Relation Regularized Subspace Recommending for Related Scientific Articles

Qing Zhang[1], Jianwu Li[*2], Zhiping Zhang[1], Li Wang[1]
[1] Institute of Scientific and Technical Information of China, Beijing, 100038, China
[2] Beijing Key Lab of Intelligent Information, School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, 100081, China
zhangq@istic.ac.cn, ljw@bit.edu.cn, {zhangzp, wangli}@istic.ac.cn

ABSTRACT

Recommending related scientific articles for a researcher is very important and useful in practice but also is full of challenges due to the latent complex semantic relations among scientific literatures. To deal with these challenges, this paper proposes a novel framework with link-missing data adaption, which casts the recommendation task to subspace embedding and similarity ranking problems. The relation regularized subspace in this framework is constructed via Relation Regularized Matrix Factorization (RRMF) for well modeling both content and link structure simultaneously. However, the link structure for an article is not always available in practical recommending. To solve this problem, we further propose two alternative approaches based on Latent Dirichlet Allocation (LDA) for link-missing articles recommendation as an extension of RRMF. Experiments on CiteSeer dataset demonstrate our method is more effective in comparison with some state-of-the-art approaches and is able to handle the link-missing case which the link-based methods never can fit.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–Retrieval models.

General Terms

Algorithms, Performance, Design, Experimentation, Theory.

Keywords

Related Scientific Articles, Recommendation, Regularized Matrix Factorization, Latent Dirichlet Allocation, Link-Missing Data.

1. INTRODUCTION

With the springing of huge publications in real world, how to alleviate the overload of the overwhelming scientific information and fully utilize this heritage of wisdom is an important and significant issue. In this paper, we focus on related scientific articles recommendation (RSAR), as a fundamental problem in various scientific recommendation tasks. To solve this problem, a number of recommendation techniques originally designed for commercial applications are introduced into the field of scientific recommending. However, some differences in nature are obvious between RSAR and the most of commercial applications, which will be taken into our consideration. Firstly, the users and items in RSAR are homogeneous, i.e., both are papers rather than people and goods (services). So, the attractive methods only based on rating information in commercial field are not fully competent. Secondly, the relationships among scientific literatures are more complex with higher quality than that among uses or items in commercial case. Thus, these valuable semantic relations abundant in scientific corpus must be taken into account.

To address the above issues, we propose a novel framework with link-missing data adaption for RSAR, which casts the recommendation task to subspace embedding and similarity ranking problems. In our proposed framework, both valuable contents and relations among scientific literatures are well modeled via Relation Regularized Matrix Factorization (RRMF) simultaneously with the capacity of dealing with link-missing data based on Latent Dirichlet Allocation (LDA). More importantly, our contribution can be seen as the foundation of two other scientific recommendation tasks to some extent. In fact, measuring the relatedness of relevant papers in complex publication network is one unavoidable step to achieve their goals for personalized paper recommendation and citation recommendation. In the former task, e.g., using collaborative filtering, finding satisfying related papers is a vital step to search similar users with common interest for better predicting final recommendation. In the latter one, the key question can also be reduced to how to find the related papers with the most top-*k* similar ones for a given citation context. So, the proposed method can also be further adapted to two other recommendation tasks.

For related work, the proposed method is different from the previous research omitted here due to the page constraint. Firstly, our method is a matrix factorization based method. Secondly, no prerequisites for citation context or existing ontology terms are needed. Finally, we take the link-missing problem encountered in practical application into account, which, to our best knowledge, has never been discussed in the field of RSAR.

2. THE PROPOSED METHODS

The proposed novel framework with link-missing data adaption for RSAR is presented in this section.

2.1 Related Scientific Articles Relation Modeling for RRMF

Many previous link-based methods usually only consider one specific type of link information in scientific articles. In fact, there are at least three types of valuable relations among scientific

[*] Corresponding author

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10...$15.00.

articles, which can jointly provide diverse information of the properties reflected by the corpus. Taking advantage of RRMF, all three relational information can be directly modeled or tailored in a unique framework together.

1) Citation-Network

A citation and its reference form the most obvious link structure that exists in research papers. The link is established by the author who builds direct semantic relation with references, which offers an information flow indicating knowledge evolution. So this type of link is crucial for finding the related papers, which can help people well understand the background or the evolution process. Moreover, it is also just the foundation for uncovering the other two latent structures, i.e., cocitation-network and coupling-network. The citation relation matrix is given by

$$\mathbf{R} = \begin{cases} R_{i,j} = 1, & \text{if paper } (j) \text{ cites paper } (i) \\ R_{i,j} = 0, & \text{otherwise} \end{cases} \tag{1}$$

2) Cocitation-Network

Cocitation relation is originally analyzed in the field of scientometrics, which refers to the relation existed in two research papers appearing in a common reference list. This type of relation is more interesting than that of the direct citation relation, due to its relation built by another person instead of the author of citing paper. It can help to find related papers from the third party view. In addition, this relation type in earlier work may be better uncovered than that in later papers since citation usually lags after its publication. Based on citation relation matrix in Eq. 1, we can obtain cocitation relation matrix,

$$\mathbf{R}^{'} = \mathbf{R}\mathbf{R}^{T} . \tag{2}$$

3) Coupling-Network

Bibliographic coupling or called coupling is another approach to evaluate relatedness between research papers in scientometrics. It assumes that if there is at least one common paper in the reference lists from two candidate papers, those two papers are defined as bibliographic coupling. Thus, this type of relationship gives the opportunity to exploit those papers without being cited. Moreover, since the newly publication is more inclined to be cited, it can also offer much more information about the research frontier regarding the target paper needing recommendation. Similar to Eq. 2, the coupling relation matrix is

$$\mathbf{R}^{''} = \mathbf{R}^{T}\mathbf{R} . \tag{3}$$

2.2 Relation Regularized Subspace Constructing via RRMF

However, the problem in the linkage established in Section 2.1 is that all links adopt equal weights implicitly, because an author only gives a pure reference list without thorough citation motivations. So, the content of each paper must be taken into consideration as the complement of link structure. In the proposed framework, we employ Relation Regularized Matrix Factorization to model both contents and semantic relations in scientific corpus.

RRMF proposed by Li et. al. [1] in 2009 provides a unique framework for learning a subspace containing both relation and content information simultaneously. It is the extension of latent semantic indexing (LSI) and is originally used for classification problems in [1]. The objective of the optimization in RRMF is

$$\min_{\mathbf{U},\mathbf{V}} \frac{1}{2}\left\|\mathbf{X} - \mathbf{U}\mathbf{V}^{T}\right\|^{2} + \frac{\alpha}{2}(\left\|\mathbf{U}\right\|^{2} + \left\|\mathbf{V}\right\|^{2}) + \frac{\beta}{2}tr(\mathbf{U}^{T}\mathbf{L}\mathbf{U}) \ , \tag{4}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_{j} A_{ij}$ and

$$tr(\mathbf{U}^{T}\mathbf{L}\mathbf{U}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} \left\|U_{i,*} - U_{j,*}\right\|^{2} . \tag{5}$$

To achieve the goal of incorporating relation information into LSI, the adding regularization $tr(\mathbf{U}^{T}\mathbf{L}\mathbf{U})$ behind two former parts in Eq. 5 makes the latent representations of two scientific articles as close as possible if a relationship between them is existed. $\mathbf{X}$ represents the content matrix with row vectors expressing data elements. $\mathbf{A}$ is a relation matrix, in which non-zero elements A_{ij} denote the existing relationship between data i and data j , zero elements show no any relations. Thus, the relation modeled in Section 2.1 can be incorporated into $\mathbf{A}$ directly.

For learning $\mathbf{U}$ and $\mathbf{V}$ with alternating projection method, one parameter is fixed and the other one is updated [1] in each round alternatively. Then, the representation u_i of each article in RRMF subspace is acquired, which contains both content and relation information as the row vector of $\mathbf{U}$. Finally, we use Cosine similarity measure for our recommendation.

2.3 Link-Missing Data Adapting

In many practical occasions, however, the link-missing data problem in scientific corpus is inevitable. For example, to save computational costs, we usually use the snapshot of the whole citation network as the corpus for learning and recommending. In fact, there are many related papers out of this networked corpus. So, how to incorporate the outer articles without any links into the linked corpus is a significant issue. To address this problem, this paper proposes two alternative methods, i.e., the multi-subspace ranking and link-missing data regularized embedding in RRMF subspace via Latent Dirichlet Allocation (LDA) as a probabilistic generative model proposed by Blei et. al. [2] in 2003.

More specifically, the whole corpus $D = \{d_1, d_2, \ldots d_M\}$ with M documents includes K topics and V words. The process of writing a document d_i is modeled by LDA on the hypothesis that a person writing a document has certain topics in mind in advance. To generate each word w_n of N words in document d_M , firstly a topic z_n is selected following the $Multinomial(\theta_m)$ and then a word w_n is picked from the selected topic distribution under $p(w_n \mid z_n, \beta_{z_n})$, which is a multinomial probability conditioned on the topic z_n . $\boldsymbol{\beta}$ is a $K \times V$ matrix with the row vectors β_k as the mixture component of topic k and $\boldsymbol{\theta}$ is a $M \times K$ matrix with the row vectors θ_m as the topic mixture proportion for document d_M . α and η are two hyper-parameters for θ and β respectively. The probabilistic representation of whole corpus modeled by LDA can be obtained via Eq. 6,

$$\begin{aligned} & p(D \mid \alpha, \eta) \\ & = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \eta)) d\theta_d \ , \end{aligned} \tag{6}$$

where z_{dn} represents the topic of word w_{dn} in document d . For estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in LDA, Gibbs sampling algorithm is employed in this paper. From the view of matrix decomposition,

$\boldsymbol{\theta}$ denotes the compact representations of original documents in K dimensional subspace. Based on LDA, we propose two alternative methods to deal with link-missing data on the assumption that the image u_i of link-missing data has existed in RRMF subspace via Section 2.2 jointly modeling with linked data.

1) multi-subspace ranking (LDA-RRMF-S): For the link-missing data, we assume that if two documents belong to the same topic learned by LDA, they are more likely similar and related. According to this assumption, we propose the multi-subspace similarity measure, i.e., the similarity measure for the link-missing data d_i is defined as the combined version of two subspace cosine similarity, $0<\gamma<1$,

$$sim(d_i,d_j)=(1-\gamma)\cos(u_i,u_j)+\gamma\cos(\theta_i,\theta_j)\ . \qquad (7)$$

So, the added topic similarity can be seen as the implicit linkage relation between the link-missing data and the linked-data.

2) link-missing data regularized embedding (Mix-R-Mapping): Following the above idea, this paper further proposes a regularized embedding approach for the link-missing data. In the proposed approach, we extend the assumption in 1) to subspace level inspired by the idea of MDS [3], i.e., if two data are similar in topic subspace they are also more likely similar in RRMF subspace. Thus, we manage to preserve this relationship between two subspaces for embedding the link-missing data. We define that this relationship is preserved through its local neighborhood in two subspaces. Then, the embedded link-missing data in RRMF can be obtain via Eq. 8, $0<\lambda\leq 1$,

$$u_i^{'}=(1-\lambda)u_i+\lambda\sum_{n=1}^{N}\Phi(\theta_n^i)=(1-\lambda)u_i+\lambda(\sum_{n=1}^{N}w_n u_n^i)\ , \qquad (8)$$

where θ_n^i is the nth nearest neighbor of θ_i in topic subspace, $\Phi(\theta_n^i)=w_n u_n^i$ in which u_n^i is the corresponding representation of d_n^i with θ_n^i in RRMF subspace and w_n in Eq. 9,

$$w_n=\cos(r_i,r_n^i)\Big/\sum_{n=1}^{N}\cos(r_i,r_n^i)\ , \qquad (9)$$

is the weight as the normalized similarity between r_i and r_n^i in original Bag of Words space for u_i and u_n^i in order to take the different contribution of each u_n^i into account. Then, we use the regularized embedding representation $u_i^{'}$ for computing similarity in RRMF subspace via Cosine metric with the candidate papers.

3. EXPERIMENTS AND DISCUSSIONS

The experiments are conducted on the CiteSeer dataset[1], which consists of 3312 scientific publications classified into one of six classes and 4732 citation links. Following previous work, we employ All-But-One technique as our offline evaluation method with the F1 and NDCG (optimal DCG takes the total relevant papers into account) measures. In fact, the evaluation task for the RSAR is nontrivial, because how to select the related paper for testing is a key question which may cause bias for other baseline methods. In this paper, we define the related paper not only the papers with citation relation to the target but also with cocitation and coupling relations. Therefore, the task of finding related paper in this paper is more challenging since that if we test two cocitation-related papers we must disjoint the total papers which have the citation relations with the two papers and if we test two couping-related papers we must disjoint the whole common papers in the reference lists of two papers. So, the link is much sparser for the testing paper than that in the citation-related test scheme. In fact, we extend the task of finding only citation-related paper to all three relations in our evaluation.

3.1 Experiments for Non-Missing Link Case

In this section, we compare the proposed method (RRMF-Subspace) with other six baseline approaches, i.e., Cocitation [4]; CCIDF [5]; HITS Vector-based [6]; Katz [7]; LSI (Latent Semantic Indexing) [8]; Content-BOW (Content in original Bag of Words representation). In particular, we extract the papers whose number of relation count, i.e., the sum of row vector in **A** (including all three relations) is equal to or more than 20 times as our test data for recommendation. For evaluation, the each testing paper is randomly disjointed the relation link including all three types according the predefined proportion, 10%. In this section, we only incorporate citation network without testing related links into RRMF-Subspace approach for equal comparison with other link-based methods, e.g., Cocitation, CCIDF. The dimension of subspace is fixed to 250 for Section 3.1 and 3.2.

Table 1. Seven Methods Comparison on Citeseer for F1.

Method	F1@5	F1@10	F1@15	F1@20
Cocitation	0.0137	0.0148	0.0154	0.0145
CCIDF	0.0081	0.0106	0.0119	0.0122
HITS Vector-Based	0.0102	0.0197	0.0238	0.0287
Katz	0.0949	0.1113	0.1078	0.1040
RRMF-Subspace	0.1206	0.1516	0.1552	0.1511
LSI	0.0890	0.1009	0.0990	0.0951
Content-BOW	0.1123	0.1200	0.1125	0.1080

Table 2. Seven Methods Comparison on Citeseer for NDCG.

Method	NDCG @5	NDCG @10	NDCG @15	NDCG @20
Cocitation	0.0203	0.0229	0.0244	0.0244
CCIDF	0.0116	0.0154	0.0174	0.0181
HITS Vector-based	0.0237	0.0277	0.0300	0.0338
Katz	0.1225	0.1377	0.1475	0.1561
RRMF- Subspace	0.1457	0.1736	0.1938	0.2076
LSI	0.1197	0.1280	0.1338	0.1392
Content-BOW	0.1507	0.1541	0.1571	0.1633

From Table 1 and Table 2, it can be found that RRMF-Subspace method performs best in all seven approaches. Due to our testing scheme discussed above, the citation network is sparse in testing context. In addition, our objective of recommendation adds up to all three relations at the same time. So, Cocitation and CCIDF methods which are based on citation network show poor performance. For HITS Vector-based approach as a more general graph-based approach, computing authority values also largely depends on the directed graph with high quality. Thus, for our sparse citation network and complex relation finding task, HITS

[1] CiteSeer Dataset, http://www.cs.umd.edu/~sen/lbc-proj/data/

Vector-based approach is cornered by the acquired poor quality of citation context. In contrast, we can find the latter four methods are significantly better than the former three ones for their relatively looser restrictions. Specifically, Katz is a link-based method; LSI and Content-BOW are content-based methods; RRMF-Subspace is a link-content hybrid-based one. Therefore, it can be demonstrated that the proposed approach is more effective and robust than other methods listed in Table 2 for RSAR task.

3.2 Experiments for Link-Missing Case

In this section, several experiments to evaluate the further proposed methods in Section 2.3 are conducted for link-missing data recommendation case. More particularly, we randomly select 5% of total papers and then fully disjoin their all relations as the testing papers to seek all three missing relations built in Section 2.1. For this case that other link-based methods never can fit, RRMF-Subspace ($\gamma = \lambda = 0$), LDA-Subspace ($\gamma = 1$), LSI and Content-BOW are the baseline methods compared with the proposed LDA-RRMF-S, Mix-R-Mapping and R-Mapping ($\lambda = 1$) in Section 2.3 with best values for F1 and NDCG respectively. The K in LDA is 50 and N in Eq. 8 is 5.

Table 3. Link-Missing Data Testing on Citeseer for F1.

Method	F1@5	F1@10	F1@15	F1@20
RRMF-Subspace	0.1769	0.1813	0.1668	0.1560
LDA-RRMF-S	0.1770	0.1922	0.1810	0.1665
LDA-Subspace	0.1164	0.1065	0.1041	0.1007
Mix-R-Mapping	0.1931	0.1974	0.1882	0.1738
R-Mapping	0.1902	0.1883	0.1776	0.1661
LSI	0.1485	0.1437	0.1326	0.1216
Content-BOW	0.2021	0.1738	0.1534	0.1417

Table 4. Link-Missing Data Testing on Citeseer for NDCG.

Method	NDCG @5	NDCG @10	NDCG @15	NDCG @20
RRMF-Subspace	0.2023	0.2327	0.2433	0.2557
LDA-RRMF-S	0.2219	0.2555	0.2703	0.2799
LDA-Subspace	0.1682	0.1723	0.1800	0.1873
Mix-R-Mapping	0.3048	0.3188	0.3279	0.3359
R-Mapping	0.3189	0.3238	0.3286	0.3350
LSI	0.2113	0.2221	0.2280	0.2331
Content-BOW	0.2785	0.2773	0.2773	0.2837

From Table 3 and Table 4, we can find that LDA-RRMF-S, Mix-R-Mapping and R-Mapping perform better than original RRMF-Subspace for link-missing case under both F1 and NDCG measures. For average value comparison of these three methods, Mix-R-Mapping wins the best performance (0.1881) for F1 and R-Mapping acquires the best performance (0.3266) for NDCG. In contrast, other baseline approaches show relatively poor performance under both F1 and NDCG. Moreover, the detailed parameter sensitivity analysis for the proposed methods in Section 2.3 is presented in Fig. 1. to Fig. 4.. The smaller parameter γ and the larger parameter λ are more likely achieve relatively high performance in LDA-RRMF-S and Mix-R-Mapping respectively.

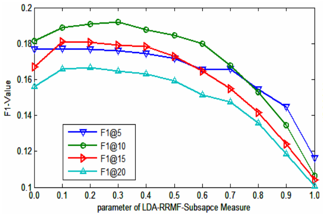

Fig. 1. LDA-RRMF-S@F1

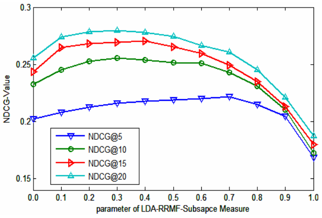

Fig. 2. LDA-RRMF-S@NDCG

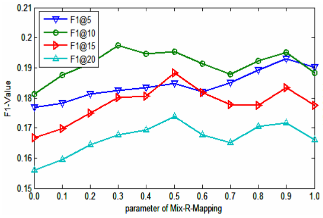

Fig. 3. Mix-R-Mapping@F1

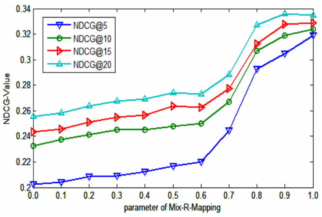

Fig. 4. Mix-R-Mapping@NDCG

4. CONCLUSIONS

This paper proposes a novel framework via RRMF with link-missing data adaption for RSAR, which is more general than citation-context oriented model and is also suited for other related item recommending problems. Moreover, in our framework, we can incorporate any link-missing data into a linked corpus as long as we jointly construct RRMF subspace and then employ the proposed LDA-based adapting methods for the isolated data. Particularly, for link-missing data regularized embedding, multi-view information from RRMF Subspace, LDA Subspace and Bag of Word Space has been jointly well considered. Furthermore, how to further unify the subspace embedding adaption and matrix decomposition simultaneously is our future work.

5. ACKNOWLEDGMENTS

This work is supported by Beijing Natural Science Foundation (4122068), Funds of Industrial Fundamental Research and National Key Technologies R&D Program of China in 12th Five-Year Plan (2011BAH10B05).

6. REFERENCES

1. W. Li and D. Yeung, Relation regularized matrix factorization. In *Proceedings of IJCAI*. 2009, 1126-1131.
2. D.M. Blei, A.Y. Ng, and M.I. Jordan, Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003, 993-1022.
3. J.T. Kwok and I.W. Tsang, The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*. 2004, 1517-1525.
4. H.Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS*. 1973, 24(4):265–269.
5. S. Lawrence, C.L. Giles, and K.D. Bollacker, Digital libraries and autonomous citation indexing. *IEEE Computer*. 1999, 67-71.
6. W. Lu, J.C.M. Janssen, E.E. Milios, N. Japkowicz, and Y. Zhang, Node similarity in the citation graph. *Knowledge and Information Systems*. 2007, 105-129.
7. D. Liben-Nowell and J.M. Kleinberg, The link prediction problem for social networks. *JASIST*. 2007, 1019-1031.
8. S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, Indexing by latent semantic analysis. *JASIS*. 1990, 391-407.