JOÃO MATEUS DE FREITAS VENEROSO

WEB DATA EXTRACTION IN SEMI-STRUCTURED DOCUMENTS

Dissertation proposal presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Berthier Ribeiro de Araújo Neto

Belo Horizonte

November 2017

Contents

Chapter 1

Introduction

Web data extraction is the task of automatically extracting structured information from unstructured or semi-structured web documents. It is a subset of the broader field of Information Extraction, and thus it faces many of the same challenges.

Structured information such as that found in a well organized relational database must conform to an underlying data model, namely an abstract model that formalizes the entities and relationships in a given application domain. Unstructured information however is not organized according to a logical model and has useful information often permeated by large chunks of irrelevant data.

Tipically, Information Extraction tasks consist of mapping unstructured or poorly structured information to a semantically well defined structure. The input is usually composed of a set of documents that describe a group of entities in a similar manner while the Information Extraction task deals with identifying those entities and organizing them according to a template.

As an example, consider a collection of novels and the task of identifying the name of the main character in each novel. For this task, the model must first identify proper nouns and then understand the text sufficiently to allow inference on the relative importance of each noun.

To achieve such a goal it is often useful to employ methods developed in the disciplines of Information Retrieval and Natural Language Processing. The former has achieved a great deal of success in the task of classifying documents according to statistical properties and the latter led to huge improvements in modelling human language. Many times, the various methods employed in these disciplines lead to different approaches in the field of Information Extraction.

In the scope of this work, we are interested in the semi-structured data usually found in HTML web documents. Web documents most often lie in between the

structured-unstructured data paradigm, meaning that they take a rather relaxed approach in regard to formal structure. Hierarchy, element disposition, class names, and other features related to document structure and indirectly associated with the data itself are valuable information in the task of identifying entities and determining relationships. So much that many times they the major source of information for classification purposes such as when extracting information from standardized tables. However, far from a structured database, web documents usually provide very limited structure to otherwise unstructured data such as that found in free text.

Take for example the staff page for the intelligent robotics laboratory of Osaka University shown in figure 1. Say we want to extract the name, position, email and picture of all members. It is easy to see there is some sort of structure to the information we want to extract from this particular website, however, parts of the information are missing, repeated or disposed differently for each particular member. If we want to go further it may be necessary to only extract information from full members, a task that may pose a harder challenge. We may use grouping similarity combined with textual information in order to properly identify the desired entities. In many cases the HTML element relative position or CSS class name is sufficient to identify occurrences of the same entity. If we do a good enough job at this first task we may extrapolate our model to extract information from similar web pages. The harder challenge lies in building a more general approach to collect data with a similar underlying structure from many different types of web pages.

Our current focused research interest regarding web data extraction is collecting computer science researcher information from university websites. We need this information in order to compare the reputations of national and international research groups in the area of computer science with the academic reputation ranking algorithm pscore.

Pscore is a recent author level metric that attempts to measure academic reputation by calculating Markov chain static distributions. It showed promising results in ranking publication venues and individual authors in the area of Computer Science using only information publically available in the DBLP database. However the DBLP database has sparse information about author affiliation and it is often outdated. To remedy this problem in the past few years researchers from UFMG have been laboriously collecting this data manually, but this process is tedious and error prone. Currently, only about 1% of the DBLP authors have associated affiliation information, being most of them from USA and Brazil. Not nearly enough to allow broad international research group comparison.

Up to now, our goal has been to build an automatic information extraction system

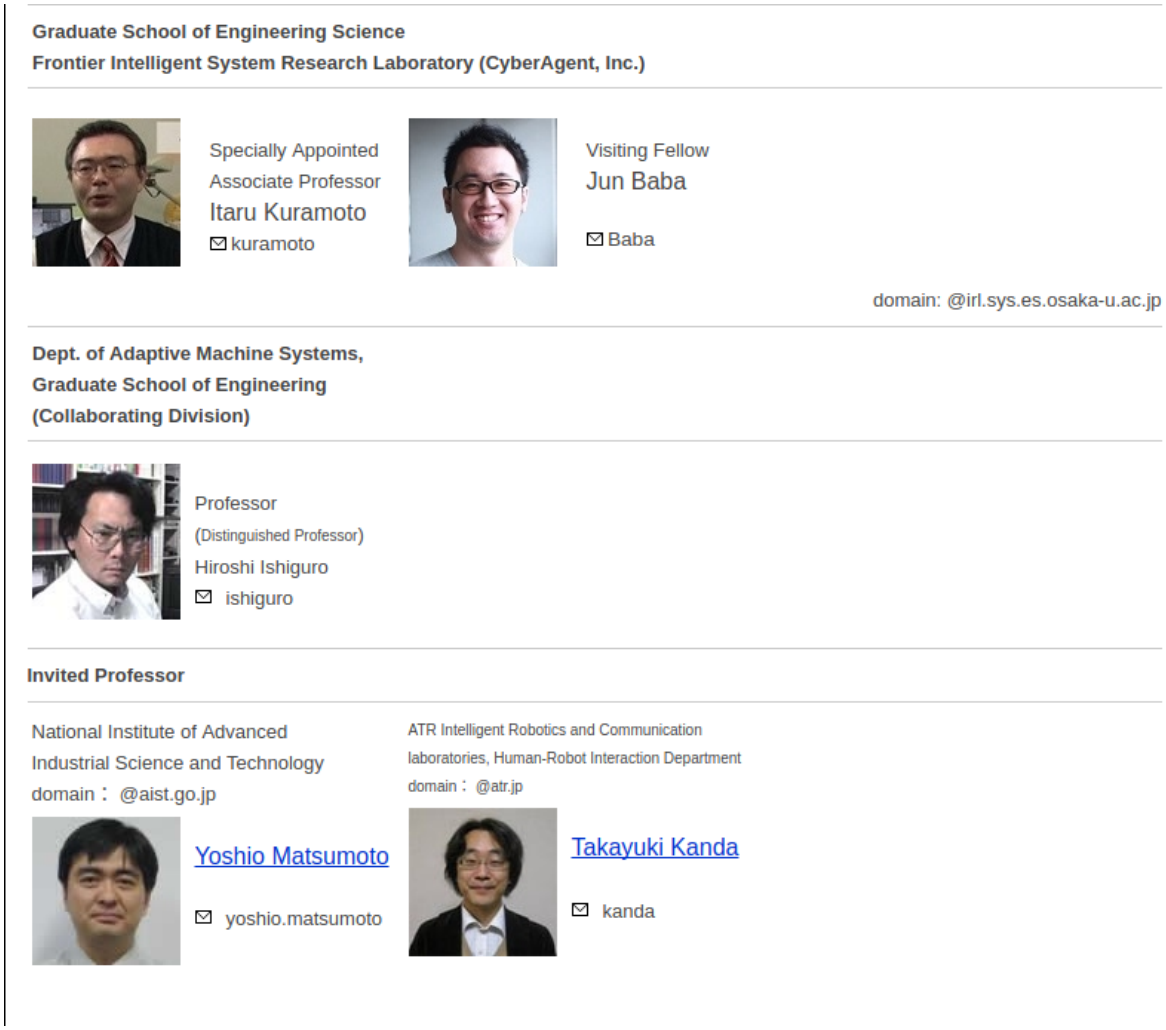

Figure 1.1. Example of semi structured information

for collecting author affiliation information from university websites. It has already achieved significant progress which shall be described further in the text. However, the ideas developed in this concrete case will be further improved in order to construct a more general approach to the broader Information Extraction task.

The research here described hopes to contribute by proposing a novel approach to the main Information Extraction task, making the computer science affiliation database available for further research and testing the quality of pscore results regarding computer science research group comparison.

Chapter 2

Related Work

Related work here.

Chapter 3

Methodology

Methodology here.

Chapter 4

Schedule

Schedule here.

Bibliography

Bichsel, M. and Pentland, A. P. (1992). A simple algorithm for shape from shading. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459--465.

Dror, R. O., Leung, T. K., Adelson, E. H., and Willsky, A. S. (2001). Statistics of real-world illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guisser, L., Payrissat, R., and Castan, S. (1992). A new 3-D surface measurement system using a structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 784--786.

Horn, B. K. P. (1986). *Robot Vision*. McGraw-Hill Book Company, Cambridge, Massachusetts. MIT Electrical Engineering and Computer Science Series.

Hougen, D. R. and Ahuja, N. (1993). Estimation of the light source distribution and its use in integrated shape recovery from stereo and shading. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 148--155.

Samaras, D. and Metaxas, D. (1999). Coupled lighting direction and shape estimation from single images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 868--874.

Sato, I., Sato, Y., and Ikeuchi, K. (1999a). Illumination distribution from brightness in shadows: Adaptive estimation of illumination distribution with unknown reflectance properties in shadow regions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 875--882.

Sato, I., Sato, Y., and Ikeuchi, K. (1999b). Illumination distribution from shadows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 306--312.

Sato, I., Sato, Y., and Ikeuchi, K. (2001). Stability issues in recovering illumination distribution from brightness in shadows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 400--407.

Shashua, A. (1997). On photometric issues in 3D visual recognition from a single 2D image. *International Journal of Computer Vision (IJCV)*, 21(1/2):99--122.