Search Computing Meets Data Extraction*

Tim Furche, Giorgio Orsi
Oxford University, Department of Computer Science, Wolfson Building, Parks Road, Oxford OX1 3QD
firstname.lastname@cs.ox.ac.uk

Alessandro Bozzon, Chiara Pasini, Luca Tettamanti, Salvatore Vadacca
Politecnico di Milano,
Via Ponzio 34/5, 20133 Milano, Italy
firstname.lastname@elet.polimi.it

ABSTRACT

Thanks to the Web, access to an increasing wealth and variety of information has become near instantaneous. To make informed decisions, however, we often need to access data from many different sources and integrate different types of information. Manually collecting data from scores of web sites and combining that data remains a daunting task.

The ERC projects SeCo *(Search Computing)* and DIADEM *(Domain-centric Intelligent Automated Data Extraction Methodology)* address two aspects of this problem: SeCo supports complex search processes drawing on data from multiple domains with a user interface capable of refining and exploring the search results. DIADEM aims to automatically extract structured data from a domain's websites.

In this paper, we outline a first approach for integrating SeCo and DIADEM. We discuss how to use the DIADEM methodology to automatically turn nearly any website from a given domain into a SeCo search service. We describe how such services can be registered and exploited by the SeCo framework in combination with services from other domains (and possibly developed with other methodologies).

1. INTRODUCTION

Recent years witnessed a paradigmatic shift in the way people deal with information. The Web provides cheap and ubiquitous access to an increasing wealth and variety of data. Yet, making informed decisions, which often require complex and articulated information retrieval tasks involving access to information from many different sources, remains a daunting task. Queries such as "Retrieve jobs as Java Developer in the Silicon Valley, nearby affordable fully-furnished flats, and close to good schools" are, unfortunately, not addressed by current search engines. From a vast list of potential sources, it is left to the user to manually *extract* and *integrate* the relevant data.

The **Search Computing (SeCo)** project [1] aims at building concepts, algorithms, tools, and technologies to support complex Web queries, through a new paradigm based on combining data extraction from distinct sources and data integration by means of specialized integration engines. Web data is typically published in two ways: as structured (and possibly linked) data accessible trough Web APIs (e.g. SPARQL, YQL, etc.), and as unstructured resources (i.e. Web pages), possibly accessible only through user-interaction such as form filling or link navigation.

Unstructured data is typically accessible to general-purpose search engines, which exploits traditional information retrieval techniques. To enable the consumption of such data by automated processes, data accessible to humans through existing Web interfaces needs to be transformed into structured information: therefore, there is the need for data extraction tools (e.g. screen scrapers); unfortunately, the interactive nature of modern Web interfaces poses a big challenge, as the dynamic nature of these user interfaces, driven by client and server-side scripting, creates challenges for automated processes to access this information.

The **DIADEM**[1] (Domain-centric Intelligent Automated Data Extraction Methodology) project aims at developing domain-specific data extraction systems that take as input a URL of a Web site in a particular application domain, automatically explore the Web site, and deliver as output a structured data set containing all the relevant information present on that site. It is based on a novel, knowledge-driven approach that combines low-level annotations with high-level domain knowledge and sophisticated analysis rules encoding common Web design patterns. The first prototype for the UK real-estate domain outperforms existing data extraction tools and validates the premise that with a thin layer of domain-specific knowledge, nearly perfect automated data extraction is feasible.

Once a web site is analyzed, the DIADEM engine can provide a one-time copy of all the data of that site, structured according to the provided schema. Alternatively, an extraction expression, formulated in OXPath [2], can be returned that extracts all the data on-demand at high-speed.

1.1 Motivations and Outline

As users get acquainted with on-line search and decision support systems, their information needs evolve, their

*The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 246858 (DIADEM) and the 2008 Call for "IDEAS Advanced Grants" as part of the Search Computing (SeCo) project.

[1] `diadem-project.info`.

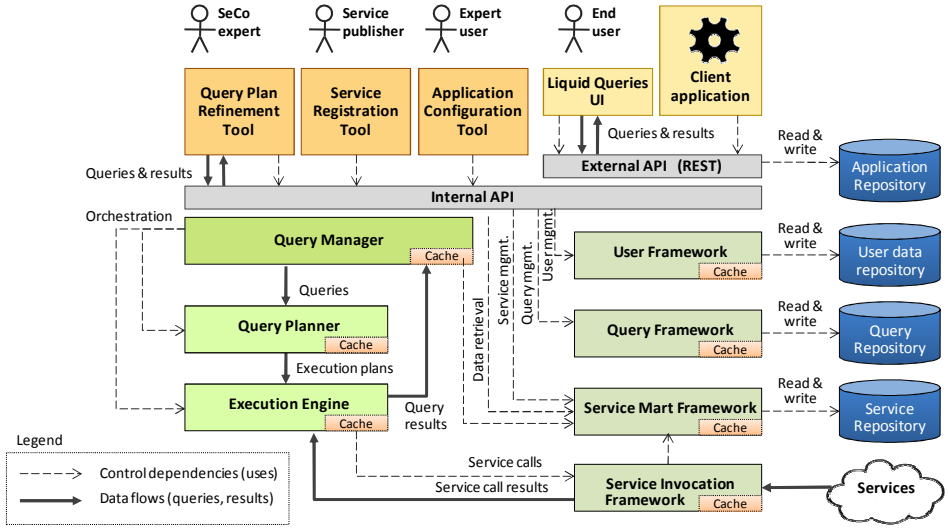

Figure 1: The Search Computing architecture

queries become more and more complex, and their demand for correct and updated data increases. Whilst data extraction approaches such as DIADEM can greatly improve the quality of available information, the need arises for systems and tools able to holistically tackle the problem of complex queries, while enabling users to select, explore and combine data sources in a customized way. A tight integration of DIADEM and SeCo can provide an answer to such need by combining high-precision data extraction, multi-domain service integration, and exploratory search interaction [3]. We demonstrate how the data extraction facilities provided by DIADEM enable the data integration performed in SeCo to easily achieve novel, multi-domain search services over large number of Web sites.

The paper is organized as follows: Section 2 describes the search computing approach to information integration, Section 3 presents the DIADEM approach to data extraction, Section 4 discusses integration issues, Section 5 concludes the paper.

2. WEB DATA INTEGRATION WITH SEARCH COMPUTING

Figure 1 shows an overview of the Search Computing framework, which comprises several sub-frameworks. The service description framework (SDF) provides the scaffolding for wrapping and registering data sources in service marts, describing the information sources at different levels of abstraction. The user framework provides functionality and storage for registering users, with different roles and capabilities. The query framework supports the management and storage of queries as first class citizens: a query can be executed, saved, modified, and published for other users to see. The service invocation framework masks the technical issues involved in the interaction with the service mart, e.g., the Web service protocol and data caching issues. The core of the framework aims at executing multi-domain queries. The query manager takes care of splitting the query into subqueries (e.g., "Which jobs as Java developer are available in the Silicon Valley?", "Where are affordable, nearby flats?", "Where are good schools?") and binding them to the respective relevant data sources registered in the service mart repository; starting from this mapping, the query planner produces an optimized query execution plan, which dictates the sequence of steps for executing the query. Finally, the execution engine actually executes the query plan, by submitting the service calls to designated services through the service invocation framework, building the query results by combining the outputs produced by service calls, computing the global ranking of query results, and producing the query result outputs in an order that reflects their global relevance.

3. AUTOMATIC DATA EXTRACTION WITH DIADEM

A framework such as SeCo allows the user to search for objects with a given specification rather than just for potentially relevant Web documents as keyword search engines. To that end, *structured data* is required, where objects and their attributes are described in a well understood schema. Unfortunately, most commercial Web sites do not provide their objects (such as job listing, properties, or products) as structured data. This is particularly true for businesses with little technical expertise.

Automatically turning existing Web sites into structured data has been mostly an unrealized dream in the past. Previous approaches to fully-automated data extraction addressed the problem by investigating general techniques that can be applied to any web site [4]. W.r.t. existing approaches, DIADEM is based on a fundamental observation: if we combine knowledge about a domain (e.g., that a four figure price is more likely a rent price than a sales price in real estate) with knowledge about the appearance of objects and search facilities in that domain (phenomenology), we can automatically derive an extraction program for nearly any web page in the domain. The resulting program produces high precision data, as we use domain knowledge to improve recognition and alignment and to verify the extraction program based on ontological constraints [5].

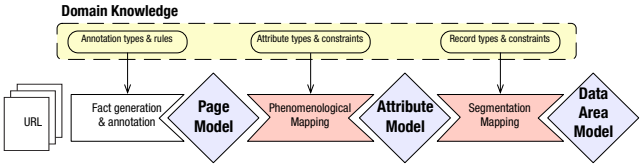

Figure 2: DIADEM's result-page analysis

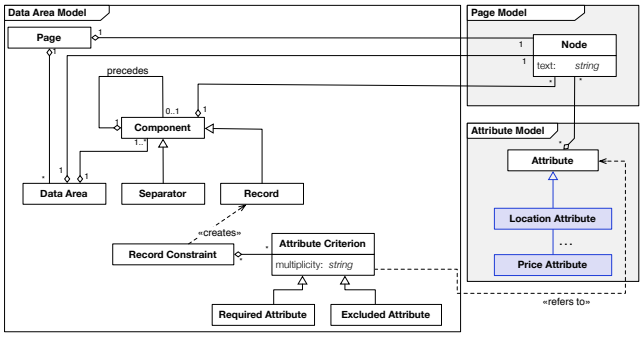

Figure 3: DIADEM's result-page model

DIADEM operates in two modes: in the analysis mode a web site is scrutinized to find relevant objects and search forms and to understand how to extract all data from that site. In the extraction mode, this knowledge is used to extract all data at high speed, assuming that the site has not changed fundamentally since the analysis.

In analysis mode, DIADEM answers primarily three questions: **(1)** How do we have to navigate the site (e.g., by clicking on links, following pagination links, etc.) to extract all the results? **(2)** Are there any forms to fill and how to fill them to find all results? **(3)** How are result records and their attributes structured and displayed? For each of these questions, DIADEM uses both domain-independent heuristics encoding typical web design patterns and domain-dependent clues and high-level knowledge to locate specific objects and their attributes and to verify and align the resulting structured data. Except for a thin browser interaction layer and some off-the-shelf machine learning tools, the whole process is encoded in logical rules maybe involving probabilistic knowledge.

Finally, all the collected models are passed to the OXPath generator that uses simple heuristics to create a generalized OXPath expression for use in extraction mode.

To illustrate how DIADEM analyses a Web site, we focus on result page analysis (the third question), see Figure 2. First we extract the page model from a live rendering of the Web page. This model logically represents the DOM tree of the page along with information on the visual rendering (e.g., CSS boxes), and linguistic annotations. The information provided by the browser model is mainly domain-independent (e.g., DOM structure and CSS boxes) while some of the linguistic annotations are generated by *domain-specific* gazetteers and rules. In the next step, we locate mandatory attributes of the records that we expect to find on a web page of a given domain; then, we proceed to the segmentation of the page into records through domain-independent heuristics. The identified records are then validated using a result-page model, see Figure 3.

Not only HTML. In many domains non-HTML data makes up a small, but significant part of the description of objects, usually as PDF documents, but sometimes just as bitmap images. Sometimes, this information is just supporting the structured data (e.g., the pictures of a car an auto-trading website); in other cases, however, these web resources carry additional information that is not present in the structured data and therefore cannot be accessed by either traditional nor object search-engines.

For instance, in almost all the UK real-estate Web sites, users cannot search for an apartment by energy efficiency or by size of the rooms despite this information is clearly present on the websites. The reason is that the energy efficiency of a house is published as an EPC (Energy Performance Certificate) chart[2] and the sizes of the rooms are published in the floor-plan images.

The automated extraction of this data is non trivial since it might require computer vision and OCR techniques. DIADEM addresses this problem by exploiting the knowledge of the domain to improve existing image and PDF/PS analysis techniques. As an example, the structure of the EPC charts is standardized by a EU directive, therefore it is easy to "reverse-engineer" their semantics. For PDF brochures, it is possible to adopt analysis techniques similar to those adopted for HTML, since the structure of such documents is also reducible to few patterns that can be easily identified by an automatic analysis.

4. TOWARD MULTI-DOMAIN, AUTOMATED WEB DATA CONSUMPTION

Our approach for the integration of structured and unstructured Web data sources is based on a service-oriented vision of the resources. The source integration operates at three levels: *wrapping*, *registration*, and *invocation*.

Service wrapping consists in implementing appropriate wrapping components that take care of invoking the services and manipulating the input and output so as to be consistent with the formats expected by the integration platform. The SeCo platform natively supports generic Web services, relational databases, YQL services, SPARQL endpoints, etc. However, the system is open to support additional data source types.

We suggest two ways for integrating DIADEM data sources into SeCo. In both cases, we assume that the schema used in SeCo matches (a fragment of) the domain ontology used in DIADEM. The first, *off-line* approach extracts all the data of a site contxtually with the analysis and stores it, e.g., in an RDF database together with the domain ontology. This database can be accessed as any other SPARQL endpoint. The advantage of this approach is that it provides very good query performance, but at the cost of storage and consistency. In domains with fast changing data, the database will often be outdated compared to the data on the live web site.

This deficit is addressed by the *on-line* approach, where an OXPath expression is generated by the DIADEM analysis and that expression is executed to extract the data at query time. A slightly specialized OXPath invoker is needed for this approach, as it needs to store the OXPath expression together with possible parameters for form filling. OXPath returns the extracted data in XML or RDF format struc-

[2] `wikipedia.org/wiki/Energy_Performance_Certificate`

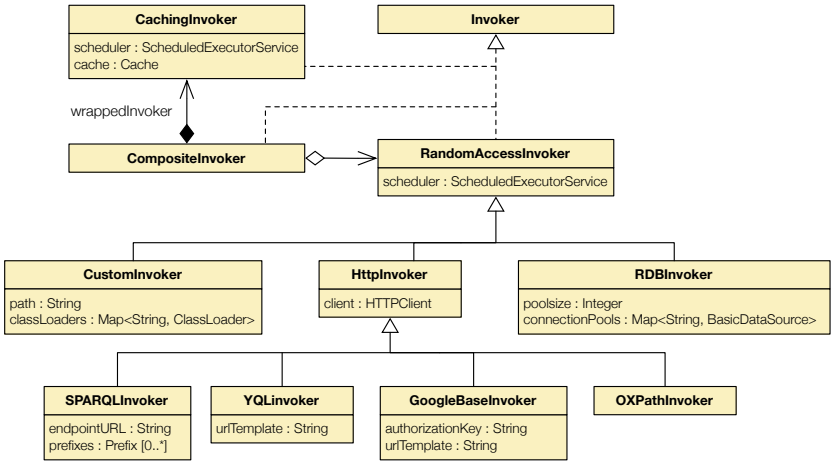

Figure 4: UML class diagram of the SeCo invokers

tured according to the SeCo schema. The latter is ensured by the construction process in the analysis, where the SeCo schema in form of the high-level DIADEM ontology is used to verify the extraction expression.

The disadvantage of this approach is that for large or complex Web sites extraction may take too long for on-line queries. This can be slightly alleviated by the high-level caching provided in SeCo. In the future, we plan to investigate techniques for incremental data extraction, where only new data is extracted. This is also useful for the off-line approach if frequent updates are desired.

Service description in SeCo is based on the registration of services within the Service Description Framework model, which describes services at three levels of abstraction: Service Marts (abstractions of several Web services dealing with the same conceptual objects available on the Web such as "flights", "hotels", "restaurants"), Access Patterns (a specific signature of the Service Mart with the characterization of each attribute as input, output, and/or ranking), and service interfaces (a description of the invocation interface of an actual source service)—leading from the conceptual representation of Web objects to the implementation of search services. If we combine SeCo with DIADEM, we can easy instantiate service descriptions for any Website of a domain. Starting from a description of conceptual objects of a domain, shared between the SeCo service marts and the DIADEM high-level ontology, DIADEM can automatically recognize existing access patterns (by form analysis) and translate them into SeCo service descriptions.

Service execution is performed by an engine, which exploits the Service Description Framework. The execution engine consists of a runtime (a Panta Rhei [6] interpreter able to translate an execution plan in a coordinated sequence of service invocations) and a set of service invokers. Low-level service invokers (one for each data source type, including the one for on-line DIADEM sources) are implemented and follow the chain of responsibility pattern (see Figure 4). There is no need for a special invoker for off-line DIADEM sources, as those reduce to SPARQL Invokers where the data is the result of the off-line extraction. An high-level caching invoker wraps the sequence of low-level invokers to read results from the cache.

5. CONCLUSIONS

Rich object search is one of the major challenges in Web research. In this paper, we show how a combination of SeCo and DIADEM has the potential to address the major challenges involved in object search: **(1)** the *integration of multi-domain* data sources including an easy interface for formulating and refining expressive, multi-domain queries. **(2)** the *automatic extraction* of highly accurate, structured data from most existing web sites.

We plan to further investigate the integration of SeCo and DIADEM. In particular, a further alignment of the conceptual descriptions, access patterns, and service interfaces would be useful. We are currently investigating the automatic extraction of rich access patterns and integrity constraints from existing Web forms. We also plan to develop techniques for incremental data extraction to allow the wrapping of time-sensitive services.

6. REFERENCES

[1] Ceri, S., Brambilla, M., eds.: Search Computing Trends and Developments. Volume 6585. Springer (2011)

[2] Furche, T., Gottlob, G., Grasso, G., Schallhart, C., Sellers, A.: Oxpath: A language for scalable, memory-efficient data extraction from web applications. In: VLDB. (2011)

[3] Bozzon, A., Brambilla, M., Ceri, S., Fraternali, P.: Liquid query: multi-domain exploratory search on the web. In: Proceedings of the 19th international conference on World wide web. WWW '10, New York, NY, USA, ACM (2010) 161–170

[4] Kayed, M., Kayed, M., Girgis, M.R., Shaalan, K.F.: A survey of web information extraction systems. IEEE Transactions on Knowledge and Data Engineering **18**(10) (2006) 1411–1428

[5] Furche, T., Gottlob, G., et al.: Real understanding of real estate forms. In: WIMS '11, New York, NY, USA, ACM (2011) 13:1–13:12

[6] Braga, D., Corcoglioniti, F., Grossniklaus, M., Vadacca, S.: Panta rhei: Optimized and ranked data processing over heterogeneous sources. In: ICSOC 2010. Volume 6470 of Lecture Notes in Computer Science. Springer (2010) 715–716