we arrive at the following definition:

DEFINITION 1. *The influence between papers u and v with respect to concept c, $Influence_c(u \leftrightarrow v)$, is the probability there exists a directed path in G_c from one paper to the other, consisting only of active edges.*

While intuitive, the exact computation of this probability is intractable, as the problem of computing connectedness in a random graph belongs to the #P-complete class of computational problems [45, 37], for which there are no known polynomial-time solutions. We can overcome this computational hurdle via approximation, by employing one of two methods: 1) a simple Monte Carlo sampling procedure with theoretical guarantees; and, 2) a deterministic, linear-time

[1] For our experiments, we use a simple tf-idf heuristic to extract informative words which we use as concepts (cf. [15]).

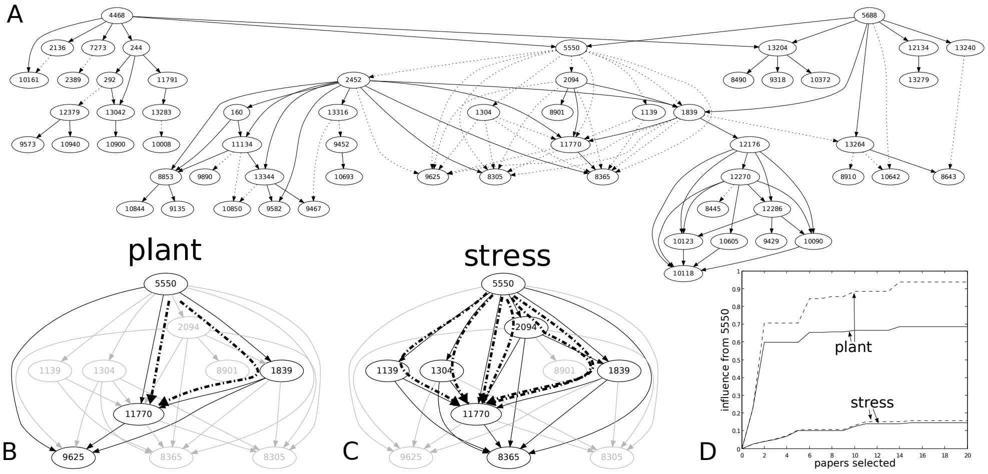

Figure 1: (A) A graph of articles from the Proceedings of the National Academy of Sciences (PNAS). Nodes represent papers, solid edges represent citations ($x \to y$ if y cites x) and dotted edges represent common authorship ($x \to y$ if x is older than y and x, y share an author). More details on the data sets used in this paper can be found in the technical report [15]. (B,C) Subgraphs of (A), limited to papers containing the concepts "plant" and "stress," respectively (other papers are grayed out). Thick dashed lines indicate paths of influence between papers 5550 and 11770. (D) Example illustrating how Equation 1 penalizes redundancy. The first two papers selected exhibit a high influence with respect to "plant," and thus subsequently adding such papers to $\mathcal{A}$ causes little increase in Equation 1 (solid lines), especially when compared to the sum of individual influences (dashed lines). The influence with respect to "stress" remains low, thus never triggering such a redundancy penalty.

dynamic programming heuristic, based on the assumption that the paths between two nodes are independent of each other.

2.2.1 Sampling

The simplest procedure for estimating the influence between two nodes is to generate samples directly based on the definition of influence. Each sample is generated as follows:

For each concept c:

1. Mark each edge $x \to y$ in G_c as active with probability $\theta^{(c)}_{x \to y}$.
2. For all pairs of nodes (u, v), record whether a path exists between them using only active edges.

After generating B samples, the probability that a node u influences a node v with respect to concept c is simply estimated as the proportion of the B samples for concept c in which an active path from u to v exists. A natural question to ask is, how many samples do we need for a reasonable estimate of influence? A short proof using Hoeffding's Inequality–found in the technical report [15]–shows us that the number of samples we need grows only *logarithmically* with the number of articles in the document collection.

THEOREM 1. *In order to estimate m influence values such that, with probability η,* each *of the m estimates is no more than δ away from its true value, a sufficient number of samples B is* $\frac{2}{\delta^2}\log(2m/\delta)$.

As the number of influence values to estimate is quadratic in the number of articles, the number of samples we need is logarithmic in the total number of articles. While this is a heartening result, we find that for large document collections, generating enough samples can still be a time-consuming process.

2.2.2 Independence heuristic

As an alternative to sampling, we describe an efficient dynamic programming heuristic based on the assumption that the paths between two nodes in G_c are independent of each other. For instance, in Figure 1B, the two influence paths between 5550 and 11770 with respect to the concept "plant" are completely independent of each other. Thus, the probability of at least one active path existing between the two nodes in this situation can be computed exactly:

$$\begin{aligned}
&\mathit{Influence}_{plant}(5550 \to 11770)\\
&\quad = \quad 1 - P(\text{there is no influence between 5550 and 11770})\\
&\quad = \quad 1 - P(\text{there is no direct influence from 5550}) \cdot\\
&\qquad\quad P(\text{there is no influence through 1839})\\
&\quad = \quad 1 - (1 - \theta^{(plant)}_{5550 \to 11770})(1 - \theta^{(plant)}_{5550 \to 1839}\theta^{(plant)}_{1839 \to 11770}).
\end{aligned}$$

The second equality follows from the independence of the two paths. On the other hand, looking at Figure 1C, we find the paths between the two nodes in G_{stress} are not independent, making such a calculation more problematic.

Based on this intuition, if we rashly assume that the paths between two nodes will *always* be independent of each other in G_c, for all c, we arrive at a simple, efficient heuristic for computing the influence between all pairs of nodes (Algorithm 1). By traversing the graph in topological order, we know that when we arrive at a node we will have already computed all the influence going to its parents. Using these influences and our independence assumption, we can then immediately compute the influence to the node itself. We note that this algorithm requires the graphs to be acyclic.[2]

[2] Based on simple chronology, one would expect a citation graph to be acyclic; after all, a researcher cannot cite a paper

Algorithm 1 Dynamic Programming Heuristic for Influence

```
$N$: number of documents
$\mathcal{C}$: vocabulary of concepts
// Initialize to empty 3D array
// $influenceEstimate[c][x][y]$ will contain influence
//     from $x$ to $y$ with respect to concept $c$.
$influenceEstimate \leftarrow array[|\mathcal{C}|][N][N]$
for all $c \in \mathcal{C}$ do
  for all nodes $y$ in $G_c$ do
    // Initialize to identity
    $influenceEstimate[c][y][y] \leftarrow 1$
  $topoOrder \leftarrow$ topological order of nodes in $G_c$
  for $y \in topoOrder$ do
    // $influenceEstimate[c][][x]$ already calculated
    //     for all $x \in parents(y)$
    if $parents(y) = \emptyset$ then
      continue
    $influenceFromParents \leftarrow array[|parents(y)|]$
    for all $x \in parents(y)$ do
      // Influence to the parent multiplied by
      //     the edge weight
      $influenceFromParents[x] \leftarrow$
          $influenceEstimate[c][][x] \cdot \theta^{(c)}_{x \to y}$
    // Product is element-wise
    $influenceEstimate[c][][y] \leftarrow$
        $1 - \prod_{x \in parents(y)}(1 - influenceFromParents[x])$
```

While the independence assumption upon which this heuristic is based certainly is not true in general, we find that, nevertheless, the values we compute are close to what we would expect from sampling. (Empirical results to this effect can be found in the technical report [15].) Thus, despite not being amenable to theoretical guarantees, we find this heuristic works well in practice.

2.3 Selecting papers

As motivated in Section 1, given a query set of papers $\mathcal{Q}$, we wish to select a small set of related papers $\mathcal{A}$ that exhibit a high degree of *influence* to or from the query set. Moreover, the set of papers we select should be both *relevant* and *diverse*.

2.3.1 Relevance

The influence between the query set $\mathcal{Q}$ and the result set $\mathcal{A}$ should be focused on the concepts that are important or prevalent in both sets of documents. First, to ensure that the selected documents pertain to the concepts most prevalent in the query set, we define a weight $\gamma^{(c)}_q$ proportional to the frequency of concept c in query document q.

Likewise, from the perspective of the result set, a document d might contain a single occurrence of the concept "plant," and that single occurrence might be heavily influenced by one of the query documents q. However, as d only tangentially mentions "plant," we do not wish this strong influence to incentivize its inclusion in the result set. Thus, we define a probability $\beta^{(c)}_d$ indicating the importance of concept c in document d. Specifically, we define this as the probability a concept c is observed in a finite number ℓ of independent samples (with replacement) from the document's word distribution: $\beta^{(c)}_d = 1 - (1 - \gamma^{(c)}_d)^{\ell}$. (Here, ℓ is a parameter of our model that we set to 20 in our experiments.)

Figure 3 provides an illustrative example of these weights.

2.3.2 Diversity

Diversity is important in this setting as it is difficult to predict the exact information need of a researcher, and thus providing a wide variety of papers increases the likelihood of query satisfaction. As such, we define the influence between a single query paper $q \in \mathcal{Q}$ and a set of documents $\mathcal{A}$ in a manner that penalizes redundancy in the result set, thereby promoting diversity. Specifically, if we define this *set influence*, $\mathit{Influence}_c(q \leftrightarrow \mathcal{A})$, as the probability influence exists between q and *at least one* document in $\mathcal{A}$, we create a disincentive for $\mathcal{A}$ to contain multiple papers with similar influence patterns to and from q; such a redundant set $\mathcal{A}$ would exhibit less influence than one composed of a broader set of documents. Formally,

$$\begin{aligned}&\mathit{Influence}_c(q \leftrightarrow \mathcal{A}) = \\ &\quad 1 - \prod_{d \in \mathcal{A}} \left(1 - \mathit{Influence}_c(q \leftrightarrow d)\beta^{(c)}_d\right). \qquad (1)\end{aligned}$$

We note the use of the probability $\beta^{(c)}_d$ here to safeguard against selecting documents that are only tangentially related to the important concepts in the query papers.

Figure 1D shows an example illustrating how the marginal gain in set influence with respect to the concept "plant" diminishes as more papers are added to the result set $\mathcal{A}$. In particular, beyond a certain level of influence, the gain observed in Equation 1 from adding additional documents to the result set is smaller than would be expected if we were naïvely summing the individual influences. We do not see the same redundancy penalty with respect to "stress," as the result set is not sufficiently influenced with respect to this concept.

2.3.3 Optimization

Given this definition of set influence, we can now define an objective function that, when maximized, returns a diverse set of papers highly relevant to the query:

$$F_{\mathcal{Q}}(\mathcal{A}) \quad = \quad \sum_{q \in \mathcal{Q}} \sum_{c \in \mathcal{C}} \gamma^{(c)}_q \mathit{Influence}_c(q \leftrightarrow \mathcal{A}). \qquad (2)$$

While, in general, solving such a combinatorial optimization problem is intractable, Equation 2 exhibits an intuitive diminishing returns property known as *submodularity*, allowing for efficient near-optimal solutions.

Definition 2 (Submodularity). *A set function F is* submodular *if, $\forall \mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}, \forall s \in \mathcal{V} \backslash \mathcal{B}, F(\mathcal{A} \cup \{s\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{s\}) - F(\mathcal{B})$.*

Intuitively, this means that the utility of adding a particular paper to a result set decreases as the result set gets larger.

Theorem 2. *Equation 2 is submodular and monotonic.*

The proof can be found in the technical report [15].

if it does not yet exist. However, this is not quite the case in practice (e.g., colleagues writing papers simultaneously may cite each other). Details on how we address this problem can be found in the technical report [15].

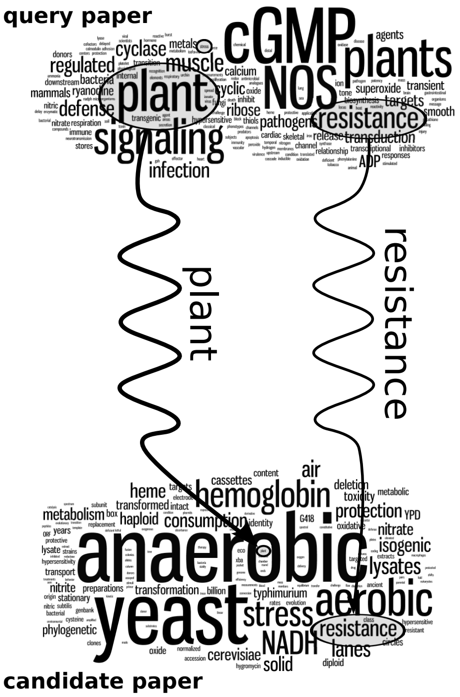

Figure 3: The top cloud represents a query paper (5550), the bottom word cloud represents a paper to be selected (11770) and the lines between them represent individual influences of varying strength. In each word cloud, the size of a word is proportional to its frequency in the corresponding article. γ is illustrated by the shaded ellipses in the top word cloud, showing a higher incentive to pick articles about "plant" or "resistance" than about "stress." However, despite its prevalence in the query document, "plant" is only tangentially present in article 11770, and thus β ensures a low degree of influence. This can be contrasted with "resistance," which is prevalent in both documents and displays a high degree of influence.

Although maximizing submodular functions is NP-hard [26], by discovering this property in our problem, we can take advantage of several efficient approximation algorithms with theoretical guarantees. For example, the classic result of Nemhauser et al. [31] shows that by simply applying a greedy algorithm to maximize our objective function, we can obtain a $(1 - \frac{1}{e})$ approximation of the optimal value. Thus, a simple greedy optimization can provide us with a near-optimal solution. However, since our set of articles is very large, a naïve greedy approach can be too costly. Therefore, we use CELF [29], which provides the same approximation guarantees, but uses lazy evaluations, often leading to dramatic speedups.

3. TRUST AND PERSONALIZATION

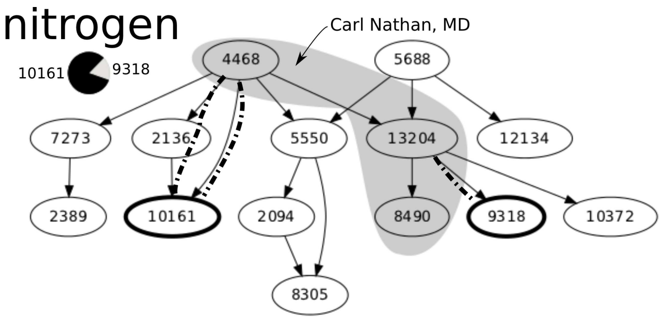

Figure 4: Example illustrating trust calculation for an immunologist asking, "How much do I trust Carl Nathan with respect to the concept 'nitrogen'?" Thick dashed lines indicate influence from Dr. Nathan to individual elements of $\mathcal{B}$, and pie chart represents relative prevalence of the word "nitrogen" in the two papers in $\mathcal{B}$.

Considering our running example of PNAS articles (Figure 1A), we can set our query set to be $\mathcal{Q} = \{4468, 5688\}$, the parents of "Nitric Oxide in Plant Immunity" (5550). Optimizing Equation 2 for this query produces a result set of articles ranging in topics from plant biology to immunology (cf. technical report [15]). While these articles may be relevant to the query, a major shortcoming is that every researcher who submits this query will receive an identical result set. For any given topic, different researchers trust different authors and publications, and the objective in Equation 2 provides no means to express these preferences. While a long line of prior work exists on summarizing the impact of an author or publication with a single number [2], often based on citation statistics [20, 25] or eigenvector methods [27, 35, 11, 38], here we wish to capture a more detailed picture of the relationship between a researcher and the authors he cites.

In order to properly model such an individual notion of *trust* in the setting of scholarly research, we consider two motivating scenarios:

1. Different authors command different levels of respect from their research communities, e.g., a Nobel laureate versus a first-year graduate student, as an extreme case.
2. Even among distinguished scientists, a particular researcher's interests may be aligned more closely with some than others. Thus, beyond simply differentiating novices from experts, a notion of trust should also capture differences in research interests. For example, asking computer scientists to name whom they most associate with the concept "network" may yield Judea Pearl (Bayesian networks), Jon Kleinberg (social networks), Geoff Hinton (neural networks) or Van Jacobson (computer networks), depending on who is answering. All are distinguished researchers, but each is associated with a distinct subfield of computer science.

At the heart of both scenarios is a personal question that is often answered differently by different researchers: *How much do I trust this author with respect to this concept?*

By answering this question, a researcher would enable us to formally incorporate his trust preferences into our objective function, allowing us to select papers tailored specifically to his tastes. However, as researchers will not be able to provide an answer for every combination of authors and concepts, we must elicit their trust preferences in a less onerous manner. In order to do so, we assume that trust is *transitive.* For example, if Alice trusts an article, and that article is heavily influenced by Bob with respect to the concept "network," then Alice is likely to also trust Bob with respect to "network." Thus, at a fundamental level, a researcher need only specify a set of trusted papers $\mathcal{B}$, from which we can infer answers to the above question. As a shortcut, a researcher may choose to define $\mathcal{B}$ indirectly by specifying a list of trusted journals and conferences, or subsets thereof (e.g., a particular conference track or article classification). $\mathcal{B}$ could also be specified as the papers *cited by* one or more trusted authors, representing a look at one's research through the eyes of another scientist, potentially in another field. Thus, a plucky physicist could ask, "What would Steven Chu recommend I read?", and obtain a set of papers related to his query, yet tailored to the research interests and trust preferences of the Nobel laureate.

With this intuition in mind, we define $\tau^{(c)}_{a|\mathcal{B}}$, the probability a researcher trusts author a with respect to concept c, given trusted articles $\mathcal{B}$. (The "$|\mathcal{B}$" notation in this section indicates personalizing with respect to trusted set $\mathcal{B}$.) Figure 4 illustrates how we compute $\tau^{(c)}_{a|\mathcal{B}}$ for a particular example from PNAS, where the concept c is "nitrogen," the author a is Carl Nathan, MD, and the researcher has specified two immunology papers as his trusted set, $\mathcal{B} = \{10161, 9318\}$. For each paper $b \in \mathcal{B}$, we compute how much the author a influenced b with respect to concept c. As our influence is now expressed from an *author* to an article, we treat all of an author's papers as a single unit.

Definition 3. *The influence from author a to article b with respect to concept c, AuthorInfluence*$_c(a \to b)$*, is the probability there exists a directed path in* G_c *from* any *article written by a to article b consisting only of active edges, where each edge is (independently) active with probability* $\theta^{(c)}_{x\to y}$.

As before, we employ sampling or dynamic programming to efficiently estimate this otherwise intractable computation (cf. technical report for more details [15]).

In our example, we first look at how much Dr. Nathan's papers influence 10161 with respect to "nitrogen," and again from Dr. Nathan's papers to 9318. We now weigh these two influences by the prevalence of the word "nitrogen" in each paper b (as indicated by the pie chart in Figure 4), and define $\tau^{(c)}_{a|\mathcal{B}}$ to be the weighted sum of the two.

More generally, we have:

$$\tau^{(c)}_{a|\mathcal{B}} = \begin{cases} \frac{1}{N^{(c)}_{\mathcal{B}}} \sum_{b\in\mathcal{B}} n^{(c)}_b \, \mathit{AuthorInfluence}_c(a \to b), & \text{if } N^{(c)}_{\mathcal{B}} > 0 \\ \tau^{(c)}_{a|\mathcal{V}}, & \text{otherwise,} \end{cases}$$

where $N^{(c)}_{\mathcal{B}}$ is the total number of occurrences of concept c in the set $\mathcal{B}$, $n^{(c)}_b$ is the frequency of concept c in paper b, and $\mathcal{V}$ is the set of all papers in the corpus. Here, the influence to each $b \in \mathcal{B}$ is weighted by the relative prevalence of concept c with respect to $\mathcal{B}$, $n^{(c)}_b / N^{(c)}_{\mathcal{B}}$. We note that if a researcher's trusted set $\mathcal{B}$ contains no occurrences of a particular concept, we assign the trust value to $\tau^{(c)}_{a|\mathcal{V}}$, as if all the papers in the corpus were trusted equally.

In order to incorporate trust into paper selection, we assume an author will trust a paper if and only if he trusts *at least one of its authors.* This intuition can be formalized by defining a modified notion of set influence, where the researcher's preferences towards the authors are directly taken into account:

$$\mathit{Influence}_c(q \leftrightarrow \mathcal{A}|\mathcal{B}) = 1 - \prod_{d\in\mathcal{A}} \left(1 - \mathit{Influence}_c(q \leftrightarrow d)\beta^{(c)}_d T^{(c)}_{d|\mathcal{B}}\right),$$

where $T^{(c)}_{d|\mathcal{B}} = 1 - \prod_{a\in authors(d)} (1 - \tau^{(c)}_{a|\mathcal{B}})$.

We can now define our *personalized* objective function as:

$$F_{\mathcal{Q}|\mathcal{B}}(\mathcal{A}) = \sum_{q\in\mathcal{Q}} \sum_{c\in\mathcal{C}} \gamma^{(c)}_q \mathit{Influence}_c(q \leftrightarrow \mathcal{A}|\mathcal{B}). \quad (3)$$

Maximizing $F_{\mathcal{Q}|\mathcal{B}}(\mathcal{A})$ subject to $|\mathcal{A}| \le k$, for some budget of k papers, leads to a personalized set of papers tailored to someone who trusts $\mathcal{B}$. This function shares the same theoretical properties as Equation 2 and can be optimized efficiently in the same manner (cf. technical report for more details [15]).

Figure 5 shows our PNAS example from before, with the same query set $\mathcal{Q} = \{4468, 5688\}$, but now incorporating the trust preferences of two hypothetical researchers: a plant biologist (A) and an immunologist (B). The figure highlights how differences in trust preferences can manifest themselves in article selection. In Figure 6, we provide another example, this time from computer science. Here, we take the famous Faloutsos, Faloutsos and Faloutsos paper, "On power-law relationships of the Internet topology" [19], and select related literature for it using the trust preferences of each author. Specifically, the visualization in the figure shows that by assuming that Michalis Faloutsos trusts SIGCOMM papers, Petros Faloutsos trusts SIGGRAPH papers, and Christos Faloutsos trusts KDD papers, we can select related work tailored to each author's perspective. While some relevant papers are common to all three points of view, other selected papers are particular to just one. For example, in Christos' data mining-focused result set, we find a few papers related to the evolution of social networks (e.g., "Microscopic evolution of social networks" by Leskovec et al.) which are not found in Michalis' and Petros' results. Moreover, these papers are not selected in the unpersonalized setting, when no trust preferences are taken into account.

4. APPROACH SUMMARY

We summarize our approach as follows:

Initialization

1. Define a vocabulary of concepts $\mathcal{C}$ (e.g., technical terms).
2. For each concept $c \in \mathcal{C}$, define a directed, acyclic graph G_c, with edge weights as in Section 2.1.
3. Compute relevance weights $\gamma^{(c)}_d$ and $\beta^{(c)}_d$, for all $c \in \mathcal{C}$ and documents d, as described in Section 2.3.1.
4. Precompute $\mathit{Influence}_c(u \leftrightarrow v)$ for all concepts c, and all pairs of documents u and v, using Algorithm 1.
5. Similarly, precompute $\mathit{AuthorInfluence}_c(a \to b)$, for all authors a, all documents b, and all $c \in \mathcal{C}$ (cf. [15]).

Per user

Given a user's trusted set of papers $\mathcal{B}$, compute $\tau^{(c)}_{a|\mathcal{B}}$ for all authors a and $c \in \mathcal{C}$.

Per query

Given query set $\mathcal{Q}$, optimize Equation 3 using CELF [29].

5. RELATED WORK

Researchers in both the library science and computer science communities have studied the shortcomings of the traditional keyword search paradigm [5, 34, 36]. In fact, our

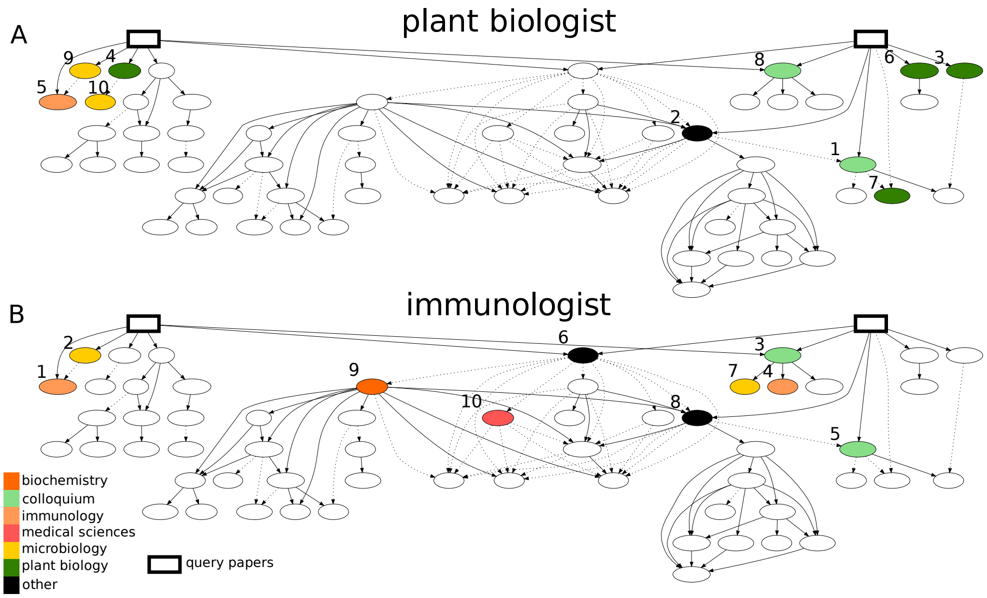

Figure 5: Top ten papers selected for $\mathcal{Q} = \{4468, 5688\}$ where $\mathcal{B}$ is defined as (A) all the plant biology papers, or (B) all the immunology papers, in the PNAS data set. Node colors correspond to article classification, as indicated by the key. (Colloquium refers to the National Academy of Sciences Colloquium on Virulence and Defense in Host-Pathogen Interactions: Common Features Between Plants and Animals. "Other" refers to unclassified papers, e.g., "From the Academy.".) Numbers indicate order of selection by optimization algorithm, roughly indicating order of importance (cf. technical report for more details [15]).

specific query model of defining a researcher's information need as a set of papers rather than as a keyword string has been described before [9, 30]. In one particularly related line of research, collaborative filtering techniques that have been successful for movie and product recommendations were adapted to the paper recommendation setting [30, 44]. Another approach uses hypothesis testing to determine the articles that most influence each paper–the paper's *Information Genealogy*–based only on article text [42]. Unlike these previous approaches, our methodology is based on a *unified* model of text and citations that places special emphasis on the different trust preferences of individual researchers.

Previous work has also considered the more general, yet related, problem of taking positive examples of membership in a set and using them to expand the set [17, 22]. While such approaches have been applied to the domain of research literature, they do not explicitly model the particular characteristics of our problem, e.g., the effect of citations, publication venues and authorship.

Moreover, it is also important to note that our algorithm is *operational* in that it describes a method for selecting papers, in contrast with many *descriptive* studies in bibliometrics, sociology and other fields [12, 39, 32, 33, 4, 40]. In particular, the large body of work on *topic modeling* in computer science and statistics focuses on fitting probabilistic models to document collections by modeling latent themes in the data [8]. While often applied to corpora of scholarly literature [18, 24, 3, 7, 41, 6, 14, 21], paper recommendation is not the primary objective of these models. Rather, our algorithm follows from a line of work that frames document selection as an explicit optimization problem (cf. [16]).

Finally, we note that the approach we describe in this paper is, in fact, agnostic to the specific definition of influence we use, and thus while we define influence to have an explicit probabilistic interpretation, other such definitions are possible. For instance, recent work by Lao and Cohen [28] provides an approach based on path-constrained random walks, which we can plug in as an alternative definition for influence.

6. EXPERIMENTAL RESULTS

While these illustrative examples provide intuition, in order to truly evaluate our methodology we must solicit feedback from real scientific researchers. To this end, we conducted a user study involving sixteen subjects (all doctoral students in computer science or related fields).

We compare two variants of our algorithm–with and without incorporating trust preferences of the participant–with three representative alternative techniques: Google Scholar [23], Information Genealogy [42] (a hypothesis testing approach based on document text), and the Relational Topic Model [10] (a state-of-the-art topic model incorporating both text and citations to model latent themes in data)[3]. For each participant, we use each of these techniques to find related work for a previously written paper–that participant's *study paper*–thereby simulating a real research scenario. We define each query set $\mathcal{Q}$ to be the references of the corresponding study paper, and we ask each participant to list up to four trusted conferences or journals, which we use to define $\mathcal{B}$.

[3] Previous work [30] has shown that Google keyword search outperforms collaborative filtering techniques for selecting useful papers, and thus we do not directly compare against these approaches.

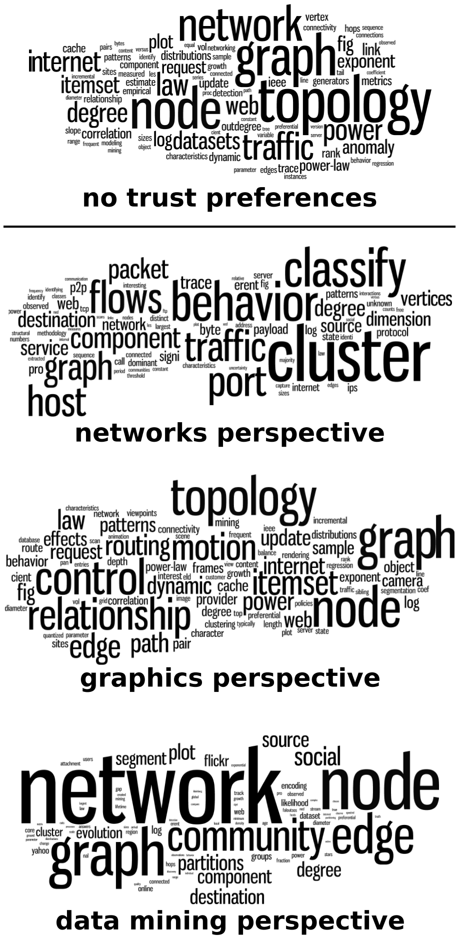

Figure 6: A visualization of related work for Faloutsos, Faloutsos, and Faloutsos' "On power-law relationships of the Internet topology." The top word cloud represents papers selected using Equation 2, with no trust preferences. (The size of each word in the cloud is proportional to its prevalence in the selected documents.) The subsequent three word clouds represent papers selected using Equation 3 with three different trusted sets $\mathcal{B}$, one representing each author's perspective. Each word cloud visualizes the selected papers that are unique to each author's result set. For example, the bottom word cloud shows the papers found in Christos' data mining-focused results, but do not exist in Petros' or Michalis' result sets.

For Google Scholar, we ask a coauthor of each participant to provide us with the keyword query he or she would use to find related work for their paper. The Relational Topic Model and Information Genealogy approaches each use the abstract of the study paper in their paper recommendation, which our algorithm does not require. Unlike many previous studies, each participant was asked to evaluate all five comparison methods, rather than just a single technique. In total, 612 distinct papers were recommended using these five techniques across all sixteen participants. (The articles used in this study come from the ACM Digital Library [1]. More details on the study can be found in the technical report.)

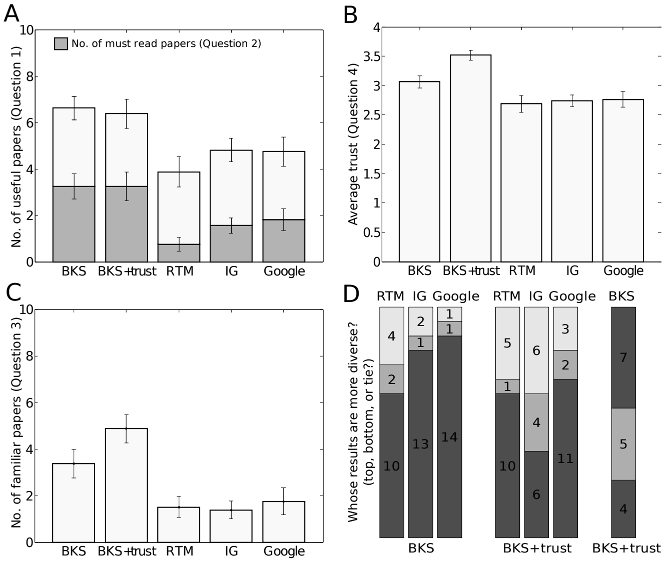

Figure 7: User study results comparing two variants of our algorithm, Beyond Keyword Search (BKS), with and without incorporating trust preferences, with the Relational Topic Model (RTM), Information Genealogy (IG) and Google Scholar. Values in bar plots (A), (B) and (C) are responses to the indicated study questions averaged over all sixteen participants, with error bars indicating one standard error. (D) shows how many participants (out of 16) found that our method produced more diverse results compared to the alternative techniques.

Each participant was presented with the recommended articles for his or her study paper in a double-blind fashion, masking the identity of the technique used to select each paper. Participants were asked to answer questions on the usefulness, novelty and trustworthiness of each paper with respect to their research. Additionally, participants were presented with entire result sets and asked to evaluate them in terms of diversity. Figure 7 shows the results of the study, from which we can glean the following main observations:

1. On average, users find the papers our algorithm selects to be more useful than those selected by the comparison techniques. The topic modeling approach performs especially poorly, with fewer than half of selected papers deemed useful.
2. Explicitly modeling the individual trust preferences of users leads to more trustworthy papers being selected. However, this comes at the expense of novelty in the selected articles, as researchers are more familiar with the work of authors they trust.
3. Our algorithm provides more diverse results than the comparison techniques, which is unsurprising, as our objective functions penalize redundancy.

7. DISCUSSION

These results illustrate the success of our approach in recommending highly relevant literature personalized to the preferences of individual researchers, acting as a promising complement to keyword search. On a personal note, employing our approach during the writing of *this article* led us to related work from another subfield of computer science that we had not discovered using more traditional search methods [30, 44]. In closing, we believe that the challenges researchers face in expressing their information needs extend beyond scientific literature to domains like patents, law and news, and the work presented herein is a significant step towards addressing this general concern.

8. ACKNOWLEDGMENTS

Research funded by NSF CNS-0625518 and ONR YIP N000140810752 and PECASE N000141010672. Word clouds generated using wordle.net. The authors are grateful to the Association for Computing Machinery for providing data from their Digital Library.

9. REFERENCES

[1] ACM Digital Library. http://portal.acm.org.

[2] R. Adler, J. Ewing, and P. Taylor. Citation statistics. *Statistical Science*, 24:1–14, 2009.

[3] E. M. Airoldi, E. A. Erosheva, S. E. Fienberg, C. Joutard, T. Love, and S. Shringarpure. Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences USA*, 2010.

[4] A.-L. Barabási. On the topology of the scientific collaboration networks. *Physica A*, 311:590–614, 2002.

[5] M. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13:407–424, 1989.

[6] D. M. Blei and J. Lafferty. Dynamic topic models. In *ICML*, 2006.

[7] D. M. Blei and J. Lafferty. A correlated topic model of *science*. *Ann. Appl. Stat.*, 1:17–35, 2007.

[8] D. M. Blei and J. Lafferty. *Topic Models*. Chapman and Hall, 2009.

[9] K. Bollacker, S. Lawrence, and C. L. Giles. Discovering relevant scientific literature on the Web. *IEEE Intelligent Systems and their Applications*, 15:42–47, 2000.

[10] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010.

[11] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google. *Journal of Informetrics*, 1:8–15, 2007.

[12] D. J. de Solla Price. Networks of scientific papers. *Science*, 149:510:515, 1965.

[13] D. Diderot. In D. Diderot and J. d'Alembert, editors, *Encyclopedia, or a systematic dictionary of the sciences, arts and crafts*, Paris, 1755. Briasson, David, Le Breton, and Durand. (tr. from French).

[14] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML*, 2007.

[15] K. El-Arini and C. Guestrin. Beyond keyword search: Discovering relevant scientific literature. Technical report, Carnegie Mellon University Machine Learning Department, 2011.

[16] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *KDD*, 2009.

[17] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *KDD*, 2008.

[18] E. A. Erosheva, S. E. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences USA*, 101:5220–5227, 2004.

[19] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *SIGCOMM*, 1999.

[20] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.

[21] S. Gerrish and D. M. Blei. A language-based approach to measuring scholarly impact. In *ICML*, 2010.

[22] Z. Ghahramani and K. A. Heller. Bayesian sets. In *NIPS*, 2006.

[23] Google Scholar. http://scholar.google.com.

[24] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences USA*, 101:5228–5235, 2004.

[25] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences USA*, 102:16569–16572, 2005.

[26] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 1999.

[27] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, 1999.

[28] N. Lao and W. W. Cohen. Relational learning using a combination of path-constrained random walks. *Machine Learning*, 81(1):53–67, 2010.

[29] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[30] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. On the recommending of citations for research papers. In *CSCW*, 2002.

[31] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[32] M. E. J. Newman. Scientific collaboration networks: I. network construction and fundamental results. *Phys. Rev. E*, 64:016131, 2001.

[33] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98:404–409, 2001.

[34] C. Olston and E. H. Chi. Scenttrails: Integrating browsing and searching on the Web. *ACM Transactions on Computer-Human Interaction*, 10:177–197, 2003.

[35] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford University InfoLab, 1999.

[36] S. Pandit and C. Olston. Navigation-aided retrieval. In *WWW*, 2007.

[37] J. S. Provan and M. O. Ball. The complexity of counting cuts and of computing the probability that a graph is connected. *SIAM J. Comput.*, 12:777–788, 1983.

[38] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80:056103, 2009.

[39] S. Redner. How popular is your paper? an empirical study of the citation distribution. *Eur. Phys. J. B*, 4:131–134, 1998.

[40] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA*, 105:1118–1123, 2008.

[41] M. Rozen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[42] B. Shaparenko and T. Joachims. Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases. In *KDD*, 2007.

[43] Thomson Reuters Web of Knowledge. http://wokinfo.com/about/whatitis.

[44] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl. Enhancing digital libraries with TechLens+. In *JCDL*, 2004.

[45] L. Valiant. The complexity of enumeration and reliability problems. *SIAM J. Comput.*, 8:410–421, 1979.