WINACS (Web-based Information Network Analysis for Computer Science) aims to reconcile the unstructured nature of the Web with the neat, semi-structured schemas of the database paradigm.

Taking computer science as a dedicated domain, WINACS first discovers related Web entity structures, and then constructs a heterogeneous computer science information network in order to rank, cluster and analyze this network and support intelligent and analytical queries.
General List Extraction
We automatically discover and extract lists on the Web using HyLiEn, an unsupervised, universal, hybrid approach. We use visual alignment of boxes employed by modern web browsers to generate list candidates. We then employ the DOM-structure to prune candidates that are not structurally aligned.

Fumarola et al, WWW’11

Entity-Page Discovery
The entity discovery module uses found lists to create parallel paths to discover the Web pages of similarly typed entities. For example, given the homepage of a CS department, and the Web page of a faculty member, we find all faculty members in the department.

Weninger et al, WWW’11

Ranking & Clustering
WinaCS utilizes NetClus, a recently proposed algorithm, to create a heterogeneous information network (HIN) from the discovered Web entities, and generate high-quality net-clusters. An HIN is composed of multiple types of objects, reflecting the structure found in many real-world networks. In addition to ranking and clustering objects, we can also discover hidden knowledge such as the roles of entities in information propagation, implicit relationships between entities, the hierarchical structure of the network organization, and useful summary information.

Sun et al, KDD’09; Wang et al, KDD’10

Record Linkage
The entity-page discovery process yields a set of Web pages representing entities of the same type. We then use the paths of anchor texts to link each entity page with a new or existing record in the offline database. Third-party structured data (e.g. DBLP) is not required.

Weninger et al, CIKM’10

Query Processing
WinaCS performs random walks on the information network to rank the probabilities of diffusion among the entities within a cluster. The system is able to quickly retrieve and rank entities in response to a query, and allows the user to intelligently navigate the search results.

Weninger et al, CIKM’10

Contact Us:
Tim Weninger - weninge1@illinois.edu
Jiawei Han - hanj@illinois.edu

Use WinaCS Today:
http://winacs.cs.illinois.edu