International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 5, Issue 3, Jun 2015, 23-30
TJPRC Pvt. Ltd.

AUTOMATIC EXTRACTION OF TOP-K LISTS FROM THE WEB PAGES

ASHISH DWIVEDI

(M.E.(Computer)) Lokmanya Tilak College of Engineering & Technology, LTCOE, NAVI Mumbai, India

ABSTRACT

The World Wide Web is currently the largest source of information. The information on the Web is structured (such as table) and unstructured or semi-structured form. Unfortunately, in most cases, context is expressed in structured text that machines can not interpret. Therefore this project is concerned with information extraction from top-k Web pages (rich and valuable source of information), which are web pages that describe top k instances of a topic which is of general interest.

Examples include “20 Most Influential Scientists Alive Today”, “the 10 hits of 2010 you don’t want to miss” etc. Compared to other structured information on the web (including web tables), information in top-k lists is larger and richer, of higher quality, generally more interesting. Therefore top-k list are highly valuable. For application such as search or fact answering, it can help enrich open-domain knowledge bases for their support. The paper concentrate on an efficient method, called Tag Path Clustering that extracts top-k list from web pages with high performance.

KEYWORDS: Web Lists, Web Information Extraction, Top-k Lists, Lists Extraction, Web Mining