Measures for Benchmarking Semantic Web Service Matchmaking Correctness

Ulrich Küster and Birgitta König-Ries

Institute of Computer Science, Friedrich-Schiller-University Jena
D-07743 Jena, Germany
{Ulrich.Kuester|Birgitta.Koenig-Ries}@uni-jena.de

Abstract. Semantic Web Services (SWS) promise to take service oriented computing to a new level by allowing to semi-automate time-consuming programming tasks. At the core of SWS are solutions to the problem of SWS matchmaking, i.e., the problem of filtering and ranking a set of services with respect to a service query. Comparative evaluations of different approaches to this problem form the base for future progress in this area. Reliable evaluations require informed choices of evaluation measures and parameters. This paper establishes a solid foundation for such choices by providing a systematic discussion of the characteristics and behavior of various retrieval correctness measures in theory and through experimentation.

1 Introduction

In recent years, Semantic Web Services (SWS) research has emerged as an application of the ideas of the Semantic Web to the service oriented computing paradigm. The grand vision of SWS is to have a huge online library of component services available, which can be discovered and composed dynamically based upon their formal semantic annotations. One of the core problems in the area concerns SWS matchmaking, i.e., the problem of filtering and ranking a set of services with respect to a service query. A variety of competing approaches to this problem has been proposed [1]. However, the relative strengths and shortcomings of the different approaches are still largely unknown. For the future development of the area it is thus of crucial importance to establish sound and reliable evaluation methodologies.

Evaluations in the area typically follow the approach taken in the evaluation of Information Retrieval (IR) systems: As a basis for the evaluation a test collection is provided. This collection contains a number of service offers, a (smaller) number of service requests and relevance judgments. These relevance judgments are provided by human experts and specify for each offer-request pair how relevant the offer is for the request, i.e., whether or to which degree the offer is able to satisfy the request. Matchmakers are then evaluated by comparing their output rankings with the one induced by the relevance judgments. This is done via retrieval correctness measures which assign an output ranking a performance score based upon the available reference relevance judgments.