Quality Prediction in Collaborative Platforms: A Generic Approach by Heterogeneous Graphs

Baptiste De La Robertie, Yoann Pitarch, Olivier Teste

Quality Prediction in Collaborative Platforms: A Generic Approach by Heterogeneous Graphs

Baptiste de La Robertie(✉), Yoann Pitarch, and Olivier Teste

Institut de Recherche en Informatique de Toulouse,
118 Route de Narbonne, 31071 Toulouse, France
{baptiste.delarobertie,yoann.pitarch,olivier.teste}@irit.fr

Abstract. As everyone can enrich or rather impoverish crowd-sourcing contents, it is a crucial need to continuously improve automatic quality contents assessment tools. Structural-based analysis methods developed for such quality prediction purposes generally handle a limited or manually fixed number of families of nodes and relations. This lack of genericity prevents existing algorithms for being adaptable to platforms evolutions. In this work, we propose a *generic* and *adaptable* algorithm, called *HSQ*, generalising various state-of-the-art models and allowing the consideration of graphs defined by an arbitrary number of nodes semantics. Evaluations performed over the two representative crowd-sourcing platforms *Wikipedia* and *Stack Exchange* state that the consideration of additional nodes semantics and relations improve the performances of state-of-the-art approaches.

Keywords: Link-analysis · Heterogeneous graphs · Quality

1 Introduction

Scientific literature has demonstrated strong correlations between users authority and contents quality on collaborative platforms [6,8,12,21]. Statistically, authoritative users are more likely to produce high quality content than others. Many state-of-the-art link analysis approaches exploit this *mutual reinforcement principle* between quality and authority for a quality assessment task [3,11,14,17,21]. However, most of them suffer from two major limitations. First, the lack of *genericity* of the formulations restricts them to a particular platform, making the solutions hardly transposable from one portal to another. Second, the lack of *adaptability* of the formulations prevents the algorithms from anticipating changes in the underlying graph. Thus, additional semantics of nodes or relations are most of the cases impossible to handle. These two limitations, shared by many structural-based algorithms, constitute the main motivations of our work. Our contributions are as follows:

– We propose a generic formulation of collaborative platforms using heterogeneous graphs and an unsupervised algorithm, *HSQ* (Heterogeneous Structural Quality), handling an unpredefined number of semantics of nodes and relations;

– We demonstrate the genericity of the proposal by instanciating three different and recent state-of-the-art algorithms and show how to easily integrate new semantics of nodes and relations;
– We conduct empirical studies on two real data sets from the *Wikipedia* and *Stack Exchange* portals that demonstrate a significant interest of considering additional entities and relations for the quality assessment task in crowdsourcing platforms.

2 Related Work

A first family of models for the quality assessment task on collaborative platforms exploit *contents signals*. Textual indices, numbers of citations or content length are some examples of content features used by *content-based* quality models. For example, on Wikipedia, it has been shown that the number of words per article [4] and the lifespan of the edits [1] are good quality predictors. However, content-based signals are too specific to a specific platform. Our work falls in the second family of approaches exploiting *structural signals* from the relations between the entities. Many works has empirically demonstrated correlations between users authority and contents quality, justifiying the wide range of PageRank [16] and HITS [13] based methods developped in the literature. On Wikipedia, a study of Dalip et al. [7] shows that structural features represent the most important family of predictors in a quality prediction task. More particularly, non considering such features leads to the greatest loss in terms of model quality. Hu et al. [10] propose to identify high quality articles on Wikipedia by exploiting this mutual dependency over a bipartite graph associating the articles to their contributors. Still on Wikipedia, a previous work [8] shows the interest of considering a co-edit relation between authors and reviewers to identify high quality articles. The study postulates that authoritative users get used to collaborate to produce high quality articles. Zhang et al. [20] apply the PageRank algorithm to on-line forums to identify authoritative users. Campbell et al. [5] and more recently Jurczyk et al. [12] make use of the HITS algorithm over a users-interaction graph to show a positive correlation between authority and quality. Recent analysis on Stack Overflow [15] and Quora [2,18] underlines the cyclic relation between content quality and producers authority.

If this mutual reinforcement principle has been extensively exploited for simple graphs considering a few types of nodes and relations, it seems that no formulation has been proposed for more complex graphs and in particular for heterogeneous graphs.

3 Approach Description

Notations. Let $\mathcal{G} = (\mathcal{H}, \mathcal{V})$ be an heterogeneous graph defined over m families of nodes $\mathcal{H} = \{\mathcal{U}_i\}_{1 \leq i \leq m}$, and a set of binary relations $\mathcal{V} \subseteq \mathcal{H} \times \mathcal{H}$. We denote by n_i the number of entities in the family $\mathcal{U}_i$. Let $(\mathcal{U}_i, \mathcal{U}_j) \in \mathcal{V}$ be a pair of families. We note $\mathcal{V}_{ij}$ the relation defined over $\mathcal{U}_i \times \mathcal{U}_j$ and A_{ij} the associated adjacency

matrix. We denote by $\mathbf{q}_i \in [0,1]^{n_i}$ the quality scores vector of the entities in the family $\mathcal{U}_i$.

Model. Firstly, for each pair of families $(\mathcal{U}_i, \mathcal{U}_j) \in \mathcal{V}$, we suppose a pair of *influence functions* (f_{ij}, g_{ji}) to model the *reinforcement principle*. Informally, the quality of the nodes in $\mathcal{U}_i$ influences the nodes quality in $\mathcal{U}_j$ and conversely, the nodes quality in $\mathcal{U}_j$ influences back the nodes quality in $\mathcal{U}_i$. This cyclic relation is illustrated in Fig. 1(a). More formally, we impose $\mathbf{x}_j = f_{ij}(\mathbf{y}_i)$ and $\mathbf{y}_i = g_{ij}(\mathbf{x}_j)$, with $\mathbf{x}_i \in [0,1]^{n_i}$ and $\mathbf{y}_i \in [0,1]^{n_i}$ being two vectors of *partial quality* scores. Note that if $(\mathcal{U}_i, \mathcal{U}_j) \notin \mathcal{V}$, we assume $f_{ij} = g_{ij} = 0$. Secondly, by considering linear aggregations of the different influences (see example in Fig. 1(b)), we have $\mathbf{x}_i = \sum_{j=1}^{m} f_{ji}(\mathbf{y}_j)$ and $\mathbf{y}_i = \sum_{j=1}^{m} g_{ij}(\mathbf{x}_j)$. In this work, we consider the case where influence functions are directly expressed by the adjacency matrices corresponding to each relation. Formally, $\forall \mathcal{V}_{ij} \in \mathcal{V}$, $f_{ij} = A_{ij}^T$ and $g_{ij} = A_{ij}$. Finally, by benoting $\mathbf{x}_i^{(t)}$ and $\mathbf{y}_i^{(t)}$ the partial quality scores at the t^{th} iteration of a label propagation process, the proposed quality model is expressed as follow:

Fig. 1. (a) Reinforcement principle between two families $\mathcal{U}_i$ and $\mathcal{U}_j$ such that $(\mathcal{U}_i, \mathcal{U}_j) \in \mathcal{V}$. (b) Linear aggregation of incoming influence functions for family $\mathcal{U}_l$.

$$\mathbf{x}_i^{(t)} = \sum_{j=1}^{m} A_{ji}^T \sum_{k=1}^{m} A_{jk} \mathbf{x}_k^{(t-1)} \text{ and } \mathbf{y}_i^{(t)} = \sum_{j=1}^{m} A_{ij} \sum_{k=1}^{m} A_{kj}^T \mathbf{y}_k^{(t-1)} \tag{1}$$

The quality q_i for each family $\mathcal{U}_i \in \mathcal{V}$ is computed as an aggregation function $\mathcal{A}_i$ of the partial quality scores $\mathbf{x}_i$ and $\mathbf{y}_i$, formally $q_i = \mathcal{A}_i(\mathbf{x}_i, \mathbf{y}_i)$. In this work, $\mathcal{A}_i$ is the average function $\forall i \in \{1, ..., m\}$.

Computation. The proposed algorithm, *HSQ* (Heterogeneous Structural Quality), is an iterative label propagation procedure propagating the adjusted scores through the relations $\mathcal{V}_{ij}$. Main steps are the following. (1) **Initialization**. For each $\mathcal{U}_i \in \mathcal{H}$, set $\mathbf{x}_i^{(0)}$ and $\mathbf{y}_i^{(0)}$ to random vectors. (2) **Propagation**. For each $\mathcal{U}_i \in \mathcal{H}$, update scores $\mathbf{x}_i^{(t)}$ and $\mathbf{y}_i^{(t)}$ with Eq. (1). (3) **Normalization**. Set $||\mathbf{x}_i^{(t)}|| = 1$ and $||\mathbf{y}_i^{(t)}|| = 1$. (4) **Return** $\mathcal{A}_i(\mathbf{x}_i, \mathbf{y}_i)$.

Steps (2) and (3) are repeated until a convergence step is reached. Convergence of the algorithm for the trivial case $m = 1$ is demonstrated in [9]. For the general case $m \geq 1$, we stop the propagation when $\sum_{i=1}^{m} ||\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}||_2 + ||\mathbf{y}_i^{(t)} - \mathbf{y}_i^{(t-1)}||_2 \leq \epsilon$. The algorithm returns a vector of scores $q_i \in \mathbb{R}^{n_i}$ for each family

of nodes $\mathcal{U}_i \in \mathcal{H}$. These scores should be ranked independently for each family in decreasing order of (predicted) quality.

Instances and Competitors. Wiki platforms are modelled with heterogeneous graphs using two families of nodes, the set of users and the set of articles (see Fig. 2(a)). Question and Answering websites are modelled with four families of nodes : users, answers, questions and comments (see Fig. 2(b)).

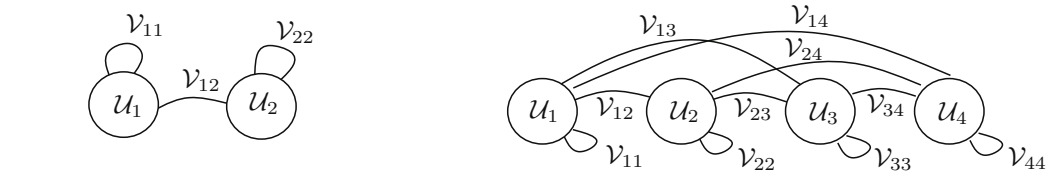

Fig. 2. (a) Wiki platform instance ($m = 2$). (b) Stack Exchange instance ($m = 4$).

On Wiki, the ***Basic*** model [10] constitutes a particular instance of the proposal, considering a bipartite graph ($m = 2$). Inter-user and inter-document relations are not considered, i.e., $\mathcal{V}_{11} = \mathcal{V}_{22} = \emptyset$. ***HSQ*** completes the previous model by considering collaborations $\mathcal{V}_{11}$ between users. Corresponding adjency matrix is such that $A_{11}(i, j)$ is the number of articles users i et j have co-edited. The degree of *collaboration* of the users is captured.

On Q&A websites, the ***HITS*** approach [11] and ***NCR*** model [21] are also particular instances of our model. In [11], a simple graph ($m = 1$) is considered, with $\mathcal{U}_1$ being the set of users. Authors assumes that $A_{11}(i, j) = 1$ if user j has answered at least once to a question formulated by user i. In [21], a graph with three families of nodes ($m = 3$) is considered, with $\mathcal{U}_1$, $\mathcal{U}_2$ and $\mathcal{U}_3$ being the set of users, answers and questions respectively. ***HSQ*** completes the *NCR* model by considering an additional set of entities $\mathcal{U}_4$ (the comments) and an inter-user relation $\mathcal{V}_{11}$. Adjency matrix associated to the inter-user relation is such that $A_{11}(i, j)$ is the number of answers i has provided *before* j to common questions. The *reactivity* of the users is captured.

4 Experiments

4.1 Datasets Description

Wikipedia.[1] A subset of roughly 23 000 articles was used. These articles were generated by 110 000 users and have been reviewed by the Editorial Team Assessment of the WikiProject. Each article is thus labelled according to the *WikiProject quality grading scheme* and belongs to one of the six class $FA \succ A \succ GA \succ B \succ C \succ S$. We assigned to each article i a numerical label y_i that respects

[1] https://en.wikipedia.org/wiki.

the user preferences. From $y_i = 0$ (class S, *Stub Articles*, i.e., very bad quality articles with no meaningful content) to $y_i = 5$ (class FA, *Featured Articles*, i.e., complete and professional articles). This scale is used as the ground truth in our evaluation. Recall we aim to rank articles by decreasing order of predicted quality. The repartition of the articles per class is summarized in Table 1.

Table 1. Statistics for the *Wikipedia* dataset.

Class	FA	A	GA	B	C	S
Label (y_i)	5	4	3	2	1	0
Number of articles	245	51	346	1 012	1 946	18 823

Stack Exchange.[2] The public dump of the *Stack Exchange* platform was used for evaluation. From October 2008 to September 2014, roughly 1 million of users, over 109 differents subplatforms, have generated more than 1.5 millions of questions, 2.5 millions of answers and 6.5 millions of comments. Numerical answers up votes, ranging from -65 to 2 182 for very popular answers are converted into integers. A first scale, noted b_s, is a binary scale where all negative answers, i.e., answers with score in $]-\infty, 0]$, constitute negative examples ($y_i = 0$) while all answers with positive scores constitute positive examples ($y_i = 1$). A second scale, used for ranking evaluation, noted r_s, is detailed in Table 2. Excepted for answers judged as bad quality (with negative scores), classes are balanced. Note that using b_s, we evaluate the capacity of the models to identify positive answers. Using r_s, the capacity of the models to rank answers in decreasing order of quality is evaluated.

Table 2. Answers scores discretization for the *Stack Exchange* dataset.

Class	A	B	C	D	E
Scores interval	$]-\infty, -1]$	$\{0\}$	$\{1\}$	$\{2, 3\}$	$]3, \infty[$
Number of answers	52 540	542 562	629 443	629 443	651 825
Label y_i	0	1	2	3	4

4.2 Evaluation Metrics

The ranking over the articles and the answers is evaluated with the *Normalized Discount Cumulative Gain at k* (NDCG@k) [19]. Let σ be the permutation ordering the documents by decreasing order of predicted quality. The DCG@k is defined as $DCG(\sigma, k) = \sum_{i=1}^{k} \frac{2^{y_{\sigma(i)}}-1}{\log(1+i)}$, where y_j is the label of document j. To compare different rankings, the normalized DCG is used,

[2] http://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/.

$NDCG(\sigma, k) = \frac{DCG(\sigma,k)}{DCG(\sigma^*,k)}$, where σ^* is the optimal ranking. On Wikipedia, σ^* places all *Features Articles* on top, then all articles belonging to class A, and so on. On *Stack Exchange*, the degree of relevance of an answer is given by scale b_s or r_s. The average *NDCG@k* is reported over all the questions. We also evaluate the precision of the solutions. On *Wikipedia*, we report the fraction of positive predictions per class. On *Stack Exchange*, the average fraction of positive answers beyond the first k answers over all the questions is reported.

4.3 Experiment Results

Results on the *Wikipedia* and *Stack Exchange* datasets are summarized in Tables 3 and 4 respectively. In both cases, user parameter ϵ is fixed to 10^{-4}.

Table 3. Evaluations of the two solutions on the *Wikipedia* dataset.

	Model	FA	A	GA	B	C	S
NDCG	***Basic***	73.77	75.14	80.76	**81.87**	**84.11**	93.11
	HSQ	**74.39**	**75.75**	**81.54**	81.19	83.16	**93.80**
Prec.	***Basic***	62.45	0	8.67	**39.03**	**34.53**	**94.17**
	HSQ	**64.9**	0	**17.92**	29.55	30.27	93.16

Table 4. Evaluations of the three solutions on the *Stack Exchange* dataset using the *NDCG* metric on scales b_s and r_s and the *Precision* metric on scale b_s.

	Model		k=**2**	k=**3**	k=**4**	k=**5**	k=**10**	k=**20**
NDCG	***HITS***	b_s	88.38	88.89	90.13	92.47	95.01	95.39
		r_s	67.27	71.26	75.64	80.29	85.21	85.98
	NCR	b_s	89.22	89.64	90.81	93	95.37	95.72
		r_s	69.33	73.07	77.26	81.26	86.21	86.91
	HSQ	b_s	**89.38**	**89.92**	**91.15**	**93.27**	**95.5**	**95.82**
		r_s	**69.49**	**74.47**	**77.85**	**82.08**	**86.41**	**87.05**
Prec.	***HITS***		81.41	80.38	79.14	77.85	49.92	26.63
	NCR		82.52	81.28	79.89	78.31	50.00	26.65
	HSQ		**82.83**	**81.85**	**80.55**	**78.77**	**50.10**	**26.66**

On *Wikipedia*, regarding classes *FA* and *GA*, experiments are very conclusive. Proposed solution clearly outperforms *Basic* [10], suggesting a non-negligible benefit (+2 % and +9 % for *FA* and *GA* articles resp.) of considering the strength of collaborations to identify high quality articles. Interestingly, the co-edit relation integrated in *HSQ* is not helpful for discriminating mid or poor quality articles (classes *B*, *C*, and *S*. On *Stack Exchange*, the interest of the proposition is immediate. For both metrics, proposed solution outperforms competitors. We conclude that both users reactivity and users engagement bring discriminating informations to identify authoritative users and, therefore, high quality answers.

5 Conclusions

In the scientific litterature, structural-based analysis approaches for quality prediction purpose rely on graphs considering a few number of families of nodes and relations. Moreover, most of them suffer from a common lack of genericity and adaptability. To tackle these limitations, an unsupervised structural based algorithm, *HSQ*, was proposed. Base on a heterogeneous graph representation of the data, the proposal enables the reformulation of various state-of-the-art methods. By instanciating *HSQ* over the two major collaborative platforms *Wikipedia* and *Stack Exchange*, we have shown the genericity of the proposed solution. Experiment results have suggested that considering additional entities and interactions in the model was beneficial. In future work, we plan to study different influence functions in order to give different strengths for each family of entities.

References

1. Adler, B.T., de Alfaro, L.: A content-driven reputation system for the wikipedia. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 261–270. ACM, New York (2007)
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of the International Conference on Web Search and Data Mining, WSDM 2008, pp. 183–194. ACM, New York (2008)
3. Bian, J., Liu, Y., Zhou, D., Agichtein, E., Zha, H.: Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 51–60. ACM, New York (2009)
4. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. In: Proceedings of the 17th International Conference on World Wide Web, WWW 2008, pp. 1095–1096. ACM, New York (2008)
5. Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise identification using email communications. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM 2003, pp. 528–531. ACM, New York (2003)
6. Chang, S., Pal, A.: Routing questions for collaborative answering in community question answering. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013, pp. 494–501. ACM, New York (2013)
7. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Automatic assessment of document quality in web collaborative digital libraries. J. Data Inf. Qual. **2**(3), 14:1–14:30 (2011)
8. de La Robertie, B., Pitarch, Y., Teste, O.: Measuring article quality in wikipedia using the collaboration network. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015. ACM, New York (2015)
9. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)

10. Hu, M., Lim, E.-P., Sun, A., Lauw, H.W., Vuong, B.-Q.: Measuring article quality in wikipedia: models and evaluation. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 243–252. ACM, New York (2007)
11. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 919–922. ACM, New York (2007)
12. Jurczyk, P., Agichtein, E.: Hits on question answer portals: exploration of link analysis for author ranking. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 845–846. ACM, New York (2007)
13. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
14. Li, B., Jin, T., Lyu, M.R., King, I., Mak, B.: Analyzing and predicting question quality in community question answering services. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012 Companion, pp. 775–782. ACM, New York (2012)
15. Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., Faloutsos, C.: Analysis of the reputation system, user contributions on a question answering website: stackoverflow. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013, pp. 886–893. ACM, New York (2013)
16. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
17. Suryanto, M.A., Lim, E.P., Sun, A., Chiang, R.H.L.: Quality-aware collaborative question answering: methods and evaluation. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM 2009, pp. 142–151. ACM, New York (2009)
18. Wang, G., Gill, K., Mohanlal, M., Zheng, H., Zhao, B.Y.: Wisdom in the social crowd: an analysis of quora. In: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013, pp. 1341–1352, Republic and Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee (2013)
19. 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)
20. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 221–230. ACM, New York (2007)
21. Zhang, J., Kong, X., Luo, R.J., Chang, Y., Ncr, P.: A scalable network-based approach to co-ranking in question-and-answer sites. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, pp. 709–718. ACM, New York (2014)