Automatic Categorization of Job Postings

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

[1] Multiposting.fr
    33 rue Réaumur, Paris, France, *jseguela@multiposting.fr*
[2] Laboratoire CEDRIC, CNAM
    292 rue Saint Martin, Paris, France

Abstract. Since the beginning of the Nineties, the increasing proportion of job vacancies which are published on the internet has led to a multiplication of on-line job search sites (job boards). Consequently, the need to assess job board performance has become a priority for recruiters. But an important issue is that each job board has a specific nomenclature to describe the type of the post. As part of the modelisation of job posting performance, we need to establish a common classification for the "function" criterion. To achieve that goal, we are working on a corpus of manually labelled texts of job offers, and we are proposing a method to categorize the texts into a two-level predefined classification of occupations.

First, a preprocessing adapted to the particularities of job offer texts is performed (stemming, use of a specific dictionnary,...). Then, we are reducing the dimensionality of the problem thanks to a feature selection method (we can see a comparative study in Yang and Pedersen (1997)). The Vector Space Model is used to represent the texts and the terms are weighted with a function depending on the position of the term in the text (title or mission description). Finally, classification of job postings is performed with SVM (e.g. Joachims (1997)). Popular performance measures such as recall and precision are used and adapted to our context with a weighting for errors according to the seriousness of misclassification. In addition, we are exploring the effects on the classification quality of the Probabilistic Latent Semantic Analysis, another dimensionality reduction method which allows to address the issue of synonymy (Hofmann (1999)).

Our method could also be applied to achieve automatic labelling of job postings according to the nomenclature of a particular job board.

Keywords: text categorization, Latent Semantic Analysis, Support Vector Machine, job posting

References

HOFMANN, T. (1999): Probabilistic Latent Semantic Analysis. In: *Proceedings of UAI'99, Uncertainty in Artificial Intelligence*. Stockholm.

JOACHIMS, T. (1997): Text categorization with Support Vector Machines: Learning with many relevant features. In: *Machine Learning: ECML-98*. Springer, Berlin, 137–142.

YANG, Y. and PEDERSEN, J. P. (1997): A comparative study on feature selection in text categorization. In: *Proceedings of ICML'97, International Conference on Machine Learning*. Nashville, US, 412–420.