Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach

Présentation
Groupe de lecture RI - 20/11/2020

Par Yann DUPERIS

Doctorant encadré par
Bernard ESPINASSE, Adrian-Gabriel CHIFU et Sébastien FOURNIER

Présentation de l'intervenant

- Yann Duperis (yann.duperis@lis-lab.fr) ;
- Thèse CIFRE
  - Un Système d'aide à la constitution de consortiums d'entreprises compétents pour un appel d'offre : une approche basée sur le traitement des langues et les ontologies ;
  - LIS > Équipe R2I
  - Encadré par :
    - Bernard ESPINASSE ;
    - Sébastien FOURNIER ;
    - Adrian-Gabriel CHIFU.
- Sujets de recherche :
  - Systèmes de recommandation adaptés au recrutement ;

Objectif de la présentation

- Contextualiser les travaux de recherche entrepris ;
- Présenter l'approche choisie ;
- Présenter l'évaluation proposée.

I. Contexte de l'article

- Publié en 2017 dans Knowledge-Based Systems ;
- Rédigé par :
  - Shuo Yang [a];
  - Mohammed Korayem [b] ;
  - Khalifeh AlJadda [b];
  - Trey Grainger [b];
  - Sriraam Natarajan [a].
- Au sein des organisations suivantes :
  - a) Université de l'Indiana ;
  - b) CareerBuilder.

I. Contexte de l'article

- CareerBuilder :
  - Job Board américains ;
  - +60 millions de CV numérisés ;
  - Contributions scientifiques pour :
    - **Recommandation d'offres d'emploi** ;
    - Construction de taxonomies standardisées : noms de métiers [4] (> 4K) et compétences (> 26K) [5] ;
    - Plongement de « Graphe économique » (travaux similaires à ceux de LinkedIn [2] et [3]) [1] ;
  - Capitalisation de grandes quantités de données :
    - CV numérisés ;
    - Données d'interaction entre profils et offres (consultation, candidature, etc.) ;

I. Contexte de l'article

- Questions scientifiques :
  - **Q1** : l'utilisation d'un système de recommandation hybride apporte-t-elle plus que le filtrage basé sur le contenu seul . ?
  - **Q2** : est-il possible de privilégier la précision sur le rappel, sans trop pénaliser les autres métriques ?

II. Approche proposée

- Système de recommandation d'offres d'emploi (*Job Recommender System*) hybride :
  - *Content-based* : exploitation des propriétés des profils, et des offres d'emploi ;
  - *Collaborative-filtering* : exploitation du graphe des interactions pour apprendre les relations statistiquement pertinentes entre profils et offres ;

- Application de l'**apprentissage statistique de relations** (*Statistical Relational Learning*) :
  - Application au domaine d'un algorithme publié en 2011 [6] ;
  - Pondération précision/rappel (précision préférée) paramétrable (**contribution**).

II. Approche proposée

- Modélisation de **dépendances statistiques** entre les prédicats

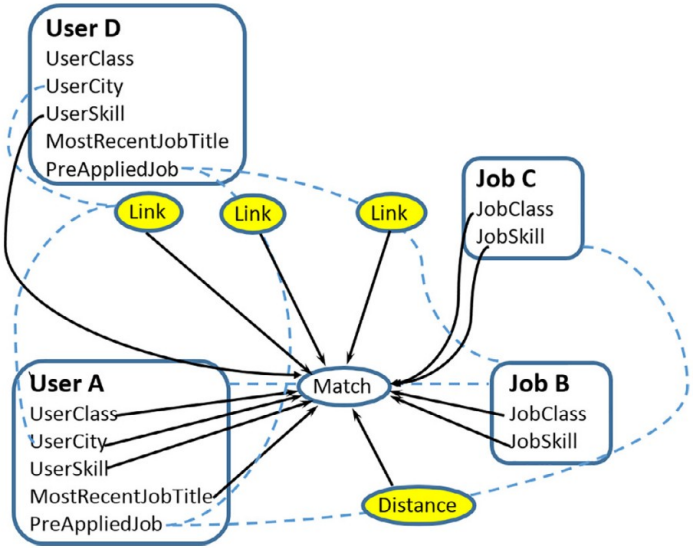

Fig. 2. Template Model of a Sample RDN. The target is *Match*(*UserA*, *JobB*) while the related objects are User D (introduced by the link nodes) and previous applied Job C. Note that D and C are first-order variables which could have multiple groundings for different target user–job pairs.

II. Approche proposée

- Prédicats utilisés :
  - **Match(UserID, JobID)** ;
  - JobSkill(JobID, SkillID) ;
  - UserSkill(UserID, SkillID) ;
  - JobClass (JobID, ClassID) ;
  - UserClass(UserID, ClassID) ;
  - PreAppliedJob(UserID, JobID) ;
  - UserJobDis(UserID, JobID, Distance) ;
  - UserCity(UserID, CityName) ;
  - MostRecentCompany(UserID, CompanyID) ;
  - MostRecentJobTitle(UserID, JobTitle).

- CommSkill(UserID1, UserID2) ;
- CommClass (UserID1, UserID2) ;
- CommCity(UserID1, UserID2) :

Filtrage collaboratif

II. Approche proposée

- Prédicats utilisés :
  - **Match(UserID, JobID)** ;
  - JobSkill(JobID, **SkillID**) ;
  - UserSkill(UserID, **SkillID**) ;
  - JobClass (JobID, **ClassID**) ;
  - UserClass(UserID, **ClassID**) ;
  - PreAppliedJob(UserID, JobID) ;
  - UserJobDis(UserID, JobID, Distance) ;
  - UserCity(UserID, CityName) ;
  - MostRecentCompany(UserID, CompanyID) ;
  - MostRecentJobTitle(UserID, **JobTitle**).

- CommSkill(UserID1, UserID2) ;
- CommClass (UserID1, UserID2) ;
- CommCity(UserID1, UserID2) :

Filtrage collaboratif

Taxonomies construites suite au travaux présentés dans [4] et [5]

- Probabilité de vérité du prédicat cible **Match(UserID, JobID)** pour un exemple ***x*** obtenue :
  - Fonction de régression ψ(x) : arbres relationnels de régression (*Relational Regression trees* [7]) ;
  - Application d'une fonction sigmoïde sur ψ pour représenter la distribution de probabilité ;

$$LL = \sum_i \log P(y_i = \hat{y}_i; \mathbf{X}_i) = \sum_i \log \frac{1}{1 + \exp(-\hat{y}_i \cdot \psi(y_i = \hat{y}_i; \mathbf{X}_i))}$$

II. Approche proposée

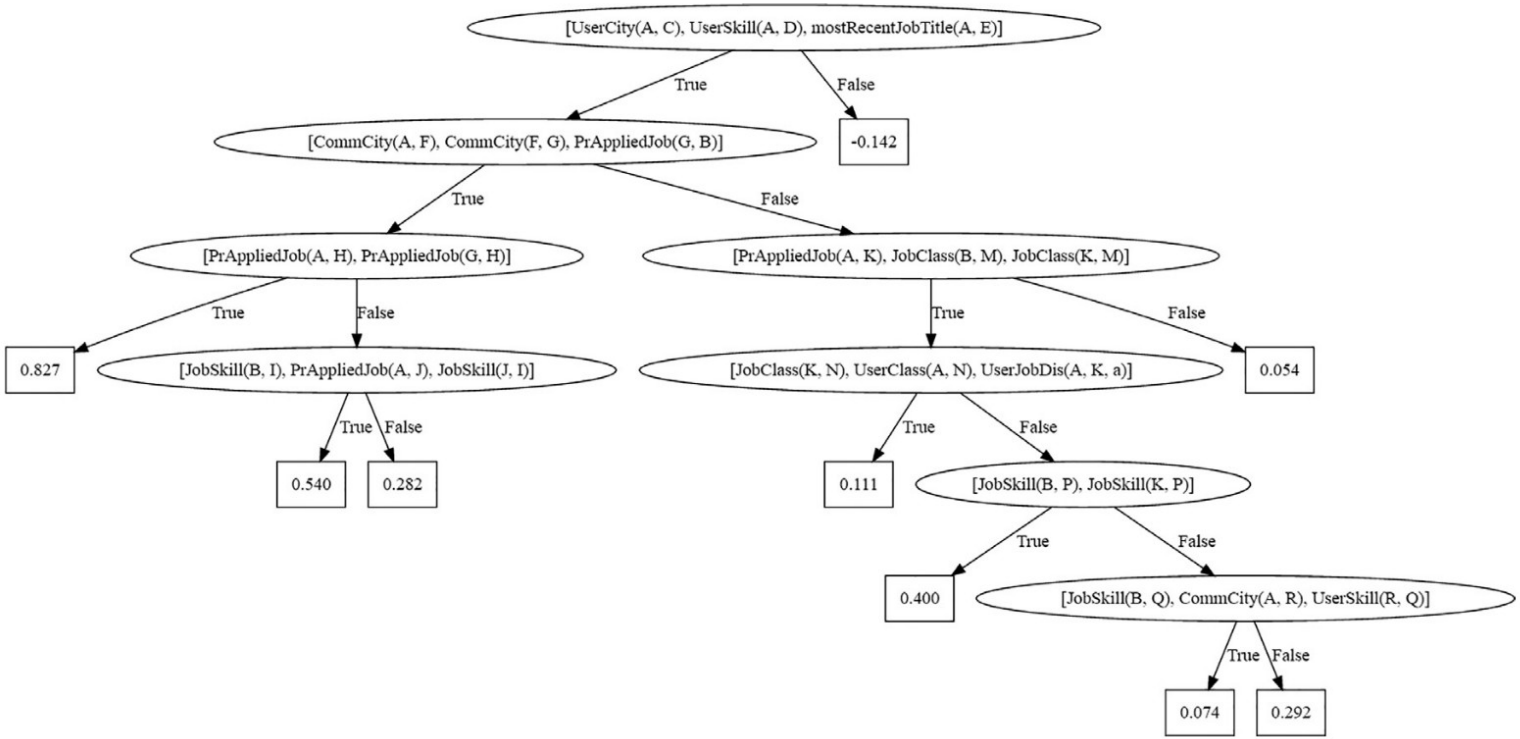

Fig. 4. Sample Regression Tree.

II. Approche proposée

- Gradient

$$\frac{\partial LL(\mathbf{x})}{\partial \psi\,(y_i = \hat{y}_i; \mathbf{X}_i)} = I(\hat{y}_i = Match) - P(y_i = Match; \mathbf{X}_i)$$

- Problème :
  - Toutes le erreurs sont pondérées de la même façon ;
  - L'expertise du domaine a révélé que **la précision était préférable au rappel**.

- Approche de gestion de coûts **paramétrable** :
  - Paramètre α : impact des faux négatifs ;
  - Paramètre β : impact des faux positifs ;

Table 1
Cost Matrix.

		Actual Class	
		True	False
Predicted Class	True	0	$I(\hat{y}_i = 1) - \frac{P(y_i=1;\mathbf{X}_i)}{P(y'=1;\mathbf{X}_i)+P(y'=0;\mathbf{X}_i)\cdot e^{-\beta}}$
	False	$I(\hat{y}_i = 1) - \frac{P(y_i=1;\mathbf{X}_i)}{P(y'=1;\mathbf{X}_i)+P(y'=0;\mathbf{X}_i)\cdot e^{\alpha}}$	0

- Données
  - Historique de **4 mois** d'utilisation de CareerBuilder (candidatures) ;
  - 707 820 offres d'emploi ;

Table 2
Feature Space.

Variable Name	SkillID	ClassID	Distance
Num of Instances	8534	1867	4
Variable Name	CityName	CompanyID	JobTitle
Num of Instances	22,137	1,154,623	823,733

III. Évaluation

- Données
  - 3 domaines retenus pour l'évaluation ;
  - Forte asymétrie des effectifs dans les classes ;

Table 3
Domains.

JobTitle	Training			Test		
	pos	neg	facts	pos	neg	facts
Retail Sales Consultant	224	6973	13,340,875	53	1055	8,786,938
Case Manager	387	35,348	13,537,324	87	5804	8,815,216
District Manager	358	16,014	13,396,635	87	3521	8,798,522

III. Évaluation

- Paramètres :
  - Régression réalisée à l'aide de **20 arbres** ;
  - **8 feuilles maximum** ;
  - 1 modèle par classe d'utilisateur (~ secteur d'activité) ;
  - Évaluation avec et sans prédicats adaptés au filtrage collaboratif (Configuration étiquetée **HR** dans les résultats pour Content-Based filtering + Collaborative filtering)

- Résultats

Table 4
Results.

Job Title	Approach	FPR	FNR	Precision	Recall	Accuracy	AUC-ROC
Retail	Content-based Filtering (CF)	0.537	0.321	0.060	0.679	0.473	0.628
Sales	Cost-sensitive CF ($\alpha 0 \beta 2$)	**0.040**	0.868	0.143	0.132	**0.921**	0.649
Consultant	Hybrid Recommender (HR)	0.516	**0**	0.089	**1.0**	0.509	**0.776**
	Cost-sensitive HR ($\alpha 0 \beta 2$)	0.045	0.906	0.096	0.094	0.914	0.755
	Cost-sensitive HR ($\alpha 0 \beta 1$)	0.113	0.623	**0.144**	0.377	0.863	0.772
Case	Content-based Filtering (CF)	0.220	0.184	0.053	0.816	0.781	0.861
Manager	Cost-sensitive CF ($\alpha 0 \beta 2$)	0.084	0.609	0.066	0.391	0.909	0.847
	Hybrid Recommender (HR)	0.239	**0**	0.059	**1.0**	0.765	**0.911**
	Cost-sensitive HR ($\alpha 0 \beta 2$)	**0.037**	0.736	**0.096**	0.264	**0.952**	**0.911**
District	Content-based Filtering (CF)	0.427	0.195	0.045	0.805	0.579	0.746
Manager	Cost-sensitive CF ($\alpha 0 \beta 2$)	0.017	0.920	**0.104**	0.080	0.961	0.745
	Hybrid Recommender (HR)	0.439	**0**	0.053	**1.0**	0.572	0.817
	Cost-sensitive HR ($\alpha 0 \beta 2$)	**0.013**	0.977	0.042	0.023	**0.964**	0.812
	Cost-sensitive HR ($\alpha 0 \beta 1$)	0.068	0.678	**0.104**	0.322	0.917	**0.825**

Merci pour votre attention !

III. Évaluation

- Résultats :
  - ☑ **Q1** : l'ajout de prédicats de filtrage collaboratif améliorent la performance ;
  - ☑ **Q2** : l'utilisation de paramètres de coût asymétriques permet d'améliorer grandement la précision ;
  - Les performances semblent homogènes sur les trois secteurs d'activité présentés ;
  - Dans ~90-95 % des cas, le système parvient à prédire l'acceptation ou le rejet d'une candidature ;
  - La forte asymétrie entre les classes et l'absence de matrice de confusion peuvent amener à se questionner sur les performances pour chacune des classes (ne prédit-on pas juste le rejet pour maximiser les métriques ?).

Bibliographie

- [1] V. S. Dave, M. Al Hasan, B. Zhang, K. AlJadda, and M. Korayem, “A **combined representation learning approach for better job and skill recommendation**,” Int. Conf. Inf. Knowl. Manag. Proc., pp. 1997–2006, 2018.

- [2] R. Ramanath et al., “**Towards deep and representation learning for talent search at LinkedIn**,” Int. Conf. Inf. Knowl. Manag. Proc., pp. 2253–2262, 2018.

- [3] J. Weiner. **The future of LinkedIn and the Economic Graph**. LinkedIn Pulse, 2012

Bibliographie

- [4] Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M., & Kang, T.S. (2015). "**Carotene: A Job Title Classification System for the Online Recruitment Domain**". 2015 IEEE First International Conference on Big Data Computing Service and Applications, 286-293.

- [5] Javed, F., Hoang, P., Mahoney, T., & McNair, M. (2017). **Large-Scale Occupational Skills Normalization for Online Recruitment**. AAAI.

- [6] Natarajan, S., Khot, T., Kersting, K., Gutmann, B., & Shavlik, J. (2011). **Gradient-based boosting for statistical relational learning: The relational dependency network case**. Machine Learning, 86, 25-56.

Bibliographie

- [7] Blockeel, H. (1998). **Top-Down Induction of First Order Logical Decision Trees**. AI Commun., 12, 119-120.