Proposition de recherche doctorale

La classification croisée dans les systèmes de filtrage collaboratif

Mots clés :

- **Directeur de thèse** : Mohamed NADIF
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'Informatique PAris DEscartes
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

{{Contexte}} Les techniques de filtrage collaboratif ont pris une place très importante dans le filtrage de l'information. Elles sont destinées principalement aux systèmes de recommandation. En se basant sur les évaluations déjà exprimées par un utilisateur à propos d'autres items, elles consistent à prédire l'intérêt d'un utilisateur pour un nouvel item, à suggérer de nouveaux items ou encore à prédire l'utilité d'items inconnus pour un utilisateur donné. Les méthodes de filtrage collaboratif peuvent être essentiellement divisées en deux groupes : approches fondées sur la mémoire et approches à base de modèles [1]. Les approches fondées sur la mémoire ont la particularité d'utiliser toutes les données disponibles pour prédire la note d'un nouvel item par un utilisateur particulier. De telles approches, qui ont été utilisées sur de nombreuses applications réelles [2, 3], ont l'avantage d'intégrer immédiatement des nouvelles données dans le système et une simplicité de l'implémentation. Par conséquent, ils fournissent en général des prévisions plus précises pour les systèmes en-ligne. Cependant, ces systèmes sont très sensibles aux données sparses. En effet, si les données s'avèrent rares, il est difficile d'identifier des voisins fiables (à partir des items co-notés) et par conséquent la performance du système décroît. En revanche, les approches à base de modèles n'utilisent pas la totalité d'informations disponibles pour faire une prédiction [4, 5, 6]. Ils apprennent généralement off-line un modèle robuste des préférences des utilisateurs. Ces approches ont l'avantage de fournir des prévisions précises d'une manière rapide et d'être moins sensibles aux données manquantes. Toutefois, elles requièrent beaucoup de temps pour apprendre le modèle, ce qui les rend moins efficaces dans un système en-ligne, où de nouvelles données sont fréquemment rajoutées au système. Le processus de construction du modèle est basé sur les techniques variées telles que les réseaux bayésiens, le clustering, les arbres de décision. Dans le domaine du filtrage collaboratif il y a eu plusieurs tentatives proposant des solutions au problème des données sparses [7, 8, 9]. Des méthodes basées sur un modèle utilisent notamment les techniques de réduction de la dimension de l'espace des items ou le clustering dans l'espace des utilisateurs dans le but d'écarter les utilisateurs ou les items non représentatifs. Ainsi l'espace de représentation utilisateur-item est réduit et le taux de données manquantes est moins important comparé à l'espace de représentation original. Notons toutefois qu'en portant la réduction de la dimension et le clustering sur un des deux espaces, on tend à privilégier un espace sur un autre. {{Travail à effectuer}} Le défi principal de ce projet de thèse est de proposer une approche permettant de surmonter les difficultés décrites précédemment. Pour ce faire, nous proposons d'adopter une approche basée sur la classification croisée co-clustering [10, 11, 12] qui consiste à partitionner les deux espaces simultanément et par conséquent ne pas privilégier un espace sur un autre. Dans cette thèse, nous allons considéré deux approches foncièrement différentes. • une première statistique basée sur les modèles de mélanges par blocs [13] et, • une seconde algébrique basée sur la tri-factorisation de matrice non négative [14]. Ces deux approches ont montré leurs intérêts dans le cadre de la classification croisée pour traiter des données de grandes tailles sparses et dans une grande dimension. Néanmoins, elles ne sont pas développées en présence de données manquantes et méritent d'être utilisées dans ce contexte. Dans une première phase, les deux approches seront étendues pour la gestion de données manquantes et la gestion en-ligne dans les systèmes de recommandation. Des méthodes hybrides combinant les deux approches pourront être proposées. Dans une seconde étape, les algorithmes issues des deux approches seront comparés aux approches existantes à partir de données simulées et de données réelles afin de mesurer la performance de l'outil proposé.

Résumé du projet de recherche (Langue 2)

De nos jours on dispose via Internet d'une large variété de ressources, qui ont la particularité d'être hétérogènes et distribués et dont le volume est sans cesse croissant. Devant cette surabondance d'information, l'utilisateur devient incapable de gérer cette masse d'information et de repérer les items qui correspondent au mieux à ses attentes et dans ce contexte, le recours à des outils permettant de faciliter l'accès aux items pertinents s'avère crucial. Parmi les techniques possibles, on peut utiliser le filtrage collaboratif qui pallie le manque de données disponibles sur l'utilisateur en transformant le problème de l'apprentissage individuel en un apprentissage collaboratif. Ce projet présente un intérêt d'ordre théorique et appliqué. Il s'intègre dans une thématique qui est portée actuellement par une communauté très active au niveau international. L'objectif de ce projet de thèse est de proposer des techniques capables d'améliorer l'efficacité des outils de recommandation et dans un temps de calcul très réduit. En moyenne, les utilisateurs ne sont pas prêts à attendre une recommandation plus de deux secondes.

Informations complémentaires (Langue 1)

Des échanges avec des équipes de recherche à l’étranger (Italie, Hollande, Belgique et Brésil) pourront être envisagés.

Informations complémentaires (Langue 2)

Ce sujet s’adresse à des étudiants souhaitant contribuer au développement des systèmes de filtrage collaboratif. Cette recherche se déroulera au sein de l’équipe GFD (Gestion et Fouille de données) sous la direction de Mohamed Nadif et en co-direction avec Nicoleta Rogovschi. Le sujet de cette thèse nécessite des connaissances dans le domaine de l’apprentissage non supervisé et supervisé. Un bagage en mathématiques appliquées et une compétence avérée en programmation seront très appréciés. {{Réferences}} [1] John S. Bresse, David Heckerman, Carl Kadie – 1998. Emperical Analysis of Predictive Algorithms for Collaborative Filtering. [2] J. L. Herlocker, J. A. Konstan, A. Borchers, andJ. Riedl. An algorithmic framework for performing collaborative filtering. In SIGIR ’99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 230–237, New York, NY, USA, 1999. ACM. [3] G. Linden, B. Smith, and J. York. Amazon.com recommendations : Item-to-item collaborative filtering.IEEE Internet Computing, 7(1):76–80, 2003. [4] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In IJCAI ’99 : Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pages 688–693, San Francisco, CA, USA,1999. Morgan Kaufmann Publishers Inc. [5] B. Marlin. Modeling user rating profiles for collaborative filtering. Advances in Neural Information Processing Systems, 16:627–634, 2004. [6]D. M. Pennock, E. Horvitz, S. Lawrence, and C. L.Giles. Collaborative filtering by personality diagnosis:A hybrid memory and model-based approach. In UAI’00 : Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, pages 473–480, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. [7] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th annual conferenceon uncertainty in artificial intelligence, pages 43–52.Morgan Kaufmann, 1998. [8] F. Fouss, J.-M. Renders, A. Pirotte, and M. Saerens.Random-walk coEnjeuxmputation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on Knowledge and Data Engineering, 19(3):355–369, 2007. [9] Z. Huang, H. Chen, and D. Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems, 22(1):116–142, 2004. [10] Inderjit S. Dhillon, Subramanyam Mallela and Dharmendra S. Modha. Information-theoretic co-clustering. KDD 2003 : 89-98. [11] Gérard Govaert and Mohamed Nadif : Clustering with block mixture models. Pattern Recognition 36(2) : 463-473, 2003. [12] Lazhar Labiod and Mohamed Nadif : Co-clustering for Binary and Categorical Data with Maximum Modularity. ICDM : 1140-1145, 2011. [13] Gérard Govaert and Mohamed Nadif : Block clustering with Bernoulli mixture models : Comparison of different approaches. Computational Statistics & Data Analysis 52(6) : 3233-3245, 2008. [14] Lazhar Labiod and Mohamed Nadif : : Co-clustering under Nonnegative Matrix Tri-Factorization. ICONIP(2) : 709-717, 2011.