Fouille des données pour l'extraction automatique des représentations

Présentation : Julien Velcin (laboratoire ERIC, Université de Lyon)

Résumé :

La manière dont nous percevons le monde qui nous entoure a longtemps été influencée par notre environnement immédiat (famille, amis, professeurs) et par les médias traditionnels (radio, presse) [1]. On assiste depuis l'arrivée d'Internet et des réseaux sociaux virtuels à un bouleversement important, non pas nécessairement des mécanismes cognitifs d'élaboration des représentations, mais des modes d'acquisition de l'information qui concourent à cette élaboration. Les chercheurs en psychologie et en sociologie ont ainsi une formidable occasion de développer de nouvelles méthodes, au-delà des techniques de sondage classiques, pour capturer l'image que les individus se font des événements, des personnages, des institutions, bref de tous ces « objets » qui peuplent leur univers mental. La création de ces nouvelles méthodes nécessite l'implication de chercheurs en intelligence artificielle, en représentation des connaissances, en analyse et fouille des données. Face à ces données disponibles sur le Web, principalement textuelles (forums de discussion, blogs, réseaux sociaux), une approche possible consiste à appliquer des algorithmes d'apprentissage automatique non supervisés afin de faire émerger, principalement par récurrence statistique, des « concepts » qui capturent l'essence de l'information qui circule sur la toile et qui serviront de base aux représentations. A ce sujet, la littérature scientifique propose de nombreuses méthodes qui cherchent toutes, fondamentalement, à re-décrire les textes en fonction d'un ensemble (réduit) de thématiques (*topics*) : décompositions factorielles [2,3], méthodes à base de centroïdes [4], modèles graphiques bayésiens [5,6]. La recherche de ces thématiques touche, sans réellement réussir à les résoudre complètement, à des problèmes classiques de classification non supervisée (*clustering*), tels que la recherche du bon nombre de classes, la description et l'évaluation des catégories (*clusters*) de textes.

Dans cette présentation, je propose de décrire les quelques contributions qui ont été récemment réalisées sur ces sujets au sein du laboratoire ERIC. Pour cela, je commencerai par rappeler la problématique, assez classique maintenant, d'extraction de thématiques, en soulignant les liens avec le problème de classification en général et plus particulièrement avec celui du regroupement conceptuel (*conceptual clustering*) [7]. Je parlerai ensuite d'une contribution sur la caractérisation des groupes de textes à l'aide d'expressions fréquentes [8], qui a notamment donné naissance à un (récent) prototype logiciel pour l'analyse des commentaires de la presse en ligne. Je présenterai également des contributions sur l'évaluation des thématiques produites, évaluation qui représente un réel verrou scientifique aujourd'hui [9,10]. Nous proposons de rapprocher les thématiques obtenues par les méthodes citées plus haut avec une base de connaissance linguistique de type WordNet [11]. Nous avons montré empiriquement que la nouvelle méthode proposée, à base de sous-arbres de concepts construits à partir d'une hiérarchie de *synsets*, permet d'obtenir automatiquement des scores de qualité fortement corrélés avec les scores obtenus lorsque l'évaluation est réalisée par des humains [12]. Enfin, je ferai le lien, naturel, avec le projet ANR ImagiWeb qui débute à peine et dans lequel nous chercherons à extraire automatiquement l'image des entités de diverses natures qui peuplent le Web.

Références bibliographiques

[1] Lippmann, W., Public opinion. Free Press 1997 (original édition, 1922).
[2] Berry, M.W., Dumais, S.T. and O'Brien, G.W., Using linear algebra for intel ligent information retrieval. 1994. Technical Report UT-CS-94-270.
[3] Paatero, P. and Tapper, U., Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, vol. 5 (2), 1994. Wiley Online Library.
[4] Velcin, J. and Ganascia, J.-G., Topic Extraction with AGAPE. Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA), 2007.
[5] Hofmann, T., Probabilistic latent semantic indexing. Proceedings of the 22th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. 50-57, 1999.
[6] Blei, D.M., Ng, A.Y. and Jordan, M.I., Latent dirichlet allocation. Journal of Machine Learning Research, vol. 3, 2003. JMLR. Org.
[7] Michalski, R.S. and Stepp, R.E., Learning from observation: Conceptual clustering. Machine Learning : An artificial intelligence approach, vol. 1, pp. 331-363. Morgan Kaufmann, 1983.
[8] Rizoiu, M.A., Velcin, J. and Chauchat, J.H., Regrouper les données textuelles et nommer les groupes à l'aide de classes recouvrantes. Actes des 10ème journées francophones en Extraction et Gestion des Connaissances (EGC 10), Hammamet, Tunisie 2010.
[9] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. and Blei, D. M., Reading Tea Leaves: How Humans Interpret Topic Models. Neural Information Processing Systems, 2009
[10] Newman, D., Lau, J. H., Grieser, K. and Baldwin, T,. Automatic Evaluation of Topic Coherence. In The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 100-108. June 2010.
[11] Miller, G. A., WordNet: a lexical database for English. Communications of the ACM. 38(11): 39-41, 1995.
[12] Musat, C., Velcin, J., Trausan-Matu, S. and Rizoiu, M.A., Improving Topic Evaluation Using Conceptual Knowledge. Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), 2011.