Abstract

As deep web grows at a very fast pace, there has been increased interest in techniques that help efficiently locate deep-web interfaces. However, due to the large volume of web resources and the dynamic nature of deep web, achieving wide coverage and high efficiency is a challenging issue. We propose a two-stage framework, for harvesting deep web interfaces. In the first stage of harvesting, performs site-based searching for center pages with the help of search engines, avoiding visiting a large number of pages. To achieve more accurate results for a focused crawl ranks websites to prioritize highly relevant ones for a given topic. In the second stage, it achieves fast in-site searching by excavating most relevant links with an
adaptive link-ranking.

References


Index Terms

Computer Science

Information Sciences

Keywords