METEOR: Metadata and Instance Extraction from Object Referral Lists on the Web

Hasan Davulcu, Srinivas Vadrevu, Saravanakumar Nagarajan, Fatih Gelgi
Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287-5406, USA
{hdavulcu,svadrevu,nrsaravana,fagelgi}@asu.edu

ABSTRACT

The Web has established itself as the largest public data repository ever available. Even though the vast majority of information on the Web is formatted to be easily readable by the human eye, "meaningful information" is still largely inaccessible for the computer applications. In this paper we present the METEOR system which utilizes various presentation and linkage regularities from referral lists of various sorts to automatically separate and extract metadata and instance information. Experimental results for the $university$ domain with 12 computer science department Web sites, comprising 361 individual faculty and course home pages indicate that the performance of the metadata and instance extraction averages 85%, 88% F-measure respectively. METEOR achieves this performance without any domain specific engineering requirement.

Categories and Subject Descriptors: H.4.m [Information Systems]: Miscellaneous; I.2.6 [Artificial Intelligence]: Learning–*Knowledge Acquisition*

General Terms: Algorithms, Performance, Experimentation

Keywords: Web, Semantic, Metadata, Object, Instance, Extraction

1. INTRODUCTION

Scalable information retrieval [1] based search engine technologies have achieved wide spread adoption and commercial success towards enabling access to the Web. However, since they are based on an unstructured representation of the Web documents their performance in making sense of the available information is also limited.

Thanks to the HTML format, unlike plain text documents, Web pages organize and present their content within nested hierarchies of HTML structures. In this paper we present an algorithm that can detect various HTML regularities [3] and utilize them to structure the Web page content itself into hierarchical $group$ structures which contains blocks of highly regularly presented $instances$.

Furthermore, many Web pages present their information in the form of labeled lists and tables of various sorts. Consider the first page in the example shown in Figure 1 that lists the faculty instances in a computer science department. Each of these faculty instance links to an individual faculty home page with detailed information. We denote these type of groups as *object referral lists* (ORL). Examples of ORLs are faculty listings in the universities, job listings in company sites, course listings in online schools, hotel and hospital listings in directories etc. Object referral lists follow a highly regular linkage pattern; first they list their instances under an informative label such as $jobs$, $faculty$, $hotels$ etc. and then each instance links to an individual detailed object page. The individual object page presents the detailed attributes of an object as shown in the second page of Figure 1. Sometimes the individual object page may present its attribute information by linking to an object attribute page as shown in the third page of Figure 1. In this paper we present algorithms that can interpret any given ORL to navigate to its instances and extract their attributes and values.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. OVERVIEW OF THE APPROACH

The METEOR system utilizes Semantic Partitioner that infers hierarchical relationships among the leaf nodes of the DOM (Document Object Model) tree of a Web page, where all the document content is stored. Semantic partitioner achieves this through a sequence of two operations: hierarchical grouping and promotion. The hierarchical grouping is based on a regular pattern mining algorithm which yields a hierarchy of groups (G) and their instances (I). After hierarchical grouping, all the content of the Web page is still at the leaf nodes of the hierarchical group tree and hence promotion of some of the leaf nodes is necessary in order to organize them into a semantic hierarchy. The promotion algorithm identifies those leaf nodes that should be promoted above their siblings.

Next the METEOR system interprets these semantic structures by utilizing linkage regularities that exist within the context of an ORL in order to separate and extract their metadata and instances. The extracted metadata and object instances are represented as F-logic [2] facts. The interpretation proceeds for the group structures in the individual object page by providing the appropriate context. Whenever, the object instance pages present their attributes using a link group, each group structure within the object attribute pages is interpreted and corresponding value types are extracted.

3. AN ILLUSTRATIVE EXAMPLE

We illustrate the inner-workings of the Hierarchical Partitioning Algorithm using the sequence $aaabcdefffghijhikhijhilhijhikmnnonmpnnqrq$ from the individual object page in Figure 1 and explain the process. The $HiearchicalGrouping$ algorithm attempts to standardize the path sequence as a regular expression which can be used to parse the original sequence into hierarchical group structures, presented in Figure 2. It utilizes the $maximize$ subroutine that

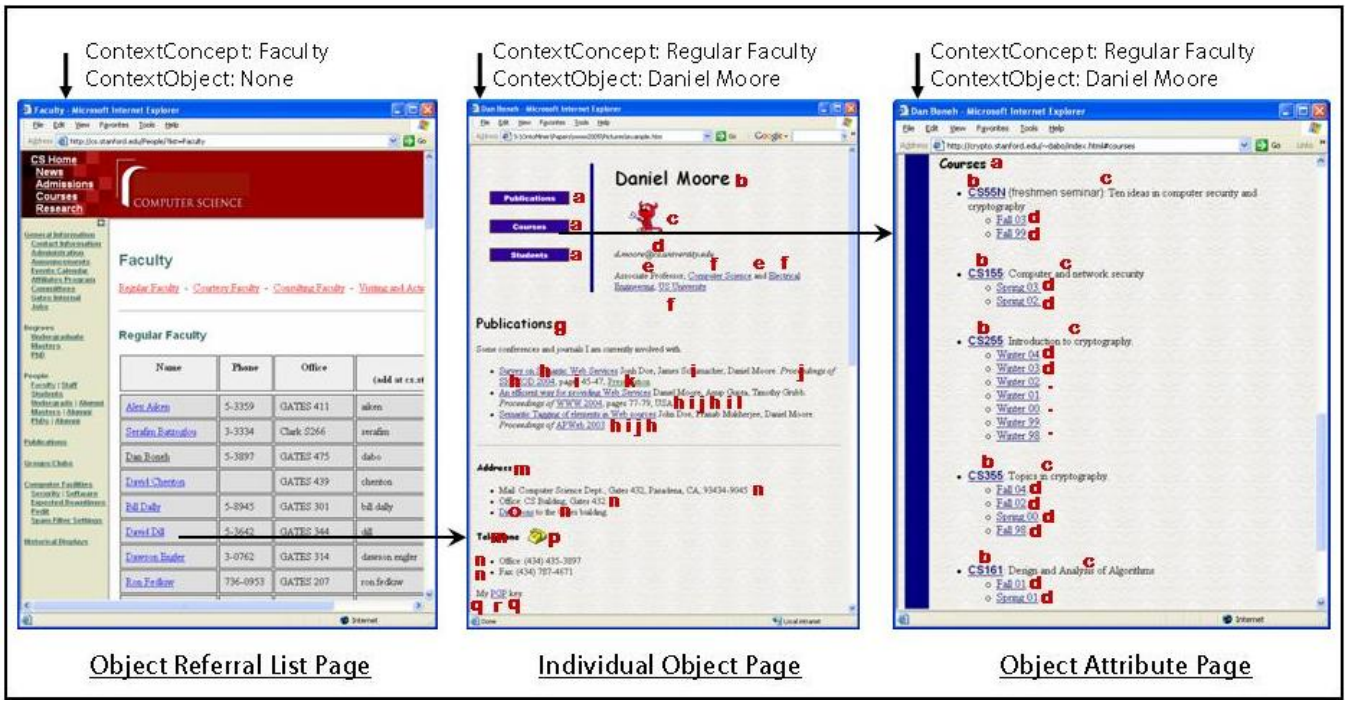

Figure 1: An example of the *object referral list* page, that links individual object pages. The figure shows an individual object page and one of its object attribute pages. The labels in individual object page and the object attribute page are marked with corresponding path identifier symbols.

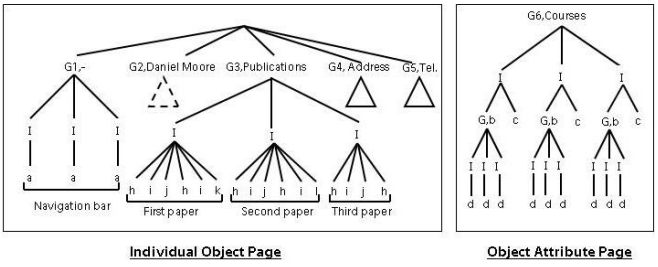

Figure 2: The group structures for the individual object page and the object attribute page in Figure 1. The groups $G_1, G_2, G_3, G_4, G_5, G_6$ correspond to the navigation bar, affiliations, publications, address, the telephone information and the courses respectively.

finds the maximum number of consecutive atoms that can be appended to the current atom. For example, initially the subpattern $(a)*$ is computed using the $maximize$ subroutine by recursively invoking the $HieararchichalGrouping$ algorithm. This pattern corresponds to the regularly presented structure, the navigation bar of the individual object page shown as G_1 in Figure 2. Then the algorithm continues to find the subpatterns $bcd(e(f)*)*$, $g(hij(hik|hil|hik))*$ that corresponds to the affiliations and the publications presented in the page shown as G_2 and G_3, and appends to the previous subpattern $(a)*$. The algorithm eventually generates nested group structures presented in Figure 2 from the final pattern. The complexity of the algorithm is $O(n^3)$ where n is the length of the input string.

For example, consider the ORL group structure G in Figure 1 that lists instances of the *Regular Faculty* concept. As the group structure presents the members of the 'Faculty' concept and its value types are found by the Hierarchical Grouping algorithm, the following F-logic statements are extracted from the ORL group G.
'Regular Faculty' : concept.
'Alex Aiken' : 'Regular Faculty'.
'Daniel Moore' : 'Regular Faculty'. ...
'Regular Faculty'['Name' $\twoheadrightarrow$ {'Alex Aiken',...}].
'Alex Aiken'['Name' $\rightarrow\!\!\!\!\!\rightarrow$ 'Alex Aiken'].
'Regular Faculty'['Phone' $\twoheadrightarrow$ {'5-3359', '3-3334', ...}].
'Alex Aiken'['Phone' $\rightarrow\!\!\!\!\!\rightarrow$ '5-3359']. ...

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the METEOR system that can automatically separate and extract metadata and instance information from *object referral lists*. The experimental results indicate that the METEOR system was able to extract the metadata and the instance information with high accuracy. In our future work, we propose to develop automated algorithms for finding all the ORL structures within any Web site.

5. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] M. Kifer, G. Lausen, and J. Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42(4):741–843, July 1995.

[3] G. Yang, W. Tan, S. Mukherjee, I.V.Ramakrishnan, and H. Davulcu. On the power of semantic partitioning of web documents. In *Workshop on Information Integration on the Web*, Acapulco, Mexico, 2003.