Scientific Referential Metadata Creation with Information Retrieval and Labeled Topic Modeling

Han Jia  
School of Library and Information Science  
Indiana University  
hanjia@indiana.edu

Xiaozhong Liu  
School of Library and Information Science  
Indiana University  
lu237@indiana.edu

Abstract

The goal of this research is to propose an innovative method of creating scientific referential metadata for a cyberinfrastructure-enabled learning environment to enhance learning experiences and to help students and scholars obtain better understanding of scientific publications. By using information retrieval, topic modeling, and meta-search approaches, different types of resources, such as related Wikipedia Pages, Datasets, Source Code, Video Lectures, Presentation Slides, and (online) Tutorials, for an assortment of publications and scientific (labeled) topics will be automatically retrieved, associated, and ranked. In order to test our method of automatic cyberlearning referential metadata generation, we designed a user experiment for the quality of the metadata for each scientific keyword and publication and resource ranking algorithms. Evaluation results based on MAP, MRR, and NDCG show that the cyberlearning referential metadata retrieved via meta-search and statistical relevance ranking can effectively help students better understand the essence of scientific keywords and publications.

Keywords: metadata generation, information retrieval, referential metadata, cyberlearning resource, user, scientific publication, labeled topic modeling

Introduction

In the past decades, rapid access to digital publications accelerated and facilitated study and research; however, several challenges should be addressed. First of all, the sheer volume of scholarly publications available online makes it impossible for a researcher to absorb all the new information available. Hence, researchers and students need to find innovative ways of quickly and effectively learning and understanding new scientific topics and publications. But existing tools are only limited to descriptive information about publications of a specific topic. Second, understanding the content of scientific publications remains daunting. For instance, in a recent survey we conducted with students in a class on “information retrieval theory and practice”, the complex models, algorithms, formulas, and methodologies in the publications were often found too difficult to understand due to their limited backgrounds in computer science, statistics, and mathematics. Third, some recent exciting developments have illustrated the possibility of utilizing multimedia content—i.e. videos and images—to facilitate students and scholars to understand scientific content. However, the cost of generating sophisticated cyberlearning resources for large-scale scientific topics or publications makes the approach prohibitive.

Metadata have traditionally centered on descriptive representation through title, author, publisher, subject keywords and other attributes of scholarly output. Descriptive metadata, however, have become increasingly inadequate as the complexity and volume of scholarly output grows. Innovative mechanisms have been developed to address these new challenges, for examples, Liu, et al. (2011) and Liu (2012) proposed referential metadata, which refers to information about any sources implicitly or explicitly cited in a publication or about artifacts associated with the publication. Referential metadata provides a context, in which the publication was created, the co-authors who collaborated, and the information and data that were used. They also link to the artifacts that the publication may have generated: a presentation video or slides, images, datasets, tutorial, source codes, and even question-answering documents. The referential
metadata may not all exist in the publication as references and some may be scattered across the researcher’s project or personal website or social media, i.e. TED, SlideShare, Sourceforge, and Wikipedia, making it more difficult to obtain them by conventional metadata generation methods. However, referential metadata can effectively help readers to better understand the essence of the publication. As detailed in the figure below (Figure 1), in order to help students and scholars understand such complex publications, it is critical to offer them resources that may enhance their learning process, such as user-friendly tutorials and video lectures.

Figure 1. Depiction of cyberlearning resources related to a scholarly publication.

For this research, instead of generating resources (referential metadata) for each topic or publication, we assumed that cyberlearning resources were available on the Internet and tried to automatically retrieve, associate, and rank resources based on their “importance” for each scientific topic and publication. In other words, we conceptualized the problem of creating cyberlearning referential metadata as an information retrieval problem that amenable to automation using meta-search and retrieval algorithms. The retrieved resources were ranked based on the content relevance (language model), topic relevance (labeled LDA inference), and a combination of the two factors (topic probability as language model prior).

In order to test our method of automatic cyberlearning referential metadata generation, we designed an experiment in which a group of graduate students learn and understand the essence of randomly sampled research topics and publications in the information retrieval domain through automatically generated cyberlearning resources. Evaluation results show that automatically generated cyberlearning resources via retrieval and meta-search can effectively help students to understand the essence of scientific topics and publications. In addition, we utilized student feedback to validate the
ranking algorithms designed to prioritize informative resources for the target scientific topics and publications.

In the remainder of this paper we (1) review relevant literature and methodologies, (2) introduce our referential metadata generation methodology, (3) describe our experiment in the information retrieval domain with a group of graduate students with respect to both sampled topics (keywords) and publications, (4) evaluate the cyberlearning resources referential metadata creation and resource ranking algorithms, and (5) discuss the contributions and limitations of our work.

Previous Research

Referential metadata (Liu, Chen & Qin, 2011) in the broadest sense include not only citations but also data about other types of scholarly output that is based on or related to the same publication. It is common today that before a paper is published in a journal, the authors of the paper may have presented it as a conference poster and/or a conference paper which produces presentation files or videos, or have made datasets, source code, or related materials available on the project website. These precursors and artifacts of a publication establish a context as well as provenance for readers to understand, evaluate, and interpret the research reported in the publication. While referential metadata are valuable for information retrieval and use, they are not usually included in publication metadata records.

Cyberlearning resources, as a specific kind of referential metadata, are highly important for e-learning environment. In this research we focused on creating referential metadata for some specific scientific topics and publication. In this section we review previous efforts on scientific metadata creation for cyberlearning resources.

Cyberlearning Resources and E-Learning Challenges

The proliferation of cyberinfrastructure and resources calls for more powerful and effective metadata representation methodologies to address information discovery and e-learning challenges. Referential cyberlearning metadata, as an emerging effort devoted to providing scientific topic- and publication-rich web context, is necessitated by the exponential growth of online open resources. The commitment of researchers in this field to education and e-learning should not be ignored.

Nevertheless, online open resources typically lack clear quality assurances. This is now recognized as a major concern with online open resources (D’Antoni, 2009). Not until fairly recently have researchers used multimedia Web resources, such as videos, audios and images, as an effective means of supporting student learning. For example, DeLeng, Dolmans, & van de Wiel (DeLeng, Dolmans, & Wiel, 2007) examined students’ views on the added value for problem-based learning (PBL) of using video resources in contrast with exclusively text-based approaches during the pre-clinical phase of undergraduate medical education. In the experiment with undergraduate students, they found that videos were generally perceived as a valuable stimulus for group discussions in PBL. Similarly, Maniar et al. (2008) examined the possibility of using mobile phones for video-based learning, a.k.a. m-learning.

Persson, Fyrenius, & Bergdahl (2010) used multimedia resources to enhance problem-based learning across the entire curriculum, making education more realistic and thereby more motivating and stimulating for students. Furthermore, Agazio and Buckley (2009) used YouTube for nursing education. In their research, YouTube was used to illustrate theoretical content, involve students, and inspire innovative teaching methods. Videos were presented on YouTube to stimulate student discussion, share information, and create a learning community. Duffy (2008) similarly, investigated the possibility of using YouTube, Podcasting, Blogs, Wikis and RSS to create a ubiquitous user-centric, user-content generated and user guided learning experience.

Typically, however, it is extremely costly to generate sophisticated cyberlearning resources for specific scientific topics and publications. Consequently, there is a need for a more effective and efficient set of tools and methods to create, associate, and manage online cyberlearning resources. To the best of our knowledge, there are few studies focusing on the task of bridging the gap between existing resources and scientific topics and publications.
Scientific Metadata for E-Learning

Metadata has traditionally been used for finding, identifying, selecting, and obtaining information objects. In its short history, however, metadata research has split into two camps with different perspectives and paradigms: 1. The description paradigm found in library and information science; 2. The processability and executability paradigm rooted in computer science (Zeng & Qin, 2008). Research on metadata representation and generation over the last few decades has drawn techniques and methods from a wide variety of research fields, including natural language processing and machine learning.

Obviously, for most existing library or document repository systems, professional metadata creators or domain experts (e.g., catalogers and indexers) are the ideal candidates (Milstead & Feldman, 1999) to create metadata, as they are familiar with the systems and terminologies. However, this approach is costly and may be limited in availability. It is hard to apply this approach to large amounts of data across different domains. Other researchers, for instance, Greenberg et al. (Greenberg, Pattueli, & Parsia, 2001), have found that authors can sometimes provide higher quality metadata for web resources. This approach is adopted by most digital libraries. Nevertheless, many authors are only willing to provide relatively simple descriptive metadata. In the medical domain, researchers have found that explicit structural abstracts are not entirely reliable (Demner-Fushman & Lin, 2007).

For these reasons, user- or author-generated and professional- or expert-generated metadata can hardly cope with the need for complex referential metadata generation at a large scale across different domains. Accordingly, the automatic approach, an economical and effective alternative, has become popular over the past few years. A wide variety of techniques have been used to process digital texts to generate metadata records. Some existing meta-search engines, such like NECI (Lawrence & Giles, 1998), SavvySearch (Dreiligner & Howe, 1996), and MetaCrawler (Selberg & Etzioni, 1995), have proved meta-search a successful technology for enhancing the user search experience. Unfortunately, generating metadata for cyberlearning resources for research topics and publications that were neither directly created by authors or publishers nor explicitly referenced in the content of the papers is a very demanding task.

Tremendous efforts have been made to improve the quality and efficiency of search engines over the years. However, relying on only one of them is insufficient. Meta-search is an information retrieval method that sends queries to multiple search engines or digital libraries simultaneously and aggregates the results into a single list. The basic assumptions are that a single search engine can only index a small portion of the web, and that aggregated retrieved results are likely to be more comprehensive. Some existing meta-search engines, such like NECI (Lawrence & Giles, 1998), SavvySearch (Dreiligner & Howe, 1996), and MetaCrawler (Selberg & Etzioni, 1995), have proved meta-search a successful technology for enhancing the user search experience.

Cyberlearning Referential Metadata Creation

As domain knowledge in most disciplines expands at a frenetic pace, disconnected research artifacts and other resources need to be connected in innovative ways, as through more effective use of metadata. In this research referential metadata were generated for six different types of cyberlearning resources--Wikipedia pages, video lectures, (presentation) slides, datasets, (online) tutorials, and source code--to help students and scholars better understand the scientific topic and publication.

In this section, we will explain our approach through three individual questions step by step: 1) How to collect different types of resources for a certain topic; 2) How to perform topic modeling on publications; 3) After gathering resources and differentiating topics, how to associate resources with each publication by using topics as intermediary and rank these resources.

Collecting Resources for Topics

In this research, we treated author-assigned keywords as a representation of each scientific research topic in a given domain of study, namely, information retrieval. In order to generate referential metadata for each topic effectively, information retrieval and meta-search approaches were applied, in which a Boolean query was sent to one or more search engines for each scientific keyword and
cyberlearning resource type. The detailed query for each resource type is listed in Table 1. The first column is the cyberlearning resource type. For each type the target query was sent to one or more search engines. For example, in order to get tutorial resources for a keyword, the query "[Keyword] AND Tutorial NOT Video" (column 3) was sent to Google (column 2), where [Keyword] was replaced with the target keyword content. Similarly, for video resources, the query "[Keyword]" was sent to YouTube1, TED2, and Videolecture3. For slides and source code, different queries were sent to different search engines. For example, for slides the queries "[Keyword] AND (filetype:ppt OR slides)" and "[keyword]" were sent to Google and Slideshare4, respectively. For performance reasons, we indexed the Wikipedia page dump locally.

Table 1
Meta-search query for different kind of resources

<table>
<thead>
<tr>
<th>Resource type</th>
<th>Search engines</th>
<th>Query (Keyword task)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia page</td>
<td>Wikipedia dump (local database)</td>
<td>[Keyword]</td>
</tr>
<tr>
<td>Tutorial</td>
<td>Google</td>
<td>[Keyword] AND Tutorial NOT Video</td>
</tr>
<tr>
<td>Slides</td>
<td>Google, Slideshare</td>
<td>Google: [Keyword] AND (filetype:ppt OR slides) Slideshare: [Keyword]</td>
</tr>
<tr>
<td>Video</td>
<td>YouTube, TED, Videolecture</td>
<td>[Keyword]</td>
</tr>
<tr>
<td>Dataset</td>
<td>Google</td>
<td>[Keyword] AND Dataset</td>
</tr>
<tr>
<td>Source code</td>
<td>Google, SourceForge</td>
<td>Google: [Keyword] AND (Source Code OR Toolkit OR Java OR C++ OR Python) SourceForge: [Keyword]</td>
</tr>
</tbody>
</table>

Based on the Table 1, a list of queries was sent to different search engines to retrieve candidate resources. In our experiment, we used the top 15 retrieved results from each search engine to aggregate the final result collection for each resource category. In most cases, the result collection was a combination of informative resources and noisy results. Experience in information retrieval reveals that since different search engines return very diverse and sparse results for the same query or for a similar one, irrelevant data may pollute the search results and mislead users. For instance, as the following diagram shows, for the topic labeled “Question Answering,” if we use the topic label as our query, two Wikipedia pages get high content relevance scores (i.e. BM25 or TF-IDF). However, users are only interested in the first one, “Open domain question answering,” for this scientific topic, while the second one, “Question and Answer (album),” should be removed or ranked lower as a noisy resource. Current information retrieval ranking methods based on bag-of-words, like language model and BM25, can hardly detect the topic level match.

To address this problem, we used the topic modeling algorithm to generate the word probability distribution for each scientific topic (keyword). For each scientific keyword, key, a topic-word distribution P(word|key) needs to be trained from scientific literature. Then, we can enhance the ranking algorithm to prioritize those informative resources for the target scientific keywords by using resource topic prior probability. In more detail, if we use P(resource|key) to rank each candidate resource:

\[
P(resource|key) = \frac{P(key|resource) \cdot P_{prior}(resource)}{P(key)}
\]
Figure 2. Statistical match ≠ topic match (for resource)

where $\text{key}_i$ is the keyword string (topic label), and $P(\text{key}_i|\text{resource})$ is the language model matching score, the likelihood of the keyword given the resource content. $P(\text{resource})$ in this formula is the resource topic prior probability:

$$P_{\text{prior}}(\text{resource}) = P(z_{\text{key}_i}|\text{resource})$$

which is the topic $z_{\text{key}_i}$ inference probability score given the resource content. Unlike the keyword string based language model, $P(z_{\text{key}_i}|\text{resource})$ employed all the possible terms in the resource to “vote” for the topic level match as the resource prior. Then, for Figure 2 case, the second Wikipedia page, about the music album, will get a very low topic score, resulting in a low rank.

Ideally, we should also consider search engine rank as a kind of prior. However, unlike other meta-search problems, the ranking lists of some cyberlearning types in this research were totally disjointed. For example, for videos we sent queries to YouTube, TED, and Videolecture. Their indexes are almost disjoint, a feature that some meta-search ranking fusion algorithms do not appreciate, e.g., Borda’s method and Markov chain methods (Dwork, Kumar, Naor & Sivakumar, 2001).

**Topic Modeling on Publications**

Blei et al., (2003) proposed Latent Dirichlet Allocation (LDA) as a promising unsupervised topic modeling algorithm. LDA employs a generative probabilistic model in the hierarchical Bayesian framework, and extends PLSI by introducing a Dirichlet prior on $\theta$. As a conjugate prior for the multinomial topic distribution, the Dirichlet distribution assumption has some advantages, including simplification of the problem. The probability density of a T-dimensional Dirichlet distribution over the multinomial distribution $p = (p_1, p_2, ..., p_T)$, where $\sum \alpha_j = 1$, is defined by:

$$\text{Dir}(\alpha_1, \alpha_2, ..., \alpha_T) = \frac{\Gamma(\sum \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

where $\alpha_1, \alpha_2, ..., \alpha_T$ are parameters of the Dirichlet distribution. These parameters can be simplified to a single value $\alpha_{\text{LDA}}$, the value of which is dependent on the number of topics.
However, one limitation of LDA is the challenge of interpreting and evaluating the statistical topics. For example, in this research, it is hard to assign a keyword to each statistical topic automatically. In addition, arbitrary numbers of topic may not be appropriate for this study because, while some topics may be very sparse (covering several keywords), others may focus only on quite detailed knowledge of the same scientific topic (covering part of a keyword). These limitations led us to use a supervised or semi-supervised topic modeling algorithm, one stem from LDA, which employs existing scientific keywords as the topic labels.

In this research, we assume that each (author-assigned) scientific keyword is a topic label and that each scientific publication is a mixture of its author-assigned topics (keywords). As a result, both topic labels and topic numbers (the total number of keywords in the metadata repository) are given. The labeled LDA (LLDA) (Ramage, Hall, Nallapati, & Manning, 2009) was used in training the labeled topic model. Unlike the LDA method, LLDA is a supervised topic modeling algorithm that assumes the availability of topic labels (keywords) and the characterization of each topic by a multinomial distribution $\beta_{\text{key},i}$, over all vocabulary words. For example, the following table is an example of the keyword–word topic probability:

<table>
<thead>
<tr>
<th>Search engine</th>
<th>Semantic web</th>
<th>Directed graph</th>
<th>Image retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>google</td>
<td>0.0173</td>
<td>0.0339</td>
<td>0.0151</td>
</tr>
<tr>
<td>log</td>
<td>0.0116</td>
<td>0.0257</td>
<td>cycle 0.0060</td>
</tr>
<tr>
<td>site</td>
<td>0.0037</td>
<td>0.0109</td>
<td>flow 0.0046</td>
</tr>
<tr>
<td>visit</td>
<td>0.0034</td>
<td>0.0102</td>
<td>minimum 0.0039</td>
</tr>
<tr>
<td>page</td>
<td>0.0011</td>
<td>0.0064</td>
<td>edges 0.0036</td>
</tr>
<tr>
<td>focused</td>
<td>0.0010</td>
<td>0.0061</td>
<td>node 0.0024</td>
</tr>
</tbody>
</table>

During the Bayesian generative topic modeling process, each word ($w$) in a publication is chosen from a word distribution associated with one of that paper's labels (keywords). The word is picked in proportion to the publication’s preference towards the associated label $\theta_{\text{paper.key},i}$ and the label's preference for the word $\beta_{\text{key},w}$. Figure 3 visualizes the LLDA generative process. For each topic (keyword) $\text{key}_{k}$, one draws a multinomial distribution $\beta_{\text{key},k}$ from the symmetric Dirichlet prior $\eta$. Then, for each publication, one builds a label set $\Lambda$ paper for the deterministic prior $\Phi$. Finally, one selects a multinomial distribution $\theta_{\text{paper}}$ over the labels $\Lambda$ paper from Dirichlet prior $\alpha$.

**Figure 3. LLDA Algorithm**

**Associating Resources with Publications**

For scientific publication metadata resource generation, the most straightforward method is to search for the paper title (exact match) in different search engines. However, this method has two major limitations. First, given a relatively long publication title, if we use exact string match, search engines will find very few results. Based on our experiment on 70 publications, on average only 0.35 resources were retrieved based on the queries in Table 1, replacing [Keyword] with [Paper Title]. We also found that the quality of resources was not good. A large portion consisted of publisher or digital library access pages for the target publication, which do not help scholars and students understand the paper. Second, the
cost of this method is quite high. Given the very large potential publication collection, we needed to send multiple queries for each publication. For example, based on Table 1, for each keyword or publication, we needed to send four queries to Google. However, most search engines restrict the number of automatic visits, which makes this method inefficient.

In this research we used a more economical method to cope with this problem. We assumed each publication was composed of a list of topics, with each topic represented by an author-assigned keyword. We then used the results from the previous section to estimate the publication resources. As Figure 4 shows, for each resource type, we first aggregated all resources retrieved using the publication keywords and then used a ranking algorithm to identify the most important resources based on publication content. Such resources are highly likely to be relevant to the publication content and could assist students or scholars to capture the essence of the publication. For this method, we assumed that improving the understanding of scholars or students on the topics (keywords) of the paper would eventually help them to understand the paper itself.

![Figure 4. Publication resources generation](image)

For all publication resources, we can use the language model to rank all the resources. Specifically, we used the language model with Dirichlet smoothing (Zhai & Lafferty, 2001) to rank the resource, the likelihood of paper content given a resource:

$$
logP(\text{paper}|\text{resource}) = \sum_{w_i \in \text{paper}} log \left( \frac{c(w_i; \text{resource}) + \mu P(w_i|C)}{|\text{resource}| + \mu} \right)
$$

Where the ranking score is the sum of log likelihood of each word $w_i \in \text{paper}$. The smoothing function used word resource frequency, $c(w_i; \text{resource})$, and the resource collection probability, $P(w_i|C)$. In practice, the paper can be represented by either publication title or abstract. When we compare this method with the task introduced previously in the resource collection section -- keyword relevance computation, we have stronger confidence in the language model usage, because keyword string are usually very short, only 2 or 3 tokens, but paper title and abstract provide much richer content, and the ranking performance should be much better. However, we also want to test and compare the topic prior performance in this study, so we define

$$P_{\text{prior}}(\text{resource}) = P(z_{\text{paper}}|\text{resource})$$

where $z_{\text{paper}}$ is a distribution of all the possible topics (mixture) of paper. Unlike LDA-based topic inference, we don’t have to project the paper onto the entire possible topic space, as the paper topic labels (publication keywords) are already available. Also, we assume the keyword is not equally important
for each publication; that is, some keywords are more specific than others, and may deserve higher “importance” scores. We used KF-IDF to calculate this importance, where:

\[ KF = \frac{C(key, paper)}{\log |paper|} \]

\[ IDF_{key_i} = \log \frac{N}{C(key_i, paper)} \]

As a variation on TFIDF, KF is the normalized keyword frequency in the paper’s title and abstract. IDF (inverse document frequency) has been used in IR as a measure of the general importance of a term in a collection. Similarly, we used IDF to assess keyword importance for our IR domain publication collection. In the IDF function, N is the total number of domain publications, and \( C(key_i, paper) \) is the number of publications with the target keyword, \( key_i \). We assume that if a keyword is rare in a collection (large IDF), this keyword could be more important for the paper. Then the resource (topic) prior can be characterized as:

\[ P(z_{paper}|resource) = \frac{\sum_{i=1}^{n} KFIDF_i \cdot P(z_{key_i}|resource)}{\sum_{i=1}^{n} KFIDF_i} \]

In this formula, the target paper has n keywords (topics), and the resource topic prior is the normalized sum of all the topic inference probability. The weight of each topic inference score is the KFIDF value.

**Experiments**

To test and compare different methods of cyberlearning resource metadata generation, we designed an experiment with information retrieval topics and publications. As we needed to judge whether a resource could effectively help scholars or students understand the essence of a scientific topic or publication, we invited graduate students with basic knowledge in information retrieval as volunteers for this experiment.

Unfortunately, some participants in this experiment did not have enough programming experience to provide judgments for source code resources. As a result, participants only evaluated five categories of resources in this experiment: Wikipedia pages, video lectures, tutorials, datasets, and slides.

**Dataset and Topic Modeling**

In this experiment, a total number of 20,799 information retrieval publications from 1965 to 2010 were used. We first used purposive sampling to identify 15 core conference proceedings and journals covering information retrieval, such as SIGIR, TOIS, CIKM, and others. Publications in these proceedings and journals were used as seed publications. Cited publications in these were then investigated to expand the corpus. If a paper was cited more than twice by these seed publications, we embody it into the test collection.

In the metadata repository, some publications did not have keyword metadata. To solve this problem, we first created a popular keyword (frequency > 3) list from the existing keywords in the test collection. We then searched for each keyword in the paper title and abstract by using greedy matching. For example, if “music information retrieval” was in the title, we didn’t use the keyword “information retrieval”. Matched keywords were used as “pseudo-keyword” metadata for the target publication if author-assigned keyword metadata was unavailable. The keyword and publication collections were used to calculate IDF_{keyword}.

A total of 7,293 publications were sampled for LLDA topic modeling. Author-provided keywords were selected as topic labels. For instance, this paper has six author-assigned keywords. Thus, our LLDA training would have assumed that this paper is a multinomial distribution over these six topics. During pre-processing we also clustered similar keywords if the edit distance between them was very small (as in “k-
means” and “k means”) or if two keywords shared the same stemmed root (as in “web searches” and “web search”).

The resource topics may or may not focus on the scientific topics we extracted from the papers. For instance, Figure 2 shows a Wikipedia page about the topic “music.” In order to model the “noisy” topics (distributions), 4,668 web pages, labeled with ODP categories, were randomly sampled from the ODP (Open Directory Project) 5 database. The top page categories were used as the noisy topic labels: Arts, Games, Kids and Teens, Reference, Shopping, Business, Health, News, Regional, Society, Home, Recreation, and Sports. Two categories, Computers and Science, were not used for noisy modeling, as they are related to information retrieval research.

If a keyword appeared less than 10 times in the selected publications, we removed it from the training topic space. For publication content we first used tokenization to extract words from the title, abstract, and publication full text. If the word contained fewer than three characters, this word was removed. Snowball stemming was then employed to extract the root of the target word. We also removed the most frequent 100 stemmed words and words appearing less than three times in the training collection.

Finally, we trained an LLDA model with 605 topics (594 scientific keywords + 11 ODP categories). These topics were used to infer the resource topic distribution, $P(z_{key} | resource)$. We then sampled 70 publications from SIGIR and CIKM conference proceedings for evaluation (this was the “publication task”). A total number of 401 keywords were contained in these publications. Finally, we sampled 45 keywords for the topic or keyword evaluation task (the “keyword task”).

**Experimental Setting**

As stated in the previous section, two tasks need to be performed in this experiment: a topic (keyword) evaluation task, and a publication evaluation task.

For topic evaluation, we sampled 45 keywords from the collection of 401. For each keyword, we designed a web evaluation page. Likewise, we designed an evaluation for each publication. On each evaluation page, the keyword or publication title and abstract were presented at the top, and the top cyberlearning resources that had been retrieved were offered with actual resource links. Users could access these resources via a pop-up window by clicking the hyperlink. Reading (or watching) and understanding these resources could be a time-consuming job for users, especially for the publication task, since participants needed to read and understand the topics of the paper first. As a result, we only offered users the top five resources for each resource type for each keyword page (hence, there were approximately 25 resources on each page), and the top three resources for each resource type for each publication page (hence, approximately 15 resources on each page). Each user was asked to evaluate ten publications and seven keywords. At this stage, we didn’t have a chance to find the best ranking method so we used the tentative method, language model without topic prior, for resource ranking for both keyword and publication tasks.

Seven graduate (masters or doctoral) students with basic knowledge in information retrieval participated in this evaluation. Each student can evaluate resource quality by using a dropdown menu. In this evaluation, students were requested to judge whether the cyberlearning resource could help them understand the essence of the scientific keyword or publication. Possible evaluations were “not relevant” (score of 0), “low relevance” (1), “good relevance” (2), and “high relevance” (3). Before performing this evaluation task, we trained participants with examples of helpful and unhelpful resources. The small number of participants is a major limitation for this research, and we will address it in the future work section.

---

5 http://www.dmoz.org/
Experimental Results

In this section, we report two kinds of experimental results based on analyzing students’ judgments of relevance. First, we evaluate whether the automatically generated cyberlearning resource referential metadata helped students to better understand the essence of scientific topics and publications by calculating mean average precision (MAP) and mean reciprocal rank (MRR). MAP calculates the average percentage (precision) of the “good relevance” (2), and “high relevance” (3) in all the listed resources, and MRR measures the average rank of the first “good relevance” or “high relevance” in the ranking list. Higher MAP and MRR scores indicate that users are more likely to get the useful resources and to understand the essence of the topic and publication.

Second, we evaluate the resource-ranking algorithm(s) for both keyword and publication tasks. In this research, we used normalized discount cumulative gain (NDCG) (Järvelin & Kekäläinen, 2002) to validate the ranking algorithms. NDCG estimates the cumulative relevance gain the user receives by examining retrieval results up to a given rank on the list. NDCG is based on two assumptions: first, highly relevant documents are more valuable than marginally relevant documents (graded relevance); second, the lower the ranked position of a relevant document, the less valuable it is for the user. A ranked vector $V$ of results $[\text{label}(v_1), \text{label}(v_2), ..., \text{label}(v_n)]$ can be generated for each query $q$, where each item in the vector is a judgment of degree of relevance. In this research, 1 is not relevant and 4 is perfect relevance. With this vector, calculation of the discount cumulative gain (DCG) is possible:

$$DCG@k(V) = \sum_{i=1}^{k} \frac{1}{log_2(1+i)} (\text{label}(v_i) - 1)$$

The normalized DCG (NDCG) of $V$ is defined as the DCG vector divided by the ideal permutation of $V$. A perfect ranking algorithm returns $DCG@k(V) = \text{ideal DCG@k}(V)$ and NDCG score = 1. All NDCG calculations are then relative values on the interval 0 to 1.

Keyword Task Results

For the keyword task, the result of each kind of resource ranking result is presented in Table 3 and Figure 5. The best-performed number for each row is highlighted in the table. LM used language model and Dirichlet smoothing with keyword string as query. LLDA used topic inference score, $P(z_{key}[\text{resource}])$, for ranking. LM + LLDA used topic inference score as the resource prior, and language model + Dirichlet smoothing for ranking.

Figure 5. Keyword Task resource quality chart (MAP, MRR, and nDCG)
Table 3
Resource quality for the keyword task

<table>
<thead>
<tr>
<th></th>
<th>LM</th>
<th>LLDA</th>
<th>LM+LLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>MAP</td>
<td>NDCG</td>
<td>MRR</td>
</tr>
<tr>
<td>Wiki</td>
<td>0.7576</td>
<td>0.7820</td>
<td>0.8248</td>
</tr>
<tr>
<td>Wiki</td>
<td>0.7820</td>
<td>0.8088</td>
<td>0.8333</td>
</tr>
<tr>
<td>Wiki</td>
<td>0.7639</td>
<td>0.8070</td>
<td>0.8241</td>
</tr>
<tr>
<td>Wiki</td>
<td>0.7753</td>
<td>0.7361</td>
<td>0.7879</td>
</tr>
<tr>
<td>Wiki</td>
<td>0.8582</td>
<td>0.8594</td>
<td>0.8608</td>
</tr>
<tr>
<td>Wiki</td>
<td>0.7904</td>
<td>0.7510</td>
<td>0.7677</td>
</tr>
<tr>
<td>Video</td>
<td>MAP</td>
<td>NDCG</td>
<td>MRR</td>
</tr>
<tr>
<td>Video</td>
<td>0.7156</td>
<td>0.7667</td>
<td>0.6865</td>
</tr>
<tr>
<td>Video</td>
<td>0.7835</td>
<td>0.8199</td>
<td>0.7638</td>
</tr>
<tr>
<td>Video</td>
<td>0.7539</td>
<td>0.8012</td>
<td>0.7054</td>
</tr>
<tr>
<td>Video</td>
<td>0.7002</td>
<td>0.7410</td>
<td>0.6979</td>
</tr>
<tr>
<td>Video</td>
<td>0.7698</td>
<td>0.8120</td>
<td>0.7803</td>
</tr>
<tr>
<td>Video</td>
<td>0.7413</td>
<td>0.7822</td>
<td>0.7205</td>
</tr>
<tr>
<td>Video</td>
<td>0.6243</td>
<td>0.6504</td>
<td>0.5974</td>
</tr>
<tr>
<td>Dataset</td>
<td>MAP</td>
<td>NDCG</td>
<td>MRR</td>
</tr>
<tr>
<td>Dataset</td>
<td>0.7959</td>
<td>0.8204</td>
<td>0.7846</td>
</tr>
<tr>
<td>Dataset</td>
<td>0.6667</td>
<td>0.6905</td>
<td>0.6381</td>
</tr>
<tr>
<td>Dataset</td>
<td>0.6771</td>
<td>0.7236</td>
<td>0.6831</td>
</tr>
<tr>
<td>Dataset</td>
<td>0.6613</td>
<td>0.7151</td>
<td>0.6835</td>
</tr>
<tr>
<td>Dataset</td>
<td>0.7300</td>
<td>0.8285</td>
<td>0.7372</td>
</tr>
</tbody>
</table>

Overall, the average MAP across different resources was 0.6831 and the average MRR was 0.7372, which means most students in this evaluation believed that the cyberlearning resources (which were automatically generated via meta-search) could help them understand the keyword-based scientific topic, and that the quality of those highly ranked resources was good. Among different resource types, Wikipedia pages and video lectures performed better than datasets and slides, and datasets performed the worst.

Form a ranking perspective, NDCG suggests the ranking’s normalized discount cumulative gain, which is more sensitive to the degree of relevance (or resource usefulness). Clearly, considering topic prior will boost the ranking algorithm via prioritizing those useful resources based on user feedback.

We also found that, for wiki and video resources, LM + LLDA outperforms LLDA. However, for all other types of resources, LLDA is better than LM + LLDA. The main reason for this is the quality of the resource content. For wiki resources, we used Wikipedia dump, and the quality for each wiki page is very high. Similarly, all video results come from specific websites, such YouTube and Videolecture, which enable us to use regular expression to extract the video tags and description from HTML code. As a result, textual representation for a video resource is also accurate. However, for all other resource types, the results were mainly from Google, and we didn’t have an opportunity to identify the essential content from the HTML code because of the variations of web page structures. Subsequently, the content of HTML used for resource representation could be so noisy that the performance of the language model will be threatened. We will address this problem in the future work section.

Publication Task Results

Table 4 and Figure 6 shows the results of five different ranking methods for the publication task. LM Title (Title + Prior in Figure 6.) used the paper title as the query to search the content of the resource with language model combined with Dirichlet smoothing, and LM Abstract (Abstract + Prior in Figure 6.) used the paper abstract as query. LLDA (LLDA only) used the paper and topic match probability, \( p(z_{\text{paper}}|r_{\text{resource}}) \), for ranking proposed in the “Associating resources with publication” section, and we also considered KF-IDF weighting. Title + LLDA (Title) used topic match probability as resource prior probability and title based language model for ranking. Abstract + LLDA (Abstract) used topic prior with abstract based language model.
Overall, in terms of ranking, LM Abstract performs the best, and LLDA-based topic modeling as resource prior did not provide stable performance gain. LLDA alone worked worst. The main reason for this could be the quality of query. Compared with the keyword task, the publication task’s query length is much larger and query quality could be much better. For instance, the average length of a keyword query was 2.02 tokens, while the average title query was 8.86 tokens and the average abstract query was 232.75 tokens. It is clear that a long query contains much richer information (that is, a long query also contains the topic information itself), and the language model-based content match itself works well. This is the reason why LM Abstract also outperforms LM Title.

In terms of resource quality, the MAP and MRR in Table 4 illustrate that the performance of the publication task is lower than that of the keyword task, but students usually suggest that the highly ranked resources could help them understand the essence of the target publication. For instance, the MRR for all resources was 0.7187, which means students find the first or second recommended resource on the ranking list is helpful or very helpful. Similarly, datasets performed the worst while Wikipedia Pages, Video Lectures, and Slides are more useful for participants.

Table 4:
Resource quality for the publication task

<table>
<thead>
<tr>
<th></th>
<th>LM Title</th>
<th>LM Abstract</th>
<th>LLDA</th>
<th>Title + LLDA</th>
<th>Abstract + LLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>MAP 0.6042</td>
<td>0.6250</td>
<td>0.5125</td>
<td>0.5375</td>
<td>0.5021</td>
</tr>
<tr>
<td></td>
<td>NDCG 0.8891</td>
<td>0.9033</td>
<td>0.8186</td>
<td>0.8643</td>
<td>0.8279</td>
</tr>
<tr>
<td></td>
<td>MRR 0.6083</td>
<td>0.6375</td>
<td>0.5208</td>
<td>0.5458</td>
<td>0.5042</td>
</tr>
<tr>
<td>Video</td>
<td>MAP 0.7019</td>
<td>0.6704</td>
<td>0.5926</td>
<td><strong>0.7185</strong></td>
<td>0.7005</td>
</tr>
<tr>
<td></td>
<td>NDCG 0.8733</td>
<td>0.8546</td>
<td>0.7688</td>
<td><strong>0.8896</strong></td>
<td>0.8780</td>
</tr>
<tr>
<td></td>
<td>MRR 0.7111</td>
<td>0.6778</td>
<td>0.6667</td>
<td><strong>0.7185</strong></td>
<td>0.7037</td>
</tr>
<tr>
<td>Slides</td>
<td>MAP 0.6156</td>
<td>0.6361</td>
<td>0.5001</td>
<td>0.6190</td>
<td>0.6139</td>
</tr>
<tr>
<td></td>
<td>NDCG 0.8313</td>
<td>0.8427</td>
<td>0.5982</td>
<td><strong>0.8513</strong></td>
<td>0.8362</td>
</tr>
<tr>
<td></td>
<td>MRR 0.6122</td>
<td><strong>0.6463</strong></td>
<td>0.5001</td>
<td>0.6259</td>
<td>0.6156</td>
</tr>
<tr>
<td>Tutorial</td>
<td>MAP 0.6007</td>
<td><strong>0.6163</strong></td>
<td>0.4201</td>
<td>0.5932</td>
<td>0.588</td>
</tr>
<tr>
<td></td>
<td>NDCG 0.8393</td>
<td>0.8374</td>
<td>0.5465</td>
<td>0.8128</td>
<td>0.8103</td>
</tr>
<tr>
<td></td>
<td>MRR 0.6215</td>
<td><strong>0.6319</strong></td>
<td>0.6423</td>
<td>0.6215</td>
<td>0.6111</td>
</tr>
<tr>
<td>Dataset</td>
<td>MAP 0.5474</td>
<td><strong>0.5779</strong></td>
<td>0.4116</td>
<td>0.5166</td>
<td>0.5048</td>
</tr>
<tr>
<td></td>
<td>NDCG 0.8045</td>
<td>0.7879</td>
<td>0.6198</td>
<td>0.7722</td>
<td>0.7657</td>
</tr>
<tr>
<td></td>
<td>MRR 0.5725</td>
<td><strong>0.6014</strong></td>
<td>0.4659</td>
<td>0.5362</td>
<td>0.5217</td>
</tr>
<tr>
<td>ALL</td>
<td>MAP 0.5991</td>
<td><strong>0.5904</strong></td>
<td>0.4694</td>
<td>0.5557</td>
<td>0.5494</td>
</tr>
<tr>
<td></td>
<td>NDCG 0.6772</td>
<td>0.6607</td>
<td>0.5507</td>
<td>0.6607</td>
<td>0.6434</td>
</tr>
<tr>
<td></td>
<td>MRR 0.7172</td>
<td><strong>0.7187</strong></td>
<td>0.6631</td>
<td>0.6847</td>
<td>0.6595</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

This research focused on the problem of generating cyberlearning referential metadata for scientific keywords and publications. It was conceptualized as an information retrieval and use problem amenable to automation using meta-search methodologies.

Based on preliminary evaluation, we found that automatically generated cyberlearning resources can help students better understand the essence of the scientific topic and publication. Meanwhile, the cost of this approach is very low. The ranking algorithms based on sophisticated content match (language model) and topic modeling (LLDA) can also effectively enhance the quality of resource ranking. For the keyword task, given short keyword queries, topic modeling can effectively help students better locate the most informative resources and remove the noisy ones. The publication task, on the contrary, used paper title and abstract as queries, which quality is much better, and topic modeling in this task couldn’t provide stable performance gain.

We also found that some types of resources are more helpful than the others. For instance, most students favored Wikipedia pages, video lectures, and presentation slides, instead of datasets.

There are two major limitations for this paper. First, we only recruited seven graduate students for the evaluation, which is mainly because of the cost and requirements of the tasks. For instance, in order to participate in this task, students should understand the basic concepts in information retrieval, and they are required to read (or watch) a large number of publications and resources for evaluation. In the future, we should generalize this experiment to other scientific domains, while finding a larger number of participants.

Second, the quality of the resource text is still questionable. Right now, we find Wiki and video performance are better than other types, which may be because of the quality of text. In the future, we should use some other algorithms to identify the important content from the HTML code for all resource pages, which could further improve the ranking performance.

References


287

Liu, X., (2012). Generating Metadata for Cyberlearning Resource through Information Retrieval and Meta-search, Journal of the American Society for Information Science and Technology (This is a preprint of an article accepted for publication in Journal of the American Society for Information Science and Technology copyright 2012)


