Abstract:
A very large number of Web sites expose their content via query interfaces (HTML forms), many of them offering the same type of products/services (e.g., flight tickets, car rental/purchasing, hotel room booking). Providing a uniform access to these sources is of practical importance as it facilitates users in searching and comparing services/products of multiple providers. The goal of my research is to construct integration systems that make the access to individual sources transparent to users. To achieve this goal a number of problems need to be addressed. The first is the problem of query interface extraction and understanding. Second, for a certain domain of discourse (e.g., real estate) a uniform query interface to the different data sources has to be constructed. Third, a query formulated on the integrated interface needs to be translated into queries against interfaces of specific sources. Last, returned data by individual sources need to be correctly extracted and the results ranked in descending order of certain desirability property (e.g. price).

In this talk, I will first give an overview of the general architecture of a deep Web integration system. I will then present two such systems that I have developed. One is VisQI (VISual Query interface Integration system). VisQI is capable of (1) extracting Web query interfaces, (2) classifying query interfaces into application domains, and (3) matching the elements of different interfaces. In this talk, I focus on the algorithms for (1) and (3). The other system is Yumi (www.yumi-meta.com), an integration system for local search engines for Geo-referenced objects. Typical queries submitted to local search engines include not only information about "what" a user is searching for (such as cuisine) but also "where", such as neighborhood. I will present three key algorithms implemented in Yumi: a neighborhood query processing algorithm, a business listing resolution algorithm and a ranking algorithm with active weights computation. Finally, I will briefly discuss future research directions. At the end of my talk, I will talk about my recent research endeavors into (1) Web data cleaning, (2) sentiment analysis, and (3) cyber-infrastructure for scientific research.

Bio:
Eduard Dragut is currently a Postdoctoral Research Associate at Purdue University, Discovery Park, Cyber Center. He completed his Ph.D. degree in Computer Science in July 2010 from the University of Illinois at Chicago. His main area of research is Web data management (e.g., extraction, representation, integration, etc.), with particular emphasis on the data from Web databases. Other projects he is actively pursuing include: Web data cleaning (e.g., Web-scale entity disambiguation, verifying the truthfulness of fact statements), sentiment analysis, and cyber-infrastructure for scientific research.