Prédiction de la qualité dans les plateformes collaboratives : une approche générique par les graphes hétérogènes

Baptiste de La Robertie*, Yoann Pitarch*, Olivier Teste*

*Université de Toulouse, IRIT UMR5505
prenom.nom@irit.fr

Résumé. La qualité des contenus sur les plateformes collaboratives est très hétérogène. Dans la littérature scientifique, les algorithmes d'analyse structurelle appliqués à la tâche de détection de contenu de qualité reposent généralement sur des graphes définis à partir d'un seul type de nœuds et de relations. Pourtant les graphes sur lesquels reposent ces récentes plateformes présentent de nombreuses sémantiques de nœuds et relations différentes, e.g., producteurs/consommateurs, questions/réponses, etc. Ces solutions souffrent d'un manque de généricité et ne peuvent s'adapter facilement à l'évolution des plateformes. Nous proposons une modélisation générique de ces platformes par les graphes hétérogènes pouvant intégrer automatiquement de nouvelles sémantiques de nœuds et de relations. Un algorithme de prédiction de qualité des contenus reposant sur ce modèle est proposé. Nous montrons qu'il généralise plusieurs travaux de la littérature. Enfin, en intégrant certaines relations inter-utilisateurs, nous montrons que notre solution, évaluée sur Wikipedia et Stack Exchange, améliore la tâche de détection de contenu de qualité.

1 Introduction

Le web voit apparaître de nombreuses plateformes collaboratives sur lesquelles l'internaute est sollicité pour enrichir différents types de contenus. Comme de nombreuses sociétés [1], les Wikis et sites de questions/réponses (Q&R) constituent les exemples les plus connus de portails collaboratifs tirant profits des collaborations à grande échelle. Le nouveau statut de producteur de l'utilisateur le place au cœur de la plateforme et savoir identifier influenceurs et productions de qualité est rapidement devenu d'un intérêt majeur. L'évolution de ces plateformes est la première motivation de notre travail. Par exemple, certains sites de Q&R comme *Yahoo ! Answers* [2] ou *Stack Exchange* [3] ont vu leur modèle évoluer en intégrant successivement de nouveaux types de contenus et donc de nouvelles relations entre les entités du modèle. Les solutions qui n'ont pas anticipé cette évolution ne peuvent tirer profit des nouvelles informations. Il est donc nécessaire de développer des formalisations génériques capables d'une part, d'anticiper la prise en compte de nouvelles sémantiques de nœuds et de relations, c'est à dire pouvant

1. https ://www.forrester.com/The+Collaborative+Economy+Will+Drive+Business+Innovation+And+Growth/fulltext/-/E-res113683
2. https ://fr.answers.yahoo.com/
3. http ://french.stackexchange.com/

accompagner l'évolution des plateformes, et d'autre part, capables de s'appliquer à différentes plateformes. C'est ce que l'on se propose de faire dans ce travail. Pour résumer, nos contributions donc sont les suivantes : (1) Nous proposons une formalisation générique par les graphes hétérogènes permettant de modéliser de nombreux portails collaboratifs différents ; (2) nous proposons un algorithme non-supervisé, *QGH* (Qualité dans les Graphes Hétérogènes), reposant sur cette formalisation capable de s'adapter à l'évolution des plateformes. Par ailleurs, nous montrons que notre modèle généralise de manière uniforme au moins deux travaux de l'état de l'art ; (3) des expérimentations conduites sur les deux plateformes Wikipedia et Stack Exchange confirment l'efficacité de l'algorithme. En particulier, nous montrons l'intérêt des relations inter-utilisateurs pour identifier les contenus de qualité.

2 État de l'art

Plusieurs études Chang et Pal (2013); Dalip et al. (2011, 2013); Movshovitz-Attias et al. (2013); Wang et al. (2013) réalisées sur *Wikipedia*, *Stack-Exchange* et *Quora* montrent que les caractéristiques structurelles sont d'une grande importance dans une tâche de prédiction de qualité des contenus et que la non-considération de cette sous-famille de signaux peut conduire à une perte significative en terme de qualité de modèle. De plus, d'autres analyses Hu et al. (2007); Jurczyk et Agichtein (2007) s'entendent sur l'intérêt d'exploiter une dépendance mutuelle entre qualité des contenus et autorité des utilisateurs. Des travaux récents de La Robertie et al. (2015) montrent même un intérêt réel à considérer les relations entre utilisateurs pour améliorer l'identification des contenus de qualité. Mais dans tous ces travaux, basés sur des modèles d'analyse structurelle non-supervisés qui exploitent cette dépendance mutuelle, seulement un ou deux types de contenus sont considérés, i.e., utilisateur/article ou question/réponse. Il semblerait qu'aucun travail n'ait proposé de formulation générique exploitant cette dépendance mutuelle dans un graphe défini sur un nombre quelconque de types de nœuds et relations. C'est l'objectif de ce travail de proposer une modélisation par les graphes hétérogènes exploitant plusieurs dépendances mutuelles permettant ainsi à l'algorithme de s'adapter à l'apparition de nouvelles familles d'entités et de relations.

3 Modélisation

Notations. Soit $\mathcal{G} = (\mathcal{H}, \mathcal{V})$ un graphe hétérogène défini sur m familles de nœuds $\mathcal{H} = \{\mathcal{U}_i\}_{1 \leq i \leq m}$, et un ensemble de relations binaires $\mathcal{V} \subseteq \mathcal{P}(\mathcal{H} \times \mathcal{H})$ entre les familles de $\mathcal{G}$. Soit $(\mathcal{U}_i, \mathcal{U}_j) \in \mathcal{V}$ une paire de familles de $\mathcal{G}$, on note $\mathcal{V}_{ij}$ la relation définie sur $\mathcal{P}(\mathcal{U}_i \times \mathcal{U}_j)$ et A_{ij} la matrice d'adjacence associée.

Modèle. On postule d'abord l'existence d'une dépendance mutuelle entre les qualités des éléments de chaque paire $(\mathcal{U}_i, \mathcal{U}_j) \in \mathcal{V}$. Autrement dit, on suppose que la qualité des nœuds de la famille $\mathcal{U}_i$ influe sur la qualité des nœuds de la famille $\mathcal{U}_j$ (première partie de l'influence) et inversement que la qualité des nœuds de la famille $\mathcal{U}_j$ influe en retour sur la qualité des nœuds de la famille $\mathcal{U}_i$ (deuxième partie de l'influence). Par ailleurs, on postule que ces influences cycliques sont distributives. Autrement dit, que la qualité des nœuds de $\mathcal{U}_i$ est directement influencée, d'une part, par la qualité des nœuds de toutes familles $\mathcal{U}_k$, tels que $(\mathcal{U}_k, \mathcal{U}_i) \in \mathcal{V}$

(agrégation des premières parties de l'influence sur $\mathcal{U}_i$) et d'autre part, directement influencée par la qualité des nœuds de toutes familles $\mathcal{U}_k$, tels que $(\mathcal{U}_i,\mathcal{U}_k) \in \mathcal{V}$ (agrégation des secondes parties de l'influence sur $\mathcal{U}_i$). En notant $\mathbf{x}_i$ et $\mathbf{y}_i$ les deux vecteurs de scores résultant de l'agrégation des premières et secondes parties des influences sur $\mathcal{U}_i$, les relations cycliques d'influence peuvent être exprimées par $\mathbf{x}_i = \sum_j f_{ij}(\mathbf{y}_j)$ et $\mathbf{y}_i = \sum_j g_{ij}(\mathbf{x}_j)$, où f_{ij} et g_{ij} expriment les influences que les nœuds $\mathcal{U}_i$ exercent sur les nœuds de $\mathcal{U}_j$. Dans ce travail, nous supposons que les influences f_{ij} et g_{ij} sont linéaires et exprimées par la matrice d'adjacence A_{ij} associée à $\mathcal{V}_{ij}$. En particulier, on pose $f_{ij}(\mathbf{y}_i) = A_{ji}^T\mathbf{y}_i$ (première partie de l'influence) et $g_{ij}(\mathbf{x}_i) = A_{ij}\mathbf{x}_i^T$ (seconde partie de l'influence). Ainsi, en remplaçant $\mathbf{x}_i$ et $\mathbf{y}_i$ de façon à couper la relation circulaire, le modèle d'influence proposé s'exprime par l'équation (1) :

$$\mathbf{x}_i^{(t)} = \sum_j A_{ji}^T \sum_k A_{jk}\mathbf{x}_k^{(t-1)} \text{ et } \mathbf{y}_i^{(t)} = \sum_j A_{ij} \sum_k A_{kj}^T\mathbf{y}_k^{(t-1)} \quad (1)$$

Finalement, le vecteur de scores de qualité des nœuds de la famille $\mathcal{U}_i$ est donné par $q_i(\mathbf{x}_i, \mathbf{y}_i)$, avec q_i une fonction d'agrégation (la fonction somme est utilisée dans cet article).

Algorithme (QGH). Les étapes de l'algorithme *QGH* sont les suivantes : **(a)** Initialisation aléatoire des scores $\mathbf{x}_i$ et $\mathbf{y}_i$; **(b)** Pour chaque $\mathcal{U}_i$, mise à jour des qualités avec l'équation (1) ; **(c)** Normaliser les vecteurs $\mathbf{x}_i$ et $\mathbf{y}_i$ indépendamment pour chaque famille de nœuds. Ces deux dernières étapes sont répétées jusqu'à stabilité des scores. Autrement dit, lorsque, d'une itération à l'autre de l'algorithme, la somme sur tous les ensembles $\mathcal{U}_i$ des variations des scores est significativement faible, i.e., quand $\sum_i ||\mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}||_2 + ||\mathbf{y}_i^{(t)} - \mathbf{y}_i^{(t-1)}||_2 \leq \epsilon$.

4 Expérimentations

4.1 Données

Wikipedia.[4] Plus de 22 000 articles produits par 111 000 auteurs issus de Wikipedia sont utilisés pour l'évaluation. L'équipe éditoriale a étiqueté ces articles selon plusieurs critères de qualité sur l'échelle $FA \succ A \succ GA \succ B \succ C \succ S$. La classe S correspond aux articles de mauvaise qualité et la classe FA aux articles de très bonne qualité. Cette échelle est utilisée comme vérité terrain pour l'évaluation. Les statistiques sur le jeu de données sont fournies dans le Tableau 1.

TAB. 1 – *Jeu de données Wikipedia.*

Classe	**FA**	**A**	**GA**	**B**	**C**	**S**
Articles	245	51	346	1 012	1 946	18 823
Label (y_i)	5	4	3	2	1	0

Stack Exchange.[5] Le jeu de données de Septembre 2014 est utilisé pour l'évaluation. Il contient toutes les questions et réponses publiées par les membres de la plateforme sur une

4. https ://en.wikipedia.org/wiki/Wikipedia :Version_1.0_Editorial_Team/Assessment
5. http ://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/

période de six ans (Octobre 2008 à Septembre 2014), soit environ 1,5 million de questions, 2,5 millions de réponses et 1 million d'utilisateurs. Les votes des utilisateurs sur les réponses sont utilisés comme vérité terrain pour évaluer les modèles. Ces scores variant de -65 à 2 182, nous les discrétisons de façon à obtenir des classes équilibrées (voire Tableau 2).

TAB. 2 – *Jeu de données Stack Exchange.*

Classe	**A**	**B**	**C**	**D**	**E**
Intervalle	$]-\infty,-1]$	$\{0\}$	$\{1\}$	$\{2,3\}$	$]3,\infty[$
Réponses	52 540	542 562	629 443	664 707	651 825
Label (y_i)	0	1	2	3	4

4.2 Instances

Le modèle proposé permet de formaliser de manière uniforme les modèles suivants :
Wikipedia. (A) *Basic* Hu et al. (2007) : $\mathcal{U}_1$ est l'ensemble des utilisateurs, $\mathcal{U}_2$ l'ensembe des articles. $\mathcal{V}_{12}$ associe à chaque utilisateur les articles qu'il a édités. Formellement, $A_{12}(i,j)$ est le nombre de mots que l'utilisateur i a écrit dans l'article j. **(B) *QGH*** : nous complétons la modélisation précédente en considérant une relation inter-utilisateurs. $\mathcal{V}_{11}$ est construite à partir de tous les couples d'utilisateurs qui ont participé à l'élaboration d'un même article. Formellement, $A_{11}(i,j)$ est le nombre d'articles que les utilisateurs i et j ont modifié ensemble.
Stack Exchange. (A) *NCR* Zhang et al. (2014) : $\mathcal{U}_1$ est l'ensemble des utilisateurs, $\mathcal{U}_2$ l'ensemble des réponses et $\mathcal{U}_3$ l'ensemble des questions. $\mathcal{V}_{12}$ associe à chaque utilisateur les réponses qu'il a formulées i.e., $A_{12}(i,j) = 1$ si i est l'auteur de la réponse j. $\mathcal{V}_{13}$ associe à chaque utilisateur les questions qu'il a posées i.e., $A_{13}(i,j) = 1$ si i est l'auteur de la question j. Enfin, $\mathcal{V}_{23}$ associe à chaque question les réponses qui y répondent i.e., $A_{23}(i,j) = 1$ si i répond à la question j. **(B) *QGH*** : nous complétons la modélisation précédente en considérant une relation inter-utilisateurs. Deux utilisateurs i et j sont en relation si i a répondu au moins une fois plus vite que j à une même question. Formellement, $A_{11}(i,j)$ est le nombre de questions auxquelles l'utilisateur i a répondu avant j. **(C) *HITS* Jurczyk et Agichtein (2007)** : seul $\mathcal{U}_1$, l'ensemble des utilisateurs, est considéré. $A_{11}(i,j) = 1$ si l'utilisateur j a répondu à une question posée par i, 0 sinon.

4.3 Évaluation

Métrique. Les modèles sont évalués sur leur capacité à ordonner les éléments des familles $\mathcal{U}_i$ par ordre de qualité décroissante. La métrique NDCG@k Yining et al. (2013), qui mesure une similarité avec l'ordonnancement optimal, est utilisée. Formellement, si σ est la permutation retournée par le modèle, $DCG(\sigma,k) = \sum_{i=1}^{k} \frac{2^{y_{\sigma(i)}}-1}{\log(1+i)}$ et $NDCG(\sigma,k) = \frac{DCG(\sigma,k)}{DCG(\sigma^*,k)}$ avec σ^* l'ordonnancement optimal. Sur Wikipedia, le classement optimal place tous les articles de la classe FA en première position, puis tous les articles de la classe A, etc. Sur Stack Exchange, c'est la capacité à placer, pour chaque question, les meilleures réponses en première position qui est évaluée. Dans ce cas, la valeur moyenne de la métrique sur toutes les questions est reportée.
Résultats. La valeur du paramètre ϵ controlant la convergence de l'algorithme est fixée à 10^{-4}

(au délà, les performances ne varient plus). Les résultats pour les jeux de données *Wikipedia* et *Stack Exchange* sont résumés dans les Tableaux 3 et 4. L'intérêt d'intégrer les relations inter-utilisateurs est immédiat. Sur Wikipedia, un gain de 6% est observé pour le NDCG@40. Notre solution place donc davantage d'articles ou de réponses de bonne qualité en tête de liste. On note que pour les articles de moyenne ou mauvaise qualité ($k \geq 1\,654$), les relations n'améliorent pas les performances : les auteurs de ces articles ne suscitant sans doute pas assez de collaborations.

TAB. 3 – *NDCG@k pour le jeux de données Wikipedia.*

Model	k=**40**	k=**50**	k=**245**	k=**296**	k=**642**	k=**1 654**	k=**3 600**	k=**22 423**
Basic Hu et al. (2007)	93.73	92.13	73.97	75.14	80.76	**81.85**	**83.92**	93.11
QGH	**100**	**98.93**	**75.66**	**76.24**	**81.69**	80.78	81.71	**93.15**

TAB. 4 – *NDCG@k moyen pour toutes les questions du jeu de données Stack Exchange.*

Model	k=**1**	k=**2**	k=**3**	k=**4**	k=**5**	k=**10**	k=**20**
QGH	**67.88**	**70.64**	**74.38**	**78.48**	**82.64**	**86.87**	**87.49**
NCR Zhang et al. (2014)	66.76	69.31	73.07	77.26	81.62	86.20	86.91
HITS Jurczyk et Agichtein (2007)	67.43	70.15	73.78	77.90	82.11	86.55	87.14

5 Conclusions

Les modèles de qualité existant développés pour les portails collaboratifs souffrent d'un manque de généricité. Dans ce travail, un nouvel algorithme générique *QGH* est proposé. Il suppose des dépendances mutuelles entre les qualités des familles de nœuds du graphe. Une modélisation par les graphes hétérogènes permet la prise en compte d'un nombre quelconque de familles de nœuds et de relations, rendant notre algorithme adaptable aux évolutions des plateformes. De plus, sa généricité permet de reformuler divers travaux de l'état de l'art et permet une utilisation pour de nombreuses plateformes différentes. A l'avenir, nous prévoyons d'utiliser de nouvelles sémantiques de nœuds pour discriminer les réponses de qualité sur les sites Q&R (commentaires sur les question/réponses). Nous prévoyons d'étudier les propriétés distributives du modèle pour proposer une version parrallèle de l'algorithme et d'instancier *QGH* avec d'autres fonctions d'influences propres à chaque famille de nœuds.

Références

Chang, S. et A. Pal (2013). Routing questions for collaborative answering in community question answering. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, New York, NY, USA, pp. 494–501. ACM.

Dalip, D. H., M. A. Gonçalves, M. Cristo, et P. Calado (2011). Automatic assessment of document quality in web collaborative digital libraries. *J. Data and Information Quality 2*(3), 14 :1–14 :30.

Dalip, D. H., M. A. Gonçalves, M. Cristo, et P. Calado (2013). Exploiting user feedback to learn to rank answers in q&a forums : A case study with stack overflow. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, New York, NY, USA, pp. 543–552. ACM.

de La Robertie, B., Y. Pitarch, et O. Teste (2015). Measuring article quality in wikipedia using the collaboration network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, New York, NY, USA, pp. 464–471. ACM.

Hu, M., E.-P. Lim, A. Sun, H. W. Lauw, et B.-Q. Vuong (2007). Measuring article quality in wikipedia : Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, New York, NY, USA, pp. 243–252. ACM.

Jurczyk, P. et E. Agichtein (2007). 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 '07, New York, NY, USA, pp. 919–922. ACM.

Movshovitz-Attias, D., Y. Movshovitz-Attias, P. Steenkiste, et C. Faloutsos (2013). Analysis of the reputation system and user contributions on a question answering website : Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, New York, NY, USA, pp. 886–893. ACM.

Wang, G., K. Gill, M. Mohanlal, H. Zheng, et B. Y. Zhao (2013). Wisdom in the social crowd : An analysis of quora. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, Republic and Canton of Geneva, Switzerland, pp. 1341–1352. International World Wide Web Conferences Steering Committee.

Yining, W., W. Liwei, L. Yuanzhi, H. Di, C. Wei, et L. Tie-Yan (2013). A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory*.

Zhang, J., X. Kong, R. J. Luo, Y. Chang, et P. S. Yu (2014). Ncr : 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 '14, New York, NY, USA, pp. 709–718. ACM.

Summary

Most of the link analysis methods developped for the Quality Assessment task consider only one set of nodes and a single relation. However, recent crowdsourcing platforms lay on various different semantics of nodes and relations. In this work, we propose a model of crowdsourcing platforms using heterogeneous graphs. Based on this representation, we propose an algorithm that can easily take advantage of various semantics of nodes and relations and make it adaptable to the evolution of the platforms. We show that our proposition generalizes some state-of-the-art models. Furthemore, experiments conducted on the two platforms, Wikipedia, and Stack Exchange, show a real interest to consider user interactions in this task.