A semi-supervised recommender system to predict online job offer performance

Julie Séguéla[1,2], Gilbert Saporta[1,⋆]

1. CNAM-CEDRIC, 292 rue Saint-Martin, 75003 Paris

2. Multiposting.fr, 23 rue d'Aumale, 75009 Paris

⋆ Corresponding author: gilbert.saporta@cnam.fr

Keywords: recommender system, similarity measure, PLS, feature extraction

In domains such as Marketing, Advertising or even Human Resources (sourcing), decision-makers have to choose the most suitable channels according to their objectives when starting a campaign. With the expansion of internet to advertise, the number of potential channels (and targets) is exponentially growing. Today, a great challenge common to several research domains is the development of intelligent tools to support (or to replace) users in their choices.

In this work, we are presenting a recommender system predicting the ranking of job boards (job search web sites) in terms of job posting performance (or return). A job posting is a job offer published on the internet, possibly on several job boards simultaneously. Performance is assumed to be the number of applications received on the job board. Given the complexity of our data, the recommender system has to be very specific, and can be considered as a particular case of recommender systems commonly encountered in the literature (Adomavicius, Tuzhilin, 2005). The aim of such systems is to help users to find items that they should appreciate from huge catalogues. Most of applications concerns very large datasets (thousands of users and items) where users have already rated several items according to their preferences. Our application concerns only about thirty users (job boards) and we focus on the new item (job posting) problem: ratings (returns) are estimated for items which have never been rated by users yet. Moreover, rating variability inside and between users is high, instead of being limited to integer values between 0 and 5 (or 0 and 10) as usually.

Experiments are made on a real dataset provided by *Multiposting.fr*, an online job posting solution which distributes jobs simultaneously on several dozens of job boards. The ratings to be estimated are the number of applications received per day on the job board studied for a new job posting. Job postings are described by thousands of features including structured and unstructured data which have to be handled simultaneously. Unstructured data are job description texts from which features are extracted thanks to an information retrieval technique. Structured data are job characteristics (contract type, industry, occupation, etc.) represented by categorical variables and location characteristics represented by quantitative variables.

The recommender system is presented in Figure 1. PLS components, linear combination of posting features resulting from NIPALS algorithm (Wold , 1966), are computed so as to explain as much as possible actual return. Then, supposing that similar postings have similar performance for a given job board, similarity measures based on PLS components are computed between job postings. Finally, expected return of item i_1 for user u is estimated thanks to an aggregating function computed on item neighborhood given by:

$$R_{u,i_1} = \frac{\sum_{i_k \in K} sim(i_1, i_k) \times r_{u,i_k}}{\sum_{i_k \in K} sim(i_1, i_k)} \qquad (1)$$

where K is the set of items defined by the $|K|$ nearest neighbors of i_1 with respect to the used similarity function, and r are actual returns. The system can be considered as a semi-supervised algorithm insofar as only relevant features are taken into account to estimate posting performance.

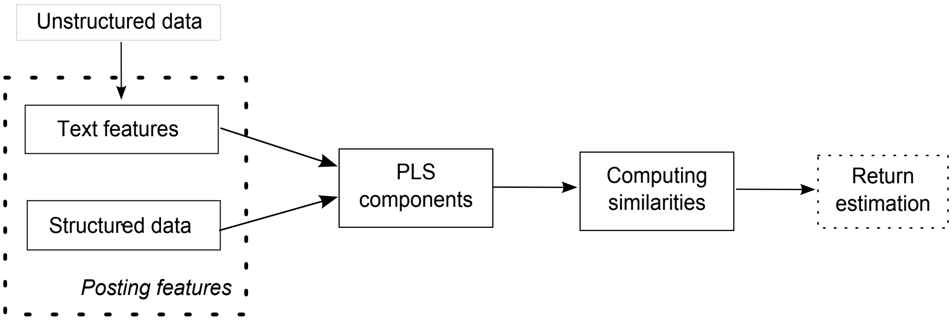

Figure 1: Recommender system overview

In a previous work (Séguéla, Saporta, 2011), in which only textual data were used, we have shown that this semi-supervised system was more efficient (regarding to the comparison criterion) than a basic PLS regression model or a "naive" system where similarities are computed directly on text feature vectors without a model. Among feature extraction techniques that were tested, Latent Semantic Indexing (Deerwester *et al.*, 1990), was the best representation technique for this kind of data whatever approach used.

According to the previous experiments, the system uses LSI to extract features from offer texts. Structured data are added to the model which allows to improve the quality of predictive algorithm. The choice of similarity measure is discussed by comparing several weighting functions (in particular gaussian and exponential) of euclidean distance and several parameter values.

References

Adomavicius G., Tuzhilin A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734-749.

Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41, 391-407.

Séguéla J., Saporta G. (2011). A semi-supervised hybrid system to enhance the recommendation of channels in terms of campaign ROI, in: *CIKM'11: Conference on Information and Knowledge Management*, Glasgow, UK.

Wold H. (1966). Estimation of principal components and related models by iterative least squares, in: *Multivariate Analysis*, P. R. Krishnaiaah Editor, New York: Academic Press, 391-420.