ABSTRACT

The proliferation of the World Wide Web and the other Internet services in the past few years intensifies the need for developing systems that help users to cope with the enormous amount of text that is available online. *Information extraction* systems, that is, systems that locate and pull relevant fragments out of domain-specific collections of text documents, seem to be a promising way to deal with the information explosion. *Machine learning* techniques facilitate the development of information extraction systems and their portability to new domains of interest. Information extraction using machine learning techniques is a typical *Web mining* problem, since the task is to learn extraction rules that can effectively recognize relevant text fragments within Web documents.

This dissertation demonstrates the effectiveness of combining information extraction systems using *voting* and *stacked generalization (stacking)*. The motivation derives from the opportunity to obtain higher extraction performance at meta-level, by exploiting the disagreement in the predictions of the information extraction systems that are employed at base-level. Existing combination techniques primarily focus on classification. However, information extraction is not naturally a classification problem. A new methodology is proposed for combining information extraction systems through voting and stacking. The proposed methodology facilitates the combination of a wide range of systems, since only their output is combined, without taking into account how each system is implemented or models the extraction task. Information extraction is transformed to a common classification problem at meta-level, allowing the applicability of voting and stacking techniques.

The effectiveness of *voting* is initially investigated for combining multiple information extraction systems at meta-level. Extensive experiments were performed in a variety of domains using well known information extraction systems at base-level. The results demonstrate the effectiveness of voting with probabilistic estimates of correctness in the output of the base-level systems, as long as a probability threshold is set for deciding whether to accept a prediction at meta-level. Voting was effective on most domains in the experiments, outperforming the best base-level systems.

The effectiveness of *stacking* is then investigated for combining multiple information extraction systems at meta-level. The basic idea is to combine well known information extraction systems with a common classifier at meta-level, such as a decision-tree classifier or a Naïve Bayes classifier. Cross-validation takes place on the base-level dataset, which consists of text documents annotated with relevant information, in order to create a meta-level dataset that consists of feature vectors. A common classifier is then trained using the new vectors. Results demonstrate the effectiveness of stacking using probabilities in the output of the base-level systems. Stacking was consistently effective in all examined domains, always outperforming the best base-level information extraction systems. Comparing against voting, stacking performs comparably or better, while always obtaining more accurate predictions at meta-level.

Particular emphasis was also given to analyzing the results obtained by voting and stacking, aiming to investigate the sources of their success in information extraction tasks. The analysis showed that voting and stacking successfully exploit the disagreement in the output of the base-level systems, towards better results at meta-level. Stacking, however, proved to be better than voting, even when all base-level systems predict identically.

This dissertation contributes to the direction of realizing the high potential of combination methods in the context of accurately identifying relevant items of information on the abundant of computerized text, aiming at a method easily adapted to new domains.

SUBJECT AREAS: information extraction, machine learning

KEYWORDS: extraction, learning, voting, stacked generalization, web