Recommendation of Newly Published Research Papers using Belief Propagation

Jiwoon Ha, Soon-Hyoung Kwon,
Sang-Wook Kim*
Dept. of Computer and Software
Hanyang University
Seoul, Korea
{jiwoonha, rikar, wook}@hanyang.ac.kr

Dongwon Lee
College of Information Sciences and Technology
Penn State University
PA, USA
dongwon@psu.edu

ABSTRACT

The problem to retrieve most relevant research papers for a given academic is studied. Existing solutions cannot adequately address the recommendation of *new* papers due to their lack of history information, the so-called *cold start* problem. Using the graphical model built from citation information between a new paper p_i and published papers, toward this challenge, we propose a novel approach based on a probabilistic inference algorithm, the *Belief Propagation (BP)*, to predict the likelihood of p_i's relevance to a target academic. Compared to item-based collaborative filtering method using a DBLP data set, the empirical validation shows an improvement in accuracy up to 26% in *F1 score*.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Algorithms

Keywords

Paper recommendation, data mining, belief propagation

1. INTRODUCTION

With the rapid development of technologies and expansion of disciplines, it becomes increasingly difficult for academics to locate "interesting" research papers, matching their own research interests or tasks in hand. While popular bibliographic digital libraries such as PubMed, DBLP, and arXiv provide a sophisticated search interface to help such academics, it requires the proactive initiation from users–i.e., users have to search for what they want. As such, a tool is greatly useful to actively locate and recommend the most interesting research papers to an academic, considering the general trends of communities and personalized interests.

In general, whether a research paper p_i is interesting to an academic a_j depends on many factors. If a_j is surveying on a particular research problem, for instance, a representative and authoritative paper p_i would be most interesting. However, if a_j wants to investigate a neighboring problem, an overlapping but diversified paper p_i might be more interesting. Since the level of "interesting-ness" of p_i to a_j is inherently subjective, therefore, we propose to use the *citation* as an indirect way to express one's interest toward papers.

Definition 1 (Interesting Paper) *When a paper p_i is cited in a research paper authored by an academic a_j, p_i is considered as an* interesting paper *to a_j.* □

Due to its applicability and usefulness in real settings, the problem of finding interesting papers for an individual has been extensively studied [4, 7, 9]. While effective in their own rights, however, existing solutions tend *not* to work for "newly published" papers. Existing methods often rely on ratings, usage history, or citation information of a target paper for recommendation. Therefore, when newly published papers lack such information, existing methods become inapplicable. This is the so-called *cold start* problem in recommender systems, the main focus of this paper. Formally, the problem is formulated as follows.

Problem 1 (Cold Start Paper Recommendation) *Given a set of published papers $\mathcal{P}$, a set of papers $\mathcal{A}$ that an academic a_j authored in past ($\mathcal{A} \subset \mathcal{P}$), and a set of newly published papers $\mathcal{N}$ ($\mathcal{P} \cap \mathcal{N} = \emptyset$), retrieve top-k newly published papers $\mathcal{K}$ ($\subseteq \mathcal{N}$) that are most interesting to a_j.* ■

We note that although a newly published paper p_n ($\in \mathcal{N}$) does not have accrued any incoming citations or usage data to p_n yet (and thus may not use existing recommendation solutions), in the reference section of p_n, there are many outgoing citations to other already published papers in $\mathcal{P}$. Exploiting these outgoing citations of p_n, therefore, we propose to model the entire citation relationship among all papers in $\mathcal{P}$, and those between $\mathcal{P}$ and p_n as a graph, and apply a probabilistic inference algorithm, the *Belief Propagation (BP)*, to predict the likelihood of p_n's interesting-ness toward a_j.

To probabilistically infer the state of a node in a graph, the *BP* is widely used [1, 3, 6, 10, 8]. By employing the *BP*

*Corresponding author.

RACS'14 October 5-8, 2014, Baltimore, MD, USA.

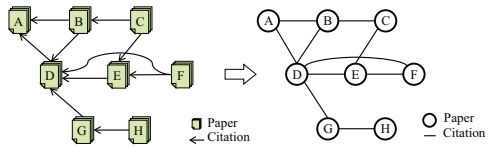


Figure 1. Generating an undirected graph from papers and their inter-paper citation information.

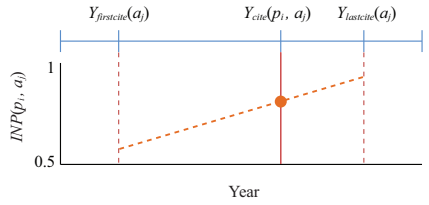


Figure 2. Assigning node potential.

to recommend newly published papers, we can infer whether a target academic is interested in each paper or not. Furthermore, the *BP* can reflect the interest of academics more accurately by considering heterogeneous relationships ("interesting" and "not-interesting").

2. THE PROPOSED METHOD

In this section, we present our proposed *BP* based method to accurately recommend top-k most interesting newly published papers for an individual. The *BP* is widely used to infer the state of a node based on the status of its neighboring nodes [1, 3, 6, 10, 8]. Our method probabilistically determines a target academic's interest on papers based on a undirected graph where nodes and edges indicate papers and citations, respectively. By employing the undirected graph, there is no distinction between "cite" and "be cited." Possible states of nodes are then defined as "interesting" and "not-interesting." Figure 1 illustrates this idea.

Once the graph is built, then, the proposed method runs as follows: For each target academic, (1) the node potential is assigned based on the target academic's interests (i.e., previous citation information); (2) the propagation matrix is defined; (3) messages are passed between nodes iteratively until convergence; (4) the "interest" belief scores are computed for all nodes; and (5) top-k newly published papers in the order of the "interest" belief score are recommended. These steps are further elaborated below.

2.1 Node Potential

In order to provide the personalized recommendation through the *BP*, we assign the node potentials based on the target academic's interest. The node potentials are represented as *vectors* that have possible states as elements. Each node potential is allocated as a value between *0.1* and *0.9*[1]. The sum of node potentials of a node is normalized to *1*. We refer to the "interesting" node potential of a paper p_i cited by a_j as $INP(p_i, a_j)$ and "not-interesting" node potential of a paper p_i cited by a_j as $NINP(p_i, a_j)$. $NINP(p_i, a_j)$ is defined as $1 - INP(p_i, a_j)$.

A naïve approach to reflect an academic's interest on node potentials is to assign the same value to node potentials of all the papers cited by her/him. Her/his interest, however, could be changed over time. To address this phenomenon, we assign a higher initial node potential to a more recently cited paper as in Equation 1.

$$INP(p_i, a_j) = \alpha + \beta \times \frac{Y_{cite}(p_i, a_j) - Y_{firstcite}(a_j)}{Y_{lastcite}(a_j) - Y_{firstcite}(a_j)} \quad (1)$$

α and β are constants to control the range of the $INP(p_i, a_j)$ value. In order to control this $INP(p_i, a_j)$ value within the range of *0.5* and *0.9*, we assign *0.5* to α and *0.4* to β. We control the minimum of the $INP(p_i, a_j)$ value of the paper cited by a_j as *0.5* to keep the $INP(p_i, a_j)$ value of the paper higher than the $NINP(p_i, , a_j)$ value ($NINP(p_i, a_j) = 1 - INP(p_i, a_j)$).

In addition, $Y_{cite}(p_i, a_j)$ indicates the year when a_j cited the paper p_i, which is to assign the node potential; $Y_{lastcite}(a_j)$ indicates the year of the most recent paper cited by a_j; $Y_{firstcite}(a_j)$ indicates one year before the oldest paper cited by a_j. Suppose $Y_{cite}(p_i, a_j)$ is 1997, $Y_{lastcite}(a_j)$ is 2001, and $Y_{firstcite}(a_j)$ is 1985. In this case, $INP(p_i, a_j)$ is computed as $0.5 + 0.4 \times \frac{1997-1984}{2001-1984}$, and thus becomes 0.805. Node potentials of papers that a_j has not cited before are initialized to be unbiased, i.e., $INP(p_i, a_j) = NINP(p_i, a_j) = 0.5$.

Figure 2 illustrates this idea visually. Dotted line indicates the proposed method while the orange circle indicates the actual $INP(p_i, a_j)$ values assigned. Note that, for the better understanding, the circle is represented like an absolute value (i.e., a published year of a paper). However, the assigned value is rather the ratio of the published year of a paper to the timeline of the corresponding paper set– a set of papers cited by a_j.

2.2 Belief Propagation

In general, the *BP* infers the state of a node by exchanging information between nodes. The information exchanged between nodes is defined as *messages*, the procedure to exchange the messages is *message passing*, and the information passed by message passing is aggregated as *belief scores*. The final state of a node is determined by belief scores.

Message Passing. A message is essentially a neighboring node's opinion (i.e., belief) about a target node's likelihood of being in a specific state (e.g., "interesting" or "non-interesting"). The messages sent from one node to another are represented as vectors, named as the message vectors. The elements of message vectors are the possible states of nodes. The message that node v_i sends to node v_j about the probability of node v_j being in x_c state is computed as follows:

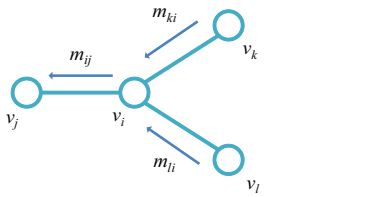
$$m_{ij}(x_c) \leftarrow \sum_{x_d \in X} \phi_i(x_d)\psi_{ij}(x_d, x_c) \prod_{k \in N(i) \setminus j} m_{ki}(x_d) \quad (2)$$

where $m_{ij}(x_c)$ denotes the message from v_i to v_j, indicating v_i's belief about v_j's likelihood of being in class x_c. The message from v_i to v_j is made up with the product of the messages from v_i's neighboring nodes except v_j. $\phi_i(x_d)$ is a node potential that represents the probability of v_i being in state x_d. $\psi_{ij}(x_d, x_c)$ represents the probability of v_j being in state x_c when its neighboring node v_i is in state x_d. It is defined by the propagation matrix in Section 2.3.

The message passing between nodes is conducted iteratively for a specific number of times or until the change of the message values between iterations is converged [8]. After the iterative message passing, the sum of every element of each message vector is normalized to 1. Figure 3 illustrates

[1]This is because *0* makes the multiplication *0*, and *1* makes another node potential *0*.

that the computation of node v_i's message about node v_j is in state x_c when there are two possible states, x_c and x_d.



$$m_{ij}(x_c) \leftarrow \begin{Bmatrix} m_{ki}(x_c) \\ \times \quad m_{li}(x_c) \\ \times \quad \psi_{ij}(x_c, x_c) \\ \times \quad \phi_i(x_c) \end{Bmatrix} + \begin{Bmatrix} m_{ki}(x_d) \\ \times \quad m_{li}(x_d) \\ \times \quad \psi_{ij}(x_d, x_c) \\ \times \quad \phi_i(x_d) \end{Bmatrix}$$

Figure 3. Computation of node v_i's message to node v_j.

Belief Score. After the message passing is converged, the belief score of each node is computed. The belief score of a node represents the probability that the node is in a particular state. Same as the message vectors, the belief scores are also represented as vectors with the possible states as the elements. The computation of the belief score for a node v_i's probability of being in state x_c is as follows:

$$b_i(x_c) = \delta \prod_{j \in N(i)} m_{ji}(x_c) \tag{3}$$

where δ is the normalization factor for making the sum of the belief scores of a node to be 1.

2.3 Propagation Matrix

Lastly, we define the propagation matrix related to $\psi_{ij}(x_d, x_c)$ of Eq. 2. Table 1 is an example of a propagation matrix with two possible states, "interesting" and "not-interesting." In the $\psi_{ij}(x_d, x_c)$ of Eq. 2, x_d and x_c indicates the row and column of the matrix, respectively. That is, the probability that the target academic is likely to have an interest in the neighboring nodes of a node v_i being in *interesting* state, ψ_{ij}(interesting, interesting) is $0.5 + \epsilon$.

We expect that a target academic is likely to have an interest in a paper if it is cited by another paper that the target academic has an interest in. In the proposed method, therefore, we assign $0.5 + \epsilon$ to ψ_{ij}(interesting, interesting). On the other hand, it is expected that the target academic would not have an interest in a paper cited by another paper that the target academic does not have an interest in. Therefore, we assign $0.5 - \epsilon$ to ψ_{ij}(not-interesting, interesting) that is smaller than the value of ψ_{ij}(interesting, interesting). In general, ϵ is set to be a very small number in order to avoid numerical underflow. In experiments, therefore, we used 0.0001 as ϵ.

Table 1. An example propagation matrix.

State	"interesting"	"not-interesting"
"interesting"	$0.5 + \epsilon$	$0.5 - \epsilon$
"not-interesting"	$0.5 - \epsilon$	$0.5 + \epsilon$

3. EXPERIMENTAL VALIDATION

3.1 Set-Up

For the empirical validation of our proposal, we used the DBLP data. We first identified *5,896* distinct academics from *6,241* papers published from 1971 to 2001 in databases and data mining fields. Note that our method uses only "citation" information between papers (i.e., no additional inputs about academics' research interests or papers' topics are used).

From the collected data set, we first generate an undirected graph G using papers and citation relationships among them from 1971 to 2001. Then, for each target academic a_j, who published at least 1 paper during 2002 to 2003 and cited at least 1 paper published in 2001 from those papers, we compute and assign a_j's personalized node potentials to G. Node potentials are assigned to G's nodes corresponding to papers cited by a_j in her previous published papers. The experimental question is then "*for each target academic a_j, how accurately can we predict top-k most interesting papers published in 2001?*"

The ground truth data set consists of the papers published in 2001 and cited by target academics during 2002 and 2003. There are 650 target academics in the ground truth data set. As the evaluation metrics, we use the well-known *precision*, *recall*, and *F1 score*. The k in top-k answers ranged from 1 to 5. The number of iterations in the *BP* was set to 20 by default.

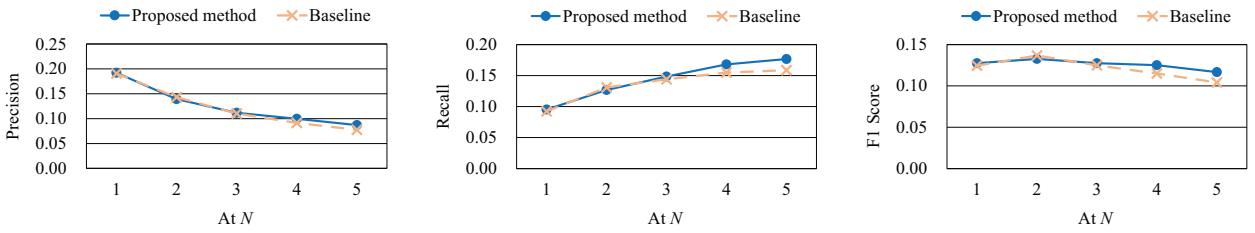
3.2 Effectiveness of Node Potential Assigning

We first analyzed the effectiveness of our node potential assigning method in Section 2.1. In order to show the effectiveness of the proposed method, we define the baseline method that assigns a fixed value (e.g., 0.9) to $INP(p_i, a_j)$ to every paper cited by the target academic. Figure 4 depicts three graphs, where x-axis indicates the number of recommended papers (i.e., k) and y-axis indicates the accuracy. The result reveals that the proposed method is slightly more accurate except when we recommend top-2 papers to the target academics. This indicates that the target academics' "interesting-ness" is reflected better when we assign node potentials based on the ratio of the published year of a paper to the timeline of the paper set that consists of papers cited by the target academic.

3.3 Comparison against Existing Method

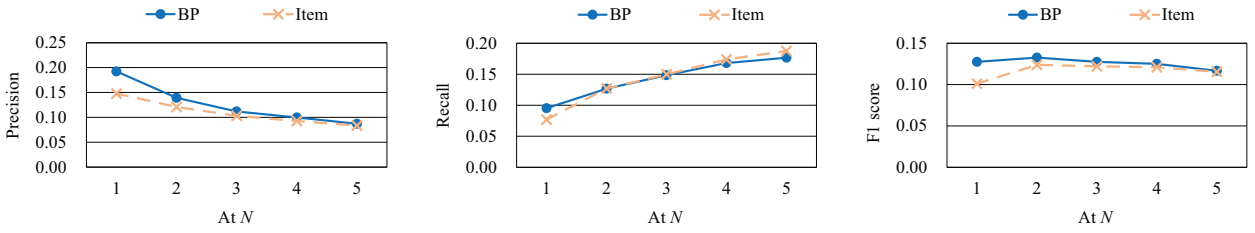
In order to evaluate our proposed method, we compared it against an existing method: item-based collaborative filtering method [2]. The item-based collaborative filtering method cannot recommend newly published papers since it relies on ratings (i.e., citation). Since newly published papers do not have incoming citations from other papers, the item-based collaborative filtering method becomes inapplicable. Therefore, we calculated the target academic's "interesting-ness" degree of each newly published paper as the average of its reference papers' "interesting-ness" degree. In this experiment, the node potentials are assigned as in Section 2.1 (Eq. 1). Figure 5 shows the *precision*, the *recall* and the *F1 score*. The x- and y-axes indicate the number of recommended papers and the accuracy of each method, respectively.

Figure 5 demonstrates that our proposed method outperforms the existing method in *precision*. In the *recall*, however, the proposed method only outperforms at top-1. Although the proposed method provides lower *recall* than the item-based collaborative filtering method from at top-2 to at top-5, in the *F1 score*, it outperforms the existing method in all cases. In the *F1 score*, the *BP* method improves the



(a) Precision (b) Recall (c) F1 score

Figure 4. Accuracy according to two node potential assignments.



(a) Precision (b) Recall (c) F1 score

Figure 5. Comparison against existing method.

accuracy by up to 26% against existing method. In the top-k recommendation, it is more important to recommend a proper research paper at the top of the recommendation list than the overall recommendation list. We claim that, therefore, the proposed method is more suitable for the practical research paper recommendation than the existing method.

From these results, we can conjecture that, by inferring a target academic's preference on newly published paper through the Belief Propagation on a graph modeled from the relationships between academics and research papers, we can infer effectively the implicit preference between academics and papers through transitivity [5]. In addition, the proposed method reflects the interest of academics more accurately by considering both *homophily* and *heterophily*. If there is a paper in which the target academic has an interest, the *BP* computes both the likelihood that the target academic has an interest in a paper and the likelihood that she does *not* have an interest in the paper.

That is, the *BP* considers the following four cases: for a target academic a_j and two papers p_1 and p_2, (1) the likelihood that a_j has an interest in p_1 cited by p_2 that a_j has an interest in, (2) the likelihood that a_j does not have an interest in p_1 cited by p_2 that a_j has an interest in, (3) the likelihood that a_j has an interest in p_1 not cited by p_2 that a_j has an interest in, (4) the likelihood that a_j does not have an interest in p_1 not cited by p_2 that a_j has an interest in. We believe that this flexible coverage of four heterogeneous cases in the *BP* is one of the main reasons of the improved accuracy.

4. CONCLUSIONS

In this paper, we proposed a *Belief Propagation* (BP) based method to recommend most interesting *newly published* papers for an academic, solving the cold start paper recommendation problem. By probabilistically inferring the target academic's interest using only citation relationship among papers without requiring any additional inputs, our proposed method was able to outperform item-based collaborative filtering method substantially in accuracy up to 26% with *F1 score*.

Acknowledgements

This research was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded by Korea Small and Medium Business Administration (Grants No. C0191469) and also by Ministry of Culture, Sports, and Tourism (MCST) via Korea Copyright Commission in 2014.

5. REFERENCES

[1] D. Chau et al. Polonium: Tera-scale graph mining for malware detection. *ACM KDD*, 2011.

[2] P. Cremonesi. Performance of recommender algorithms on top-n recommendation tasks. *ACM RecSys*, pages 39–46.

[3] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1):41–54, 2006.

[4] M. Gori and A. Pucci. Research paper recommender systems: A random-walk based approach. *IEEE/WIC/ACM WI*, pages 778–781, 2006.

[5] J. Ha et al. Top-n recommendation through belief propagation. *ACM CIKM*, pages 2343–2346, 2012.

[6] M. McGlohon et al. Snare: A link analytic system for graph labeling and risk detection. *ACM KDD*, pages 1265–1274, 2009.

[7] S. M. McNee et al. On the recommending of citations for research papers. *ACM CSCW*, pages 116–125, 2002.

[8] S. Pandit et al. Netprobe: A fast and scalable system for fraud detection in online auction networks. *WWW*, pages 201–210, 2007.

[9] K. Sugiyama and M. Kan. Scholarly paper recommendation via user's recent research interests. *ACM/IEEE JCDL*, pages 29–38, 2010.

[10] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring Artificial Intelligence in the New Millennium*, 8:236–239, 2003.