Assessing the authors of online books in digital libraries using users affinity

Baptiste De La Robertie

Assessing the Authors of Online Books in Digital Libraries Using Users Affinity

B. de La Robertie(✉)

Université de Toulouse, IRIT UMR5505, 31071 Toulouse, France
baptiste.delarobertie@irit.fr

Abstract. Information quality generated by crowd-sourcing platforms is a major concern. Incomplete or inaccurate user-generated data prevent truly comprehensive analysis and might lead to inaccurate reports and forecasts. In this paper, we address the problem of assessing the authors of users generated published books in digital libraries. We propose to model the platform using an heterogeneous graph representation and to exploit both the users' interests and the natural inter-users affinities to infer the authors of unlabelled books. We formalize the task as an optimization problem and integrate in the objective a prior of consistency associated to the networked users in order to capture the neighboors' interests. Experiments conducted over the *Babellio* platform (http://babelio.com/), a French crowd-sourcing website for book lovers, achieved successful results and confirm the interest of considering an affinity-based regularization term.

Keywords: User-generated-content · Labels propagation · Classification

1 Introduction

Over the past decade, crowd-sourcing platforms have entered mainstream usage and rapidly become valuable organizational resources, offering rich heterogeneous and relational data. However, to properly exploit the user-generated data and to produce comprehensive analysis, associated digital business must face several issues of quality and consistency. Even by clamping down signups, meta-data associated to users generated contents can be doubtfull or incomplete, justifying the needs of quality and consistency assessment tools.

In this work, the challenge of assessing the authors of unlabelled books in digital libraries is addressed. An heterogeneous graph is used to represent the platform and the relations between the different entities and a classification problem is formulated to predict the authors of unlabelled nodes. The *homophily patterns* lying between the interests of the users and their friends are first empirically demonstrated. Based on this observation suggesting that close friends tend to have similar favorite readings, an affinity-based regularization term in integrated in a dedicated objective function in order to smooth latent representations of the users.

The paper is organized as follow. Section 2 introduces previous research closely related to our problem. Section 3 motivates the general ideal of our work. Section 4 describes the proposal. Finally Sects. 5 and 6 provide experimental setup, evaluations and conclusions.

2 Related Work

Several research has empirically demonstrated [1,5,6,8] or exploited [2,4,10] many types of correlations between the structural properties of a graph and the associated users properties. Cook et al. [5] show that people's affinity networks are highly correlated with several behavioral and sociodemographic characteristics, exploring geography, family ties, education, social class and others. In [8], the social structure of the Facebook affinity network of several American institutions in studied. The authors has examined the *homophily patterns* using assortativity coefficients based on observed ties between nodes, considering both microscopic and macroscopic properties. They show different realizations of networks and, for example, observe that women are more likely to have friends within their common residence while this characteristic for male-only networks exhibit a larger variation. Backstrom et al. [1] have studied the ways in which communities grow over time, and more importantly, how different groups come together or attract new members. By taking the case of the *LiveJournal* platform, they have shown how the affinity graph structure of a member impacts his propensity to join new communities. Similar results have been suggested over the collaboration networks of scientists. For example, in [3], authors suggest that two researchers are more likely to collaborate if both have already collaborated with a third common scientist. As in [4,10], we suppose that two nodes connected in a network will tend to have similar latent representations. Thus, we propose to capture *homophily patterns* using an *affinity-based* regularization term.

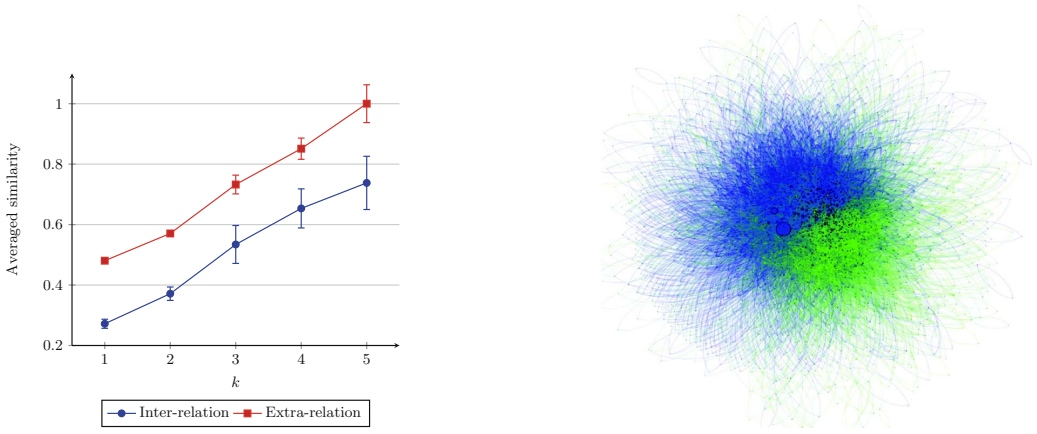
3 Motivations

In this section, we make use of the affinity graph of the members of the *Babelio* platform to demonstrate that linked users tend to have similar favorite books.

Let consider the affinity relation V such that $(i,j) \in V$ iff user i and user j are friends on the platform. Let $\boldsymbol{f}_i^k$ be a characteristics vector such that $f_{i,j}^k$ is the number of books written by author j for which user i has given k stars (from 1 to 5). From the averaged distance function S^k formalized in Eq. (1), we define the *inter-relation* and *extra-relation* distances metric as follow:

$$S^k = \frac{1}{N} \sum_i \sum_{j \in \mathcal{N}_i} \frac{||\boldsymbol{f}_i^k - \boldsymbol{f}_j^k||_2}{|\mathcal{N}_i|} \tag{1}$$

For a neighborhood $\mathcal{N}_i = \{j : (i,j) \in V\}$, i.e., the friends of the user associated to node i, the *inter-relation* metric captures the averaged distance between all nodes and their neighboors. For $\mathcal{N}_i = \{j : (i,j) \notin V\}$, i.e., users who are not



(a) (b)

Fig. 1. (a) Normalized averaged *Intra-relations* and *extra-relations* measures for ratings $k \in \{1, 2, 3, 4, 5\}$ over the experimental graphs. (b) An affinity graph extracted from the *Babelio* platform colored by users' favorite authors. (Color figure online)

friends with node i, we define the *extra-relation* metric. In practice, the latter is defined over a random subset of $\mathcal{N}_i$ such that the neighbooroods' size of both metrics are equal. Figure 1(a) reports the normalized evolution of both distances metrics in function of k over the four graphs used for the experimentations and described in Sect. 5.

Firstly, we observe that the *inter-relation* distance (in blue) is globally lower than the *extra-relation* one. In other words, connected nodes are more likely to read books of similar authors than non connected ones. This first observation constitutes the core of our proposal and justifies the regularization term proposed in Sect. 2 that constraints users to have similar latent representations. Secondly, from Fig. 1(b), which shows the main component of the affinity graph G_1 used in the experiments, we observe two distinct patterns. A color is associated to each author, and nodes are colored according to their favorite ones. Areas of uniform colors clearly reflect homophily patterns showing that users tend to naturally create communities sharing similar reading.

4 Model

Notations. Let $\mathcal{U} = \{u_i\}_{1 \leq i \leq n}$ be the set of users, $\mathcal{B} = \{b_j\}_{1 \leq j \leq m}$ the set of books and $\mathcal{A} = \{a_l\}_{1 \leq l \leq p}$ the set of authors, with $|\mathcal{U}| = n$, $|\mathcal{B}| = m$ and $|\mathcal{A}| = p$. Let $G_{pref} = (U_{pref}, V_{pref})$, with $U_{pref} = \mathcal{U} \cup \mathcal{B}$ and $V_{pref} = \{(u_i, b_j, v_{ij})\}_{i \leq n, j \leq m}$, be a bi-partite graph associating the interest $v_{ij} \in \mathbb{R}$ of user $u_i \in \mathcal{U}$ to book $b_j \in \mathcal{B}$. In addition, let $G_{friends} = (U_{friends}, V_{friends})$ with $U_{friends} = \mathcal{U}$ be an affinity graph: users u_i and u_j are friends iff $(u_i, u_j) \in V_{friends}$. Let $\boldsymbol{\alpha}_i \in \mathbb{R}^k$, $\boldsymbol{\beta}_j \in \mathbb{R}^k$ and $\boldsymbol{\gamma}_l \in \mathbb{R}^k$ be the latent representations of the users, books and authors respectively, with k being the dimension of the common latent space. Finally, let $\boldsymbol{y}_j \in \mathbb{R}^p$ be the labels vector associated to book

b_j. In particular, $y_{j,l} = 1$ if a_l is the author of book b_j, -1 otherwise. The goal is to reconstruct the labels vectors $\boldsymbol{y}_j$ for each unlabelled book.

Formulation. Predicting books' author is viewed as a classification task where the variable $y_{j,l} \in \{-1, +1\}$ has to be explained. In this work, we assume a set of linear classifiers per books, where the prediction $\tilde{y}_{j,l}$ for a pair $(b_j, a_l) \in \mathcal{B} \times \mathcal{A}$ is given by the linear model $f_l(b_j) = \langle \boldsymbol{\gamma}_l; \boldsymbol{\beta}_j \rangle$. Given a particular loss function $\Delta : \mathbb{R}^2 \to \mathbb{R}$, we propose to optimize the following objective:

$$\mathcal{L} = \sum_{(b_j, a_l) \in \mathcal{B} \times \mathcal{A}} \Delta(y_{j,l}, f_l(b_j)) + \sum_{(u_i, b_j) \in V_{pref}} d(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j) + \sum_{(u_i, u_j) \in V_{friends}} d(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j) \tag{2}$$

The first term computes the classification error related to the authors' predictions associated to each book. A Hinge loss function $\Delta(y_{j,l}, f_l(b_j)) = \max(0, 1 - y_{j,l} f_l(b_j))$, which is suitable for classification problems, was used in our experiments. The last two terms are aimed to smooth and propagate the decision variables through the different relations and capture the proposed intuition. The regularization d is done using the L_2 norm. Therefore, close friends and related favorite books tend to have similar representations in $\mathbb{R}^k$. We call the last term the *affinity regularization term*. Finding the representations of the users, books and authors such that $\mathcal{L}$ is minimized is equivalent to solve:

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*) = \arg\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}} \mathcal{L} \tag{3}$$

Since the Hinge loss is a convex function, standard approaches based on gradient descent can be used. In particular, we have:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}_i} = \sum_{(u_i, b_j) \in V_{pref}} 2(\boldsymbol{\alpha}_i - \boldsymbol{\beta}_j) + \sum_{(u_i, u_j) \in V_{friends}} 2(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j) \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}_j} = \sum_{(u_i, b_j) \in V_{pref}} 2(\boldsymbol{\beta}_j - \boldsymbol{\alpha}_i) - \sum_{1 \leq l \leq p} y_{j,l} \boldsymbol{\gamma}_l \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}_l} = \sum_{b_j \in \mathcal{B}} -y_{j,l} \boldsymbol{\beta}_j \tag{6}$$

In practice, we solved Eq. (3) using *L-BFGS* [7], a quasi-Newton method for non-linear optimizations. The parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ exhibit kn, km and kp decision variables respectively. Thus, our model parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ define a metric space in $\mathbb{R}^{k(n+m+p)}$.

5 Experiments

Dataset. For the experiments, several subsets of the *Babelio*[1] platform were used. Founded in 2007, *Babelio* is an emerging French crowd-sourcing portal

[1] http://babelio.com/.

Table 1. Four graphs used for the experiments.

	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{G_4}$
Authors	5	10	50	100
Books	525	937	5 615	11 462
Users	5 425	7 178	19 659	25 297
Preferences	25 470	28 852	156 022	251 477
Affinities	60 067	93 548	259 360	312 792

for book lovers, where internauts can share their favorite readings. Members can critic books by leaving textual comments and assigning from 1 to 5 stars. These engagement signals are made public by the platform, allowing members to network each others using a friendship functionality. Table 1 summarized the four graphs used for the experiments.

Evaluation Metric. For every book j, let σ^j be the permutation over the authors induced by the predicted scores $\tilde{\boldsymbol{y}}_j$. The ranking induced by σ^j is evaluated using the *Normalized Discounted Cumulative Gain* [9], computed as follow:

$$NDCG(\sigma^j, k) = \frac{DCG(\sigma^j, k)}{DCG(\sigma^{j,*}, k)} \text{ with } DCG(\sigma^j, k) = \sum_{i=1}^{k} \frac{2^{y_{j,\sigma^j(i)}} - 1}{\log(1+i)}$$

where $\sigma^{j,*}$ is the optimal ranking for book j, consisting in placing the real authors of a book in first positions. Thus, we capture how far the prediction is from the optimal rank. The average of the NDCG values over all the books is reported.

Protocol. For each graph, two optimizations, with identical initial values, are performed:

- **Prefs. + Aff.** The proposed objective as formalized in Equation (2).
- **Prefs. only.** The proposed objective without considering the *affinity regularization term.*

Since the initialization may affect the solution, only the best runs according to the introduced evaluation metric are reported. For each run, the dataset is randomly splitted into a train and a test datasets as follow: for each author, $x\%$ of his books are used for training, and the rest for testing. Results over the test dataset are reported.

Results. Several values of k have been tested and only the best runs are reported. Results are summarized in Table 2. Proposed solution globally improves the baseline in generalization by roughly 3 %, confirming our intuition and the interest of smoothing the users representations.

Table 2. Evaluation of the solutions using the NDCG metric over the test datasets.

	Training							
	10 %				50 %			
Graph	G_1	G_2	G_3	G_4	G_1	G_2	G_3	G_4
Prefs.	64.79	56.18	59.82	58.61	68.72	60.20	63.81	62.06
Prefs. + Aff.	65.52	59.04	61.12	60.76	69.59	64.91	67.57	64.97
Inprovement	+1.11 %	+4.84 %	+2.12 %	+3.53 %	+1.25 %	+7.25 %	+5.56 %	+4.47 %

6 Conclusions

We address the problem of assessing the authors of unlabelled books in digital libraries. To this end, the *homophily patterns* lying between the interests of the users and their friends are empirically demonstrated and incorporated as a regularization term in a dedicated objective function. By postulating that friends are more likely to share favorite readings, we force connected node to have similar representations. Experiments demonstrate significant quality improvement compared to the baseline that does not consider inter-users relationship. As future work, we will pursue our study by integrating the numerical votes in the system and new members caracteristics.

References

1. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, pp. 44–54. ACM, New York (2006)
2. He, Y., Wang, C., Jiang, C.: Discovering canonical correlations between topical and topological information in document networks. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM 2015, pp. 1281–1290. ACM, New York (2015)
3. Hou, H., Kretschmer, H., Liu, Z.: The structure of scientific collaboration networks in scientometrics. Scientometrics **75**(2), 189–202 (2008)
4. Jacob, Y., Denoyer, L., Gallinari, P.: Learning latent representations of nodes for classifying in heterogeneous social networks. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM 2014, pp. 373–382. ACM, New York (2014)
5. Cook, J.M., McPherson, M., Smith-Lovin, L.: Birds of a feather: homophily in social networks. Ann. Rev. Sociol. **27**, 415–444 (2001)
6. Newman, M.: Networks: An Introduction. Oxford University Press Inc., New York (2010)
7. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comput. **35**(151), 773–782 (1980)
8. Traud, A.L., Mucha, P.J., Porter, M.A.: Social structure of Facebook networks. CoRR, abs/1102.2166 (2011)

9. Yining, W., Liwei, W., Yuanzhi, L., Di, H., Wei, C., Tie-Yan, L.: A theoretical analysis of NDCG ranking measures. In: Proceedings of the 26th Annual Conference on Learning Theory (2013)
10. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schlkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems 16, pp. 321–328. MIT Press (2004)