Chapter XLI
Performance Evaluation
Measures for Text Mining

ABSTRACT

The purpose of this chapter is to provide an overview of prevalent measures for evaluating the quality of system output in seven key text mining task domains. For each task domain, a selection of widely used, well applicable measures is presented, and their strengths and weaknesses are discussed. Performance evaluation is essential for text mining system development and comparison, but the selection of a suitable performance evaluation measure is not a straightforward task. Therefore this chapter also
Performance Evaluation Measures for Text Mining

attempts to give guidelines for measure selection. As measures are under constant development in many task domains and it is important to take the task domain characteristics and conventions into account, references to relevant performance evaluation events and literature are provided.

INTRODUCTION

Performance evaluation is used to establish the relative performance of different approaches, to develop existing systems further, and to track the improvements over time. In this chapter, the topic is discussed from the viewpoint of the quality of the text mining system output. When selecting an evaluation protocol and a related performance evaluation measure, it is crucial to fully understand all the choices involved.

The selection of a suitable performance evaluation measure is not a straightforward task. There are a number of options available depending on the text mining task, the data and the evaluation goal in question. While there are simple measures that can be applied to a wide variety of task domains, they are in many cases not reliable indicators of actual system performance. Poor choice of a performance evaluation measure may have an adverse effect on system development. A suitable performance evaluation measure must take into account the specifics of the task domain in question.

The focus of this chapter is on text mining as represented by the task domains of text classification, information extraction, text segmentation, text ranking, information retrieval, text clustering and text summarization. For each task domain, a selection of widely used, well applicable measures evaluating the quality of the system output are presented, their properties are discussed, and references for further information are offered. If the use of performance evaluation measures in a given task domain has been implicitly directed by large international competitive evaluation events, the measures used in these events are described.

In the next section, terminology related to performance evaluation is introduced, alternatives to accomplish performance measurement are described and the chapter is positioned with respect to previous studies. Then, in the following four sections, the focus is on four distinct types of performance evaluation measures: classification, ranking, text clustering and text summarization measures. For each of these, the applicable measures are presented and, when appropriate, specific application domains and their specialized measures are discussed. When appropriate, performance evaluation measures are described as real-value functions \( p(Y, f(X)) \) comparing the system output \( f(X) \) produced from the input \( X \) to the correct output \( Y \). Finally, concluding remarks for this chapter are provided.

BACKGROUND

The evaluation of text mining systems can be broadly divided with respect to its goals into performance evaluation, adequacy evaluation and diagnostic evaluation (Hirschman & Thompson, 1997). Performance evaluation, the focus in this chapter, measures the system performance in one or more specific areas, typically producing numeric values. By contrast, adequacy evaluation aims at establishing to what extent the system satisfies user needs in a particular task and diagnostic evaluation determines the correctness of the system output, typically with respect to a test suite. These latter aspects of evaluation are not considered in more detail in this chapter; for further information, see, for example, Hirschman and Thompson (1997).