UC San Diego
Technical Reports

Title
WikiAnalytics: Disambiguation of Keyword Search Results on Highly Heterogeneous Structured Data

Permalink
https://escholarship.org/uc/item/82f168vg

Authors
Balmin, Andrey
Curtmola, Emiran

Publication Date
2010-05-18

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

WikiAnalytics: Disambiguation of Keyword Search Results on Highly Heterogeneous Structured Data

ANDREY BALMIN
IBM Almaden Research Center
and
EMIRAN CURTMOLA
University of California, San Diego

Wikipedia infoboxes is an example of a seemingly structured, yet extraordinarily heterogenous dataset, where any given record has only a tiny fraction of all possible fields. Such data cannot be queried using traditional means without a massive a priori integration effort, since even for a simple request the result values span many record types and fields. On the other hand, the solutions based on keyword search are too imprecise to exactly capture the user's intent.

To address these limitations, we propose WikiAnalytics system that utilizes a novel search paradigm in order to derive tables of precise and complete results from Wikipedia infobox records. The user starts with a keyword search query that finds a superset of the result records, and then browses clusters of records deciding which are and are not relevant. WikiAnalytics uses three categories of clustering features based on record types, fields, and values that matched the query keywords, respectively. Since the system cannot predict which combination of features will be important to the user, it efficiently generates all possible clusters of records by all sets of features. We utilize a novel data structure, universal navigational lattice (UNL), that compactly encodes all possible clusters. WikiAnalytics provides a dynamic and intuitive interface that lets the user explore the UNL and construct homogeneous structured tables, which can be further queried and aggregated using the conventional tools.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: On-line Information Services; H.2.m [**Database Management**]: Miscellaneous

General Terms: Algorithms, Design, Performance

Additional Key Words and Phrases: ad-hoc querying, heterogeneous structured data, search web data, keyword search

Request

To obtain a copy of this UCSD technical report please send an email or a letter request to:

Emiran Curtmola ecurtmola@cs.ucsd.edu
University of California, San Diego - CSE
9500 Gilman Dr.
La Jolla, California 92093, USA

Author's addresses: Andrey Balmin, IBM Almaden Research Center (ARC), San Jose, CA 95120; email: abalmin@us.ibm.com; Emiran Curtmola, Department of Computer Science at University of California San Diego, La Jolla, CA 92093; email: ecurtmola@cs.ucsd.edu;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided copies are not made or distributed for profit or commercial advantage and copies bear this notice and the full citation on the first page. To copy otherwise, republish, post on servers or redistribute to lists, requires prior specific permission and/or a fee.

Technical Report CS2010-xxxx, UC San Diego, May 2010.