WINACS: Construction and Analysis of Web-Based Computer Science Information Networks

Tim Weninger, Marina Danilevsky, *Fabio Fumarola, Joshua Hailpern, Jiawei Han, Thomas J. Johnston, †Surya Kallumadi, Hyungsul Kim, Zhijin Li, David McCloskey, Yizhou Sun, ‡Nathan E. TeGrotenhuis, Chi Wang, Xiao Yu
University of Illinois at Urbana-Champaign, *Universita' degli Studi di Bari
†Kansas State University, ‡Whitworth University
{weninge1, danilev1, jhailpe2, hanj, johnst26, hkim21, zli12, mcclosk1, sun22, chiwang1, xiaoyu1}@illinois.edu, *ffumarola@di.uniba.it, †surya@ksu.edu, ‡ntegrotenhuis12@whitworth.edu

ABSTRACT

WINACS (**W**eb-based **I**nformation **N**etwork **A**nalysis for **C**omputer **S**cience) is a project that incorporates many recent, exciting developments in data sciences to construct a Web-based computer science information network and to discover, retrieve, rank, cluster, and analyze such an information network. With the rapid development of the Web, huge amounts of information are available in the form of Web documents, structures, and links. It has been a dream of the database and Web communities to harvest such information and 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.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

WINACS, Web mining, information networks

1. INTRODUCTION

Exciting new developments in Web mining and information network analysis show promising results in the access and synthesis of heterogeneous information. To date, there has only been limited work at the intersection of information networks and the Web-at-large. Specifically, information network analysis requires data in a structured format, and Web mining researchers have been unable to provide Web information in such a format.

Despite the growing demand for such an application, there exist only a few tools which perform Web mining for entity discovery or information network analysis on the Web. Furthermore, to our knowledge, there do not exist any applications at the intersection of these two technologies. The goal of the WINACS system is to integrate Web mining with information network analysis as well as several other state-of-the-art technologies to facilitate the ranking, clustering, and retrieval of various entities from the Web.

SIGMOD'11, June 12–16, 2011, Athens, Greece.
ACM 978-1-4503-0661-4/11/06.

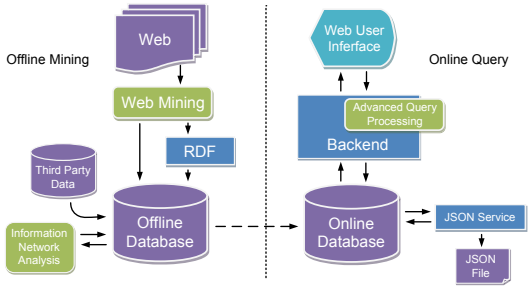

Figure 1: System Architecture of WINACS.

WINACS takes computer science as a domain of study. It discovers Web entity structures, and then constructs a heterogeneous computer science information network by integrating the contents of the DBLP database with the entities found on the Web. It also supports information browsing, query answering and mining-based search.

The system architecture of WINACS depicted in Figure 1 consists of two interconnected parts: (1) offine mining, and (2) online query retrieval. To maximize overall system performance, we employ different architectures for each part: extended operational data storage for offline mining, and model-view-controller with a lightweight service oriented architecture for online query retrieval.

The offline mining part is responsible for the gathering and general analysis of information. Multiple data sources are integrated and preprocessed for future query and mining analysis. The Web mining component gathers and preprocesses data from the Web, while the information network analysis component clusters and ranks the information in the database. The model consists of an offline database running Microsoft's SQLServer data storage engine with a dynamic schema created and maintained by a rich set of algorithms within the Web mining and information network analysis modules. When data in RDF format is found during the Web mining process it is also stored in the offline database for analysis. Although the Web mining component does use some information from the offline database, the main flow of information is from the Web to the offline database via the Web mining component as indicated by the arrows in Figure 1.

At designated update periods (*e.g.*, once per day) the offline mining and online query parts are synchronized so that freshly mined

information from the Web can be queried by users. In its current form, WINACS does not analyze user interaction information (*e.g.*, query logs, click-through data); however, this capability is planned as a matter of future work.

The online query part consists of a model-view-controller architecture. The view component is a Web application that can be accessed by any Web browser. Queries and commands input to the Web application are sent via HTTP to a backend Web server controller. Depending on the nature of the request the Web server may directly access data in the Online Database or it may require more advanced query processing (see Section 2.3).

2. MAJOR FUNCTIONAL MODULES

As we discussed in Section 1, several recent developments in data sciences have facilitated the development of WINACS. We divide the new algorithms into three categories based on the nature of the methods: (1) *Web structure mining*, (2) *information network analysis*, and (3) *advanced query processing*.

2.1 Web Structure Mining

The Web mining process consists of four consecutive steps, which are described in order here.

2.1.1 List Finding

The first module concerns the discovery and extraction of *general* lists on the Web. We find that current approaches to this problem are limited by burdensome assumptions and therefore are not universal enough for the Web-at-large [8]. Our approach, called HyLiEn is an unsupervised, Hybrid approach for automatic List discovery and Extraction on the Web. Our extraction method employs general assumptions about the visual rendering of lists, and the structural representation of items contained in them. The approach uses: (1) the visual alignment of boxes within the two dimensional visual box model used by modern Web browsers to generate list candidates, and (2) the DOM-structure of aligned boxes to prune candidates which are not structurally aligned [2].

2.1.2 Entity Discovery

The entity discovery module uses the lists found in the previous module to create *parallel paths* through Web sites in order to find the Web pages of similarly typed entities. This module operates given a Web site and an entity-page and returns all of the entity-pages of the same type as the example entity-page. For example, given the homepage of a computer science department and the Web page of a faculty member, the entity discovery algorithm discovers all faculty members in the department. Currently, entity discovery is approx. 90% accurate on faculty members and 100% accurate on research groups, courses, and other entities [7].

2.1.3 Record Linkage

The result of the entity discovery model is a collection of Web pages which represent entities of the same type. At this point, the system does not know the name or any other identification of the entity. To resolve this, we use the paths of anchor texts, a byproduct of the entity discovery module, to link each entity-page with a new or existing record in the Offline Database. Third party structured data, such as DBLP or a phonebook, can be helpful here, but is not required. Our record linkage method achieves 98% precision at 100% recall in this domain [9].

2.1.4 Schema Discovery

The final step of the Web mining controller is to attribute properties to the appropriate entities thereby automatically creating schemata. For example, professor-entities typically have universities, addresses, phone numbers, students, grants, and other academic information attributed to them. We discover these attributes automatically and use them to dynamically extend the schema for each entity type.

As a result of these four Web mining steps, the Offline Database is populated with entities and dynamic schemata. In this approach, we do not actively seek for information to fill any predefined schema, instead we let the Web tell us what information is important and how it relates to appropriate entities.

2.2 Information Network Analysis

Information network analysis investigates effective discovery of patterns and knowledge from large-scale networks that consist of interconnected components [3]. The data collected by the Web mining part is an ideal candidate for such analysis. The major themes in our study include: (1) ranking-based clustering on different types of entities (*e.g.*, faculty members, students, courses) in heterogeneous information networks [5]; (2) hierarchical network structure analysis for OLAP, multidimensional text database analysis [4], and ranking promotion [10]; (3) query-based information network extraction and analysis; and (4) link-based object resolution and disambiguation for bibliographic networks [11].

2.2.1 Ranking-based Clustering

A heterogeneous information network is an information network composed of multiple types of objects. Clustering on such a network may lead to better understanding of both hidden structures of the network and the individual role played by every object in each cluster. Real-world networks consist of many types of entities, and the interactions among multi-typed objects play a key role in disclosing the rich semantics that a network carries. WINACS uses a new algorithm, NetClus, that utilizes links across multi-typed entities to generate high-quality net-clusters. It has been shown that NetClus generates informative clusters, and good rankings and cluster membership information for each entity in each net-cluster [5].

2.2.2 Network Structure Analysis

In addition to ranking and clustering, we can discover hidden knowledge from the information network, such as the roles of the entities in information propagation, the relationships between entities, the hierarchical structure of the network organization and summary information. Specific tasks in the computer science domain include: (1) advisor-advisee relationship mining where the task is to find a subnetwork with distinguished relationships [6]; and (2) hierarchical structure mining which allows for the the network to be analyzed at multiple granularities.

With the discovered relationships, our system can support the search and ranking for linked entities according to both content relevance and structural coherence. For example, if a user wishes to find database communities formed by the advising relationship, the system will use knowledge of these relationships in order to select the appropriate entities for retrieval. On the other hand, if a user gives a topic, the system can find the most influential research papers in the field development. Generally, automatic role discovery helps the system to create clearer structure from undistinguished links between entities using only a small amount of human effort.

2.2.3 Entity Resolution and Disambiguation

Different people or objects may share identical names in the real world, which causes confusion in many applications. It is a nontrivial task to distinguish those objects, especially when there is only very limited information associated with each of them. Network-

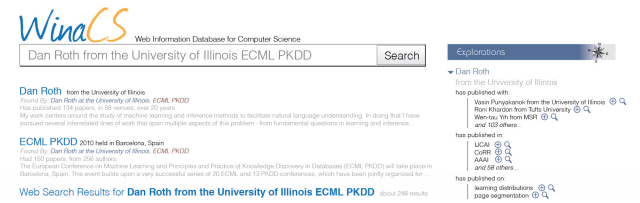

Figure 2: Illustration of WINACS on query of "Dan Roth" returning an *infobox* snippet.

based disambiguation algorithms such as DISTINCT can distinguish entities in these cases [11]. This process combines two complementary measures for relational similarity: (1) set resemblance of neighbor tuples and (2) random walk probability, in order to analyze subtle linkages effectively. We are also able to leverage information from list finding to help this process.

2.3 Advanced Query Processing

Because of the data representation within WINACS, queries including both structured dimension information and unstructured text information are able to be supported. Among the various possibilities two novel search functions found in WINACS are: (1) integrated search, and (2) promotion queries.

2.3.1 Infobox-styled Integrated Search

The integrated search functionality of WINACS allows the user to find relationships between two or more entities. For example, the query: "Jim Gray, Data Cube" will return information on how Jim Gray is related to data cube topic. Results for this query include relevant publications by Jim Gray, coauthors of Jim Gray on data cubes, institutions performing data cube research, top-k most similar pairs, and so on. Integrated search provides location and temporal search capabilities as well. If a user wishes to find people and/or universities performing formal language research in the midwest, they may issue the query: "formal methods, midwest".

2.3.2 Promotion Query

Promotion analysis is the processes by which WINACS finds properties of a particular entity which rank relatively high among the complete set of entities. For example, the query: "promote: Kansas State" will find which attributes of Kansas State University rank highly during various periods of time. Results for this query include topics that Kansas State excels at, such as programming languages and formal methods, as well as highly ranked professors, courses and papers at Kansas State during highly ranked years.

3. ABOUT THE DEMONSTRATION

The WINACS system integrates the Web mining methods presented in Section 2.1 and the Information Network Analysis functions presented in Section 2.2. The methods embedded in the system are novel, practical, and derived from recent and ongoing research. The data set used in this demo was collected from various computer science Web sites; this is to demonstrate its effectiveness on a large, varied, yet familiar domain.

WINACS provides an easy-to-use Web interface. Figure 2 is a preliminary screen shot of an example result. Notice that the result in Figure 2 is not a collection of names, terms, or 10 blue links, but rather an automatically generated *infobox* snippet of the retrieved entity. Because we cannot guarantee the accuracy of the information in all cases, we allow users to navigate to the evidence used to compose the business card results.

This system stands apart from existing Web-based computer science search engines (*e.g.*, DBLife [1]) because our Web structure mining, information network analysis and advanced query processing modules are completely automatic. That is, we have no use for hand-crafted information extraction rules, or preset domains; instead, we rely on the Web to tell the system where the entities are located and what their properties are.

Although ArNetMiner, iNextCube and other recent demo systems similarly cluster and organize computer science information, we would like to emphasize that we use data available on the general Web, and therefore do not solely rely on DBLP (nor any other structured database) for our data. This information is gathered and analyzed by the methods introduced above. We believe that, instead of relying on explicit or handcrafted data, by moving towards a Web-based information network database system like WINACS we can achieve a more informative, richer set of query results. This demonstration will therefore promote further research into new, challenging issues at the intersection of Web mining and information network analysis.

4. ACKNOWLEDGMENTS

This work is funded by an NDSEG Fellowship to the first author. Other support comes from NSF IIS-09-05215, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA).

5. REFERENCES

[1] P. DeRose, W. Shen, F. Chen, Y. Lee, D. Burdick, A. Doan, and R. Ramakrishnan. Dblife: A community information management platform for the database research community. In *CIDR*, 2007.

[2] F. Fumarola, T. Weninger, R. Barber, D. Malerba, and J. Han. Extracting general lists from web documents: A hybrid approach. In *IEA/AIE*, March 2011.

[3] J. Han, Y. Sun, X. Yan, and P. S. Yu. Mining knowledge from databases: an information network analysis approach. In *SIGMOD*. ACM, 2010. (Tutorial).

[4] C. X. Lin, J. Han, F. Zhu, and B. Zhao. Text cube: Computing ir measures for multidimensional text database analysis. In *ICDM*. IEEE, 2008.

[5] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *SIGKDD*. ACM, 2009.

[6] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *SIGKDD*. ACM, 2010.

[7] T. Weninger, F. Fumarola, R. Barber, C. X. Lin, J. Han, and D. Malerba. Growing parallel paths for entity-page discovery. In *WWW*, April 2011.

[8] T. Weninger, F. Fumarola, R. Barber, and J. H. D. Malerba. Unexpected results in automatic list extraction on the web. *SIGKDD Explorations*, 12(2), 2010.

[9] T. Weninger, F. Fumarola, J. Han, and D. Malerba. Mapping web pages to database records via link paths. In *CIKM*, 2010.

[10] T. Wu, D. Xin, Q. Mei, and J. Han. Promotion analysis in multi-dimensional space. *PVLDB*, 2(1), 2009.

[11] X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In *ICDE*. IEEE, 2007.