

Adatletöltés szemantikus fákkal

Jánosi-Rancz Katalin Tünde, Lajos Árpád

Sapientia Erdélyi Magyar Tudományegyetem
tsuto@ms.sapientia.ro, lajos.arpas85@gmail.com

Az adatletöltés egy széles körben kutatott terület. Az adatletöltőket általában nehéz fenntartani, mert meg kell birkózzanak végtelen számú lehetséges adatstruktúrával, új adatforrások integrálásával, valamint a meglévő adatforrások szerkezeti változásaival.

A mi célunk az, hogy megkönnyítsük az adatletöltők létrehozásának munkaigényes folyamatát. Kifejlesztettünk egy gyors és pontos eszközt, *RK* névvel, amely szemantikus adatletöltést támogat fa struktúrájú adatforrásokból, mint HTML, XML, JSON, stb. Ez az eszköz immúnis a strukturális módosításokra és lehetővé teszi a szemantikus keresést a letöltött adatok alapján.

Az *RK* segítségével végzett kísérletek bebizonyították, hogy a megközelítésünk javítja az adatletöltést pontosság és sebesség szempontjából.

Hivatkozások

- [1] Chang, C.-H., Kaye, M., Girgis, M. R., and Shaalan, K. F. A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10),(2006), 1411-1427
- [2] V. Crescenzi, D. Qiu, and P. Merialdo. A Framework for Learning Web Wrappers from the Crowd. *Proceedings of the 22nd WWW Conference*, (2013), 261-272
- [3] X. Zheng, Y. Gu, and Y. Li. Data extraction from web pages based on structural-semantic entropy. In *Proceedings of the 21st WWW Conference*, (2012), 93-102
- [4] T.-L. Wong and W. Lam. Unsupervised Extraction of Popular Product Attributes from Web Sites. *8th Asia Information Retrieval Societies Conference*, (2012), 437-446