Sujet : **Collecte et modélisation probabiliste d'opinions et autres données, dans le but d'optimiser un système de paris en ligne**

Stagiaire : (étudiant de M2)

Co-encadrement : SanJuan E, El-Bèze M, El Azouzi R., Cossu J-V

Dates : de début février à fin juillet 2014.

Ce stage s'inscrit dans un projet ambitieux qui vise à explorer plusieurs méthodes permettant d'optimiser des paris en lignes. Le problème est intéressant non pas tant du point de vue de l'application retenue que des méthodes qui peuvent être employées et combinées à cette fin. Faisant l'hypothèse que les paris concernent un match de football, on envisage un certain nombre d'étapes comme les suivantes

- Tirer parti de toutes les données que l'on a pu réunir sur chacun des membres de chaque équipe pour proposer un modèle simple permettant de conclure dans un premier temps à la supériorité potentielle d'une équipe sur l'autre ;
- Si on accepte l'idée qu'une équipe est plus que la réunion de ses membres, proposer une modélisation un peu plus élaborée tenant compte de la cohésion (ou non) des 2 équipes, au moins en prenant les membres 2 à 2 ;
- Effectuer une analyse automatique des articles (multi lingues) que la presse spécialisée peut produire avant une rencontre sportive internationale, pour en prédire l'issue.
- Exploiter les contenus échangés sur les réseaux sociaux afin de donner plus de poids à une hypothèse de victoire ou de défaite.

Ce sont, bien entendu, les deux derniers points qui intéressent le plus les chercheurs actifs en TALNE.

Deux points méritent d'être signalés :

Une école anglaise présente la théorie des probabilités comme l'expression de paris.

Pour donner une estimation des bornes inférieures et supérieures de l'entropie moins grossière que celle proposée par (Shannon 1951), et reprise par (Bimbot et al., 2001), l'approche dénommée « *Gambling approach* » proposée par (Cover et King 1978) s'inscrit dans une vision originale de la théorie des probabilités dont on peut s'inspirer[1] pour estimer les chances de succès et les risques pris en pariant sur un ensemble de matches selon des modèles plus ou moins complexes.

Le montage du protocole expérimental fait partie du sujet de stage. La constitution du corpus incluant la partition classique (apprentissage, développement et test) ne devrait pas poser de problèmes. En effet, sur Internet, il est assez évident de trouver, dans de multiples archives, de nombreux écrits produits avant pendant et après les matches du passé, ainsi que le résultat des rencontres. Mais pour ne pas s'éparpiller, il conviendra de suivre les consignes données dans la section suivante

[1] On consultera avec intérêt la page du wikipedia Anglais concernée : http://en.wikipedia.org/wiki/Statistical_association_football_predictions

Plan de travail

Par souci de réalisme, nous avons choisi de ramener les objectifs énoncés ci-dessus à un périmètre raisonnable. Dans le cadre de ce stage, nous nous limiterons aux paris sur les matchs de football en Ligue 1 (L1) au royaume uni (UK) et en France. Nous utiliserons comme source de données *WikiPedia* et *DbPedia* avec l'ensemble des pages référencées. On pourra s'en tenir aux sous objectifs suivants :

- Extraire de *WikiPedia* le sous-réseau concernant l'ensemble des équipes et des joueurs de L1. Cette extraction devra se faire automatiquement à partir d'un *dump* récent de *WikiPedia*. Il s'agit d'abord d'un problème de parcours orienté de graphe.
- Compléter le réseau de contenus ainsi obtenu (les sommets sont des pages, les arêtes sont les liens internes à *WikiPedia*) avec les informations structurées RDF de *DBPedia* (âge du joueur, budget d'un équipe, entraîneur etc .)
- Élaborer et évaluer un premier système de paris en ligne n'utilisant que les propriétés du réseau obtenu en les pondérant selon les informations du *DbPedia*.
- Élaborer un deuxième système de paris en ligne qui lui n'utilise que le contenu textuel des pages et en particulier la détection des opinions répertoriées dans *WikiPedia*. Cette recherche d'opinion devra s'étendre aux pages externes référencées par *WikiPedia*.
- Comparer et combiner les deux systèmes.

Néanmoins, le sujet pourra être affiné de telle sorte que l'accent sera mis sur tel ou tel aspect, selon les qualités et la motivation du stagiaire qui sera retenu.

Les performances des systèmes seront évaluées sur les deux L1. Les données UK étant bien plus volumineuses. L'utilisation de *WikiPedia* a deux avantages, d'abord il délimite le cadre de recherche et le risque de s'enliser dans les problématiques techniques et légales d'exploration du Web. Ensuite il permet de tester l'hypothèse de recherche que *WikiPedia* concentre une grande partie de « l'information utile » du Web.

Notes :

Dans le cadre du projet ODTM soumis à l'ANR, de multiples cas d'étude sont envisagés dont au moins un portant simultanément sur la production de données publiques sur le Web au travers de *Wikipedia* et des projets liés, ainsi que sur le domaine de l'événementiel médiatique que constitue la coupe du monde de football de Rio (2014). En cas d'acceptation du projet ODTM, un financement pour une prolongation en thèse serait ainsi accessible. Une deuxième piste réside dans la possibilité d'envisager une thèse en co tutelle avec une des universités brésiliennes avec lesquelles nous avons des liens dans le cadre du GdRI WebSciences.

Références Bibliographiques :

Bellot P., Doucet A, Geva S, Gurajada S, Kamps J, Kazai G, Koolen M, Mishra A, Moriceau V, Mothe J, Preminger M, SanJuan E, Schenkel R, Tannier X, Theobald M, Trappett M, Wang Q : *Overview of INEX 2013*. CLEF 2013: 269-281

Bimbot F., El-Bèze M., Igounet S., Jardino M., Smaïli K., Zitouni I. *: An alternative scheme for perplexity estimation and its assessment for the evaluation of language models*. 1-13, Computer Speech and Language, Volume 15, Issue 1, January 2001, pp. 1–13

Cover, T. M., *and* R. C. King *(1978),* "A *Convergent Gambling Estimate* of the *Entropy* of *English*", *IEEE Transactions* on *Information Theory* (24)4 (July) pp.413-419

Hill I.D. (1974), *Association football and statistical inference*. Applied statistics, 23, 203-208

C.E. Shannon. "*Prediction and Entropy of Printed English*", in *Bell System Technical Journal* 30:50-64. 1951.