Prior Scholarly Works Regarding Citation Recommendation and Prediction Systems

1. McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Riedl, J. (2002). *On the recommending of citations for research papers*. Paper presented at the Proceedings of the 2002 ACM conference on Computer supported cooperative work.

Abstract

Collaborative filtering has proven to be valuable for recommending items in many different domains. In this paper, we explore the use of collaborative filtering to recommend research papers, using the citation web between papers to create the ratings matrix. Specifically, we tested the ability of collaborative filtering to recommend citations that would be suitable additional references for a target research paper. We investigated six algorithms for selecting citations, evaluating them through offline experiments against a database of over 186,000 research papers contained in ResearchIndex. We also performed an online experiment with over 120 users to gauge user opinion of the effectiveness of the algorithms and of the utility of such recommendations for common research tasks. We found large differences in the accuracy of the algorithms in the offline experiment, especially when balanced for coverage. In the online experiment, users felt they received quality recommendations, and were enthusiastic about the idea of receiving recommendations in this domain.

Summary

Proposal: Global recommendation of additional citations in a research papers using collaborative filtering (CF).
Database: Papers contained in ResearchIndex.
Database Size: 186,000 documents.
Algorithms: CF algorithms using k-nearest neighbor, item-based algorithms and Bayesian Networks. CF: co-citation matching, user-item CF, item-item CF, Naive Bayes Classifier. Non CF: Localized Citation Graph Search, Keyword Search ('Google' Baseline).
Year: 2002
Validation: Two validations: First, citations from a test dataset were removed and then attempt to predict. The authors found the percentage of citations predicted varies with the algorithm from 30% bayesian to 80% graph search. Second, 120 users evaluated the effectiveness of the algorithms and the utility of such recommendation. Results vary; 80% found helpful recommendations by the graph engine.
Additional comments: Very popular paper with 289 citations. They designed four ways to apply CF, and 2 non CF algorithms to the domain of Citation Recommendation and validate them by measuring users experiences.

2. Strohman, T., Croft, W. B., & Jensen, D. (2007). *Recommending citations for academic papers*. Paper presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.

Abstract

Substantial effort is wasted in scientific circles by researchers who rediscover ideas that have already been published in the literature. This problem has been alleviated somewhat by the availability of recent academic work online. However, the kinds of text search systems in popular use today are poor at handling vocabulary mismatch, so a researcher must know the words used in relevant documents in order to find them. This makes serendipitous results unlikely. We approach the problem of literature search by considering an unpublished manuscript as a query to a search system. With this approach, the entire text content of the paper can be used in the search process. We use the text of previous literature as well as the citation graph that connects it to find relevant related material. We evaluate our technique with manual and automatic evaluation methods, and find an order of magnitude improvement in mean average precision as compared to a text similarity baseline.

Summary

Proposal: Recommend citations given a two page text using graph-based features in the retrieval process.
Database: Rexa database.
Database Size: Total paper entries: 964,977. Papers with text: 105,601. Total number of citations (X cites Y): 1.46 million. Total number of cited papers 675,372.
Algorithms/Method: System of 2 stages: first return the top 100 documents (R); second, all papers cited in R are added to R generating a DB of approximately 3000 papers. Finally, papers are ranked using the following features: Publication Year, Text Similarity (multinomial diffusion kernel), Co-citation Coupling (citation among papers within the DB), Same Author, Katz (graph distance measure), Citation Count.
Year: 2007
Validation/Evaluation: 1000 documents from the Rexa collection used as sample queries. Authors reported 0.06 to 0.1 mean average precision metric. depending on features considered.
Additional comments: Short paper, 2 pages.

3. Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). *Joint latent topic models for text and citations*. Paper presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

Abstract

In this work, we address the problem of joint modeling of text and citations in the topic modeling framework. We present two different models called the Pairwise-Link-LDA and the Link-PLSA-LDA models. The Pairwise-Link-LDA model combines the ideas of LDA [4] and Mixed Membership Block Stochastic Models [1] and allows modeling arbitrary link structure. However, the model is computationally expensive, since it involves modeling the presence or absence of a citation (link) between every pair of documents. The second model solves this problem by assuming that the link structure is a bipartite graph. As the name indicates, Link-PLSA-LDA model combines the LDA and PLSA models into a single graphical model. Our experiments on a subset of Citeseer data show that both these models are able to predict unseen data better than the baseline model of Erosheva and Lafferty [8], by capturing the notion of topical similarity between the contents of the cited and citing documents. Our experiments on two different data sets on the link prediction task show that the Link-PLSA-LDA model performs the best on the citation prediction task, while also remaining highly scalable. In addition, we also present some interesting visualizations generated by each of the models.

Summary

Proposal: Develop two models (Pairwise Link-LDA and Link-PLSA-LDA) that jointly model topic and citations.
Database: Two data sources: scientific literature from CiteSeer containing citations, and blog data containing hyperlinks.
Database Size: From 3312 documents in CiteSeer, 763 documents were selected after processing. From blog data, 2248 postings with at least 2 outgoing links each and 1,777 documents with at least two incoming links each were selected.
Algorithms/Method: Pairwise Link-LDA: The authors combined the Latent Dirichlet Allocation (LDA) model with the Mixed Membership Stochastic Block (MMSB) model. This model is computationally expensive, improved by assuming that the link structure is a bipartite graph. Link-PLSA(probabilistic latent semantic analysis)-LDA.
Year: 2008
Validation/Evaluation: The models were trained using part of each DB, then the model was tested using a test dataset. With parameters defined, the log-likelihood of citations was measured. Authors reported Link-PLSA-LDA performed better in terms of log-likelihood.
Additional comments: Deeply technical paper describing the algorithms.

4. Bethard, S., & Jurafsky, D. (2010). *Who should I cite: learning literature search models from citation behavior*. Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.

Abstract

Scientists depend on literature search to find prior work that is relevant to their research ideas. We introduce a retrieval model for literature search that incorporates a wide variety of factors important to researchers, and learns the weights of each of these factors by observing citation patterns. We introduce features like topical similarity and author behavioral patterns, and combine these with features from related work like citation count and recency of publication. We present an iterative process for learning weights for these features that alternates between retrieving articles with the current retrieval model, and updating model weights by training a supervised classifier on these articles. We propose a new task for evaluating the resulting retrieval models, where the retrieval system takes only an abstract as its input and must produce as output the list of references at the end of the abstract's article. We evaluate our model on a collection of journal, conference and workshop articles from the ACL Anthology Reference Corpus. Our model achieves a mean average precision of 28.7, a 12.8 point improvement over a term similarity baseline, and a significant improvement both over models using only features from related work and over models without our iterative learning.

Summary

Proposal: The authors build a retrieval system that takes abstracts as inputs and produces reference lists as output.
Database: ACL Anthology Reference Corpus (ACL-ARC).
Database Size: A set of 10,921 papers from computational linguistics workshops, conferences and journals.
Algorithms/Method: The model scores each document (article) *d* against the query (project idea) *q* using a weighted sum of feature scores. Features: Similar terms, Cited by others, Recency, Cited Using Similar Terms, Similar Topics, and Social Habits.
Year: 2010
Validation/Evaluation: The authors constructed a query by concatenating an article's title and abstract, had their retrieval model find relevant articles for this query, and compared the results against the actual references. 794 articles were used as testing. The model achieves a mean average precision of 28.7, a 12.8 point improvement over a term similarity baseline.
Additional comments: Very similar to what we are trying to do. We may need to get deep into it as it seems to be easily implemented in BlueMix environment.

5. Kataria, S., Mitra, P., & Bhatia, S. (2010). *Utilizing Context in Generative Bayesian Models for Linked Corpus*. Paper presented at the Aaai.

Abstract

In an interlinked corpus of documents, the context in which a citation appears provides extra information about the cited document. However, associating terms in the context to the cited document remains an open problem. We propose a novel document generation approach that statistically incorporates the context in which a document links to another document. We quantitatively show that the proposed generation scheme explains the linking phenomenon better than previous approaches. The context information along with the actual content of the document provides significant improvements over the previous approaches for various real world evaluation tasks such as link prediction and log-likelihood estimation on unseen content. The proposed method is more scalable to large collection of documents compared to the previous approaches.

Proposal: Global identification of topics for linked documents in a corpus using document text, citation information, and citation context.
Database: Subset of CiteSeer digital library and web-pages from webkb data set.
Database Size: 3312 documents across 6 research fields in CiteSeer; 2,877 computer science department web pages from webkb.
Algorithms: Bayesian Networks
Year: 2010
Validation: First analysis done by calculating log-likelihood on unseen text with 10-fold cross validation, and calculating the average log-likelihood. Shows significant improvement for CiteSeer data set using cite-PLSA-LDA model, which factors in citation context. Also shows improvement on webkd data set, but a less significant improvement (authors explain this is likely due to frequent links to project pages and academic pages in web sites without accompanying context). Second analysis performed by labeling citations that actually appear in documents and comparing to the ranking of the citation generated by the models. This analysis shows cite-PLSA-LDA model outperforms other evaluated models in both CiteSeer and webkd data sets.
Additional comments: General idea is that evaluating the context in which a citation appears together with the text of the query document and information about the cited document improves topic identification for the query document.

6. Yu, X., Gu, Q., Zhou, M., & Han, J. (2012). *Citation Prediction in Heterogeneous Bibliographic Networks*. Paper presented at the Sdm.

Abstract

To reveal information hiding in link space of bibliographical networks, link analysis has been studied from different perspectives in recent years. In this paper, we address a novel problem namely citation prediction, that is: given information about authors, topics, target publication venues as well as time of certain research paper, finding and predicting the citation relationship between a query paper and a set of previous papers. Considering the gigantic size of relevant papers, the loosely connected citation network structure as well as the highly skewed citation relation distribution, citation prediction is more challenging than other link prediction problems which have been studied before. By building a meta-path based prediction model on a topic discriminative search space, we here propose a two-phase citation probability learning approach, in order to predict citation relationship effectively and efficiently. Experiments are performed on real-world dataset with comprehensive measurements, which demonstrate that our framework has substantial advantages over commonly used link prediction approaches in predicting citation relations in bibliographical networks.

Summary

Proposal: Finding and predicting the citation relationship between a query paper and a set of previous papers by building a meta-path based prediction model on a topic discriminative search space, we here propose a two-phase citation probability learning approach.
Database: DBLP citation data set generated by (Tang & Zhang, 2009).
Database Size: 29,615 papers and 215,502 citation relations.
Algorithms/Method: The authors propose a data structure for capturing both document similarity and potential citation relationship, which they call discriminative term buckets. They use a meta path-based feature space to interpret structural information in citation prediction, and define citation probability within the scope of meta path-based feature space. The model scores each document (article) *d* against the query (project idea) *q* using a weighted sum of feature scores.
Year: 2012
Validation/Evaluation: The authors report improvement of the performance of baseline methods, and precision around 75% in training and testing datasets.
Additional comments: Very similar to what we are trying to do. Although the methodology used is quite hard to understand, we may need to get deep into it.

7. Ren, X., Liu, J., Yu, X., Khandelwal, U., Gu, Q., Wang, L., & Han, J. (2014). *ClusCite: Effective citation recommendation by information network-based clustering*. Paper presented at the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.

Abstract

Citation recommendation is an interesting but challenging research problem. Most existing studies assume that all papers adopt the same criterion and follow the same behavioral pattern in deciding relevance and authority of a paper. However, in reality, papers have distinct citation behavioral patterns when looking for different references, depending on paper content, authors and target venues. In this study, we investigate the problem in the context of heterogeneous bibliographic networks and propose a novel cluster-based citation recommendation framework, called ClusCite, which explores the principle that citations tend to be softly clustered into interest groups based on multiple types of relationships in the network. Therefore, we predict each query's citations based on related interest groups, each having its own model for paper authority and relevance. Specifically, we learn group memberships for objects and the significance of relevance features for each interest group, while also propagating relative authority between objects, by solving a joint optimization problem. Experiments on both DBLP and PubMed datasets demonstrate the power of the proposed approach, with 17.68% improvement in Recall@50 and 9.57% growth in MRR over the best performing baseline.

Summary

Proposal: Global identification of references relevant to paper by clustering citations into different interest group; generate recommendations based on input paper's interest group.
Database: Subsets of DBLP dataset and PubMed dataset.
Database Size: 137,298 papers from DBLP and 100,215 papers from PubMed.
Algorithms: Graph regularized co-clustering on heterogeneous bibliographic network. Parameters used for learning model: "group memberships for attribute objects; feature weights for interest groups; and object relative authority within each interest group." Also describes authors' "ClusCite" algorithm, which iteratively alternates between optimizing for interest group and optimizing for relative authority.
Year: 2014
Validation: Splits corpus into training data, validation data for parameter tuning, and testing data. Splits are based on publication year of documents in corpus. Compares effectiveness of ClusCite algorithm to other "state of the art" citation recommendation algorithms, including algorithms which evaluate only text, evaluate text and citations, or evaluate authority. Authors find that ClusCite outperforms all other methods. Authors also evaluate citations for various authors and publication venues in particular interest areas, and verify that authors and publications relating to particular interest groups are cited more frequently when document group contains more documents in the corresponding interest area.

Additional comments: Goal of system is to recommend a small number of highly relevant papers for any input document. Proposed system seeks to accomplish this goal by generating interests group and authority rankings for each interest group. Authority rankings factor in authors, publication venue, and relationship to other papers, assuming highly authoritative papers are published by high value authors in high value publication venues and related to other high value papers.

8. Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C. L. (2015). *A Neural Probabilistic Model for Context Based Citation Recommendation*. Paper presented at the Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

Abstract

Automatic citation recommendation can be very useful for authoring a paper and is an AI-complete problem due to the challenge of bridging the semantic gap between citation context and the cited paper. It is not always easy for knowledgeable researchers to give an accurate citation context for a cited paper or to find the right paper to cite given context. To help with this problem, we propose a novel neural probabilistic model that jointly learns the semantic representations of citation contexts and cited papers. The probability of citing a paper given a citation context is estimated by training a multi-layer neural network. We implement and evaluate our model on the entire CiteSeer dataset, which at the time of this work consists of 10,760,318 citation contexts from 1,017,457 papers. We show that the proposed model significantly outperforms other state-of-the-art models in recall, MAP, MRR, and nDCG.

Summary

Proposal: Learn a model of distributed semantic representations of words and cited documents in a dataset using a multi-layer neural network to estimate the probability of citing a document, based on a citation context. The focus is on local citation recommendation (i.e., recommending citations for a particular context rather than a list of citations for a work as a whole).
Database: CiteSeer dataset.
Database Size: 1,017,457 papers; 10,760,318 citation contexts.
Algorithms: Multi-layer neural network using distributed representations of words. Negative sampling and noise-contrastive estimation are used to build the model.
Year: 2015
Validation: Authors used two-fold validation to evaluate results, with documents crawled during or before 2011 as the training set and documents crawled after 2011 as the testing set. Authors deem their proposed model the “Neural Probabilistic Model.” Authors found statistically significant improvement over all other baseline models tested, which included Cite-PLSA-LDA, Restricted Boltzmann Machine, Citation Translation Model, and Word2vec Model. Authors further divided test set by number of citations to papers in data set. They found that their model was more accurate for papers that were cited less often, while all models performed similarly for papers cited very often. Based on this result, they assert their model requires fewer training examples to generate an accurate model.
Additional comments: The general idea here seems to be predicting citations based on the distribution of words that appear within a factor M of words contained in a citation context.

9. Zarrinkalam, F., & Kahani, M. (2012). *A multi-criteria hybrid citation recommendation system based on linked data*. Paper presented at the Computer and Knowledge Engineering (ICCKE), 2012 2nd International eConference on.

Abstract

Citation recommendation systems can help a researcher find works that are relevant to his field of interest. Currently, most approaches in citation recommendation are based on a closed-world view which is limited to using a single data source for recommendation. Such a limitation decreases quality of the recommendations since no single data source contains all required information about different aspects of the literature. This paper proposes a citation recommendation approach based on the open-world view provided by the emerging web of data. It uses multiple linked data sources to create a rich background data layer, and a combination of content-based and multi-criteria collaborative filtering as the recommendation algorithm. Experiments demonstrate that the proposed approach is sound and promising.

Summary

Proposal: Proposes recommendation system that "enriches" a data set by filling in gaps in metadata with data taken from other data sets, then recommends citations for input documents based on content analysis and meta-data criteria.
Database: CiteSeerX, ACM, DBLP, IEEE
Database Size: 30,000 publications from CiteSeerX collected; removed publications with no title and no abstract (about 30%); removed publications after 2007; final set of 12000 documents for background system (seems to be training data) and 600 documents for testing.
Algorithms: Content-based filtering and multi-criteria collaborative filtering.
Year: 2012
Validation: Using documents from CiteSeerX, authors compared 1) system using publication data "enriched" with information from ACM, DBLP, and IEEE and using both content-based filtering and multi-criteria collaborative filtering; 2) a system using content-based and multi-criteria collaborative filtering without the "enriched" publication data; and 3) a system using only content-based filtering. Results show method 2) is significantly more accurate than method 3), but method 1) does not show as great an improvement over method 2) as the authors anticipated. Authors explain that manual review showed linked data sources did not yet provide high quality data and had a lot of missing data, thus did not contribute as much information to initial data sources as authors anticipated.
Additional comments: Very high level, experimental system implemented in Java; less detail and focus on algorithms than other papers cited above.