INTISARI

PENGEMBANGAN XPATH WRAPPER DARI HASIL EKSTRAKSI AUTORM DAN DAG-MTM MENGGUNAKAN ALGORITME BUCAPP

Oleh

SIGIT DEWANTO
15/388499/PPA/04938

Ekstraksi data dari halaman web merupakan bidang yang cukup aktif diteliti hingga saat ini. Salah satu permasalahan dalam bidang ini adalah ekstraksi *data records* dari *list page*. AutoRM merupakan salah satu algoritme terkini untuk permasalahan tersebut. Untuk menjajarkan *data items* pada *data records* hasil ekstraksi AutoRM, digunakan suatu algoritme yang bernama DAG-MTM. Salah satu kelemahan dari AutoRM dan DAG-MTM adalah kedua algoritme tersebut tidak menghasilkan *wrapper*. Yang dimaksud dengan *wrapper* di sini adalah suatu prosedur atau aturan ekstraksi yang dapat digunakan untuk mengekstrak *data records* dari sekumpulan halaman web yang memiliki *template* yang sama. Jika *wrapper* dapat dihasilkan, maka proses ekstraksi menggunakan AutoRM dan DAG-MTM cukup dijalankan sekali pada halaman web contoh. Selanjutnya, proses ekstraksi pada halaman web lain yang memiliki *template* yang sama dengan halaman web contoh dapat dilakukan dengan menggunakan *wrapper*. Hal ini dilakukan karena proses ekstraksi menggunakan *wrapper* lebih efisien, baik dari segi waktu maupun penggunaan memori.

Pada penelitian ini, diusulkan sebuah algoritme baru yang dinamakan BUCAPP (*Bottom-Up, Class And Positional Predicate*) untuk menghasilkan *wrapper* dari hasil ekstraksi AutoRM dan DAG-MTM. *Wrapper* yang dihasilkan adalah *wrapper* berbasis XPath. Algoritme BUCAPP memanfaatkan atribut "class" dan juga posisi suatu *node* pada pohon DOM.

Hasil pengujian menunjukkan bahwa rata-rata *precision* ekstraksi menggunakan XPath *wrapper* yang dihasilkan dengan algoritme BUCAPP lebih rendah dibanding ekstraksi menggunakan AutoRM dan DAG-MTM. Sementara itu, rata-rata *recall* ekstraksi menggunakan XPath *wrapper* lebih tinggi dibanding AutoRM dan DAG-MTM. Selain itu, proses ekstraksi yang dilakukan dengan XPath *wrapper* juga lebih efisien dari segi waktu dan penggunaan memori.

Kata-kata kunci : ekstraksi data web, induksi *wrapper*, XPath.

ABSTRACT

XPATH WRAPPER DEVELOPMENT FROM AUTORM AND DAG-MTM'S EXTRACTION RESULT USING BUCAPP ALGORITHM

By

SIGIT DEWANTO
15/388499/PPA/04938

Web data extraction is a quite active research area. One of the problems in this area is data records extraction from list page. AutoRM is a state-of-the-art algorithm for this problem. To align data items on data records extracted by AutoRM, an algorithm named DAG-MTM is used. One of the limitations of AutoRM and DAG-MTM is those two algorithms don't generate wrapper. Wrapper here refers to extraction procedures or rules that can be used to extract data records from web pages having the same template. If wrapper can be generated, extraction using AutoRM and DAG-MTM only needs to be conducted once on the example web page. Extraction on other web pages with the same template as example page can be done by using wrapper. The main reason of using wrapper instead of AutoRM and DAG-MTM is efficiency. Wrapper is more efficient in term of time and memory usage.

On this research, a new algorithm named BUCAPP (Bottom-Up, Class And Positional Predicate) is proposed to induce wrapper from AutoRM and DAG-MTM's extraction result. The wrapper is based on XPath. BUCAPP algorithm exploits "class" attribute and the position of a node in the DOM tree.

Experimental results show that the average precision of extraction using XPath wrapper generated by BUCAPP algorithm is lower than AutoRM and DAG-MTM's. Meanwhile, the average recall of extraction using XPath wrapper is higher than AutoRM and DAG-MTM's. Moreover, the extraction using XPath wrapper is far more efficient in term of time and memory usage.

Keywords : web data extraction, wrapper induction, XPath.