Building an Environmental Information System for Personalized Content Delivery

Leo Wanner[1], Stefanos Vrochidis[2], Sara Tonelli[3], Jürgen Moßgraber[4], Harald Bosch[5], Ari Karppinen[6], Maria Myllynen[7], Marco Rospocher[3], Nadjet Bouayad-Agha[1], Ulrich Bügel[4], Gerard Casamayor[1], Thomas Ertl[5], Ioannis Kompatsiaris[2], Tarja Koskentalo[7], Simon Mille[1], Anastasia Moumtzidou[2], Emanuele Pianta[3], Horacio Saggion[1], Luciano Serafini[3], Virpi Tarvainen[6]

[1] Dept. of Information and Communication Technologies, Pompeu Fabra University
[2] Centre for Research and Technology Hellas, Informatics and Telematics Institute
[3] Fondazione Bruno Kessler
[4] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation
[5] Institute for Visualization and Interactive Systems, University of Stuttgart
[6] Finish Meteorological Institute
[7] Helsinki Region Environmental Services Authority

leo.wanner@upf.edu, stefanos@iti.gr, satonelli@fbk.eu, juergen.mossgraber@iosb.fraunhofer.de, harald.bosch@vis.uni-stuttgart.de, ari.karppinen@fmi.fi, Maria.Myllynen@hsy.fi, rospocher@fbk.eu, nadjet.bouayad@upf.edu, ulrich.buegel@iosb.fraunhofer.de, gerard.casamayor@upf.edu, Thomas.Ertl@vis.uni-stuttgart.de, ikom@iti.gr, Tarja.Koskentalo@hsy.fi, simon.mille@upf.edu, moumtzid@iti.gr, pianta@fbk.eu, horacio.saggion@upf.edu, serafini@fbk.eu, Virpi.Tarvainen@fmi.fi.

Abstract. Citizens are increasingly aware of the influence of environmental and meteorological conditions on the quality of their life. This results in an increasing demand for personalized environmental information, i.e., information that is tailored to citizens' specific context and background. In this work we describe the development of an environmental information system that addresses this demand in its full complexity. Specifically, we aim at developing a system that supports submission of user generated queries related to environmental conditions. From the technical point of view, the system is tuned to discover reliable data in the web and to process these data in order to convert them into knowledge, which is stored in a dedicated repository. At run time, this information is transferred into an ontology-structured knowledge base, from which then information relevant to the specific user is deduced and communicated in the language of their preference.

Keywords: environmental information service, environmental node discovery, knowledge, personalization, infrastructure, services.

1 Introduction

Citizens are increasingly aware of the influence of environmental and meteorological conditions on the quality of their life. One of the consequences of this awareness is the demand for high quality environmental information that is tailored to one's specific context and background (e.g. health conditions, travel preferences, etc.), i.e., which is personalized. Personalized environmental information may need to cover a variety of aspects (such as meteorology, air quality, pollen, and traffic) and take into account a number of specific personal attributes (health, age, etc.) of the user, as well as the intended use of the information. So far, only a few approaches have been proposed with a view of how this information can be facilitated in technical terms. All of these approaches focus on one environmental aspect and only very few of them address the problem of information personalization [1], [2], [3]. We aim to address the above task in its full complexity.

In this work, we take advantage of the fact that nowadays, the World Wide Web already hosts a great range of services (i.e. websites, which provide environmental information) that offer data on each of the above aspects, such that, in principle, the required basic data are available. The challenge is threefold: first, to discover and orchestrate these services, second, to process the obtained data in accordance with the needs of the user, and, third, to communicate the gained information in the user's preferred mode. To address this problem, we need to involve a considerable number of rather heterogeneous applications and thus an infrastructure that is flexible and stable enough to support a potentially distributed architecture. In what follows, we first outline the process of the discovery of the environmental services (also referred to as *nodes*) in the Web. This is considered as the prerequisite step for enable the retrieval capabilities of the system. Then, we describe briefly the tasks involved in the processing of the data obtained from the environmental nodes until their delivery to the user, and finally present the infrastructure designed to accommodate for both the discovery itself and the posterior tasks.

2 Discovery of Environmental Nodes

As already pointed out above, the web hosts a large amount of environmental (meteorological, air quality, traffic, pollen, etc.) services, which include both (static or dynamic) public webpages that offer environmental information worldwide, as well as dedicated environmental web services with free access. Especially the number of meteorological services that cover each major location is impressive. However, the fact that environmental information is highly distributed and available in heterogeneous forms and formats makes the problem of the discovery and extraction of information from webpages that provide environmental information a serious challenge. Still, it can be considered to be a problem of domain-specific web search, such that methodologies from this area can be applied to implement a node discovery framework.

We apply two types of methodologies of domain search: (a) the use of existing search engines for the submission of domain-specific automatically generated queries,

and (b) focused crawling of predetermined websites [4]. To perform the queries generated by combining domain information from ontologies and geographical input obtained by geographical web services, we use a web search API (e.g., as offered by Yahoo). The queries are expanded by keyword spices [5], which are domain specific keywords extracted with the aid of machine learning techniques from environmental websites. In parallel, a set of predefined environmental websites is further enriched using a focused crawler, which is capable of exploring the web in a directed fashion in order to collect other nodes that satisfy specific criteria related to the content of the source pages and the link structure of the web.

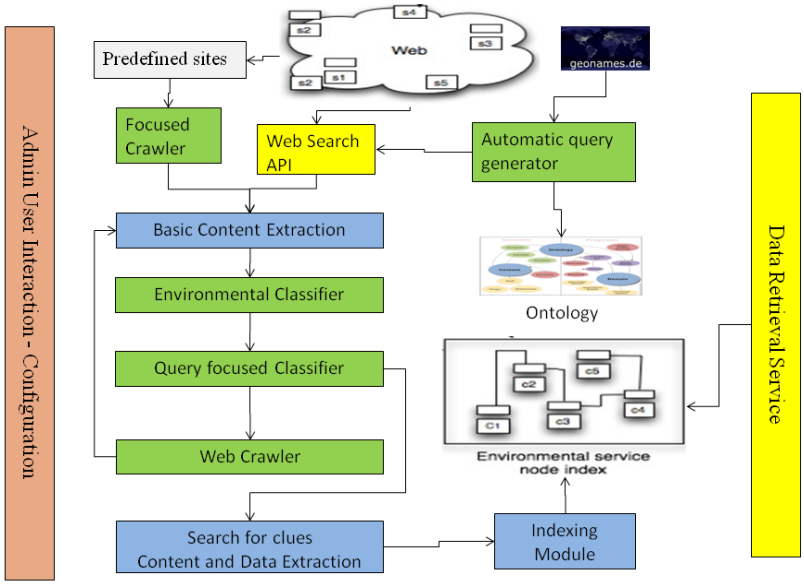

Fig. 1. Architecture for the discovery of environmental service nodes.

The output of the search is post-processed in order to:

(i) separate relevant from irrelevant nodes;

(ii) categorize and further filter the relevant nodes with respect to the types of environmental data they provide (air quality, pollen, weather, etc.);

(iii) parse the body and the metadata of the relevant webpages in order to extract the structure and the clues that reveal the information presented;

(iv) identify internal and external links of the retrieved webpages, which can be further explored by crawlers. The determination of the relevance of the nodes and their categorization is done using a classifier that operates on a weight-based vector of key phrases and concepts from the metadata and content of the webpages. The procedure of the exploration of the external links is recursive and terminated by a manually-set threshold. The information obtained with respect to each relevant node

is indexed in a repository as its finger print, which can be accessed and retrieved by the system through the data retrieval service.

The whole discovery procedure is automatic, however an administrative user could intervene through an interactive user interface, in order to select geographic regions of interest to perform the discovery, optimize the selection of keyword spices and parameterize the training of the classifiers. Figure 1 shows the architecture of the discovery of environmental service nodes.

3 Processing, Orchestration and Retrieval of Environmental Nodes

Once the environmental nodes have been detected and indexed, they are available as data sources or as active data consuming services (if they require external data and are accessible via a web service).

The following tasks still need to be resolved, in order to be able to offer a user-tailored environmental information service.

1. **Orchestration of environmental service nodes**:

Environmental service nodes encountered in the web may require input data provided by other service nodes. In order to obtain all necessary data, the environmental service nodes must thus be "orchestrated", i.e., selected and chained. This presupposes the selection of appropriate protocols and the use of appropriate data interchange formats. To decide which nodes are to be selected over which other nodes, or which nodes fit best together, node quality criteria must be taken into account that are measured by data uncertainty and service confidence metrics derived by using statistical measures, machine learning and visual analytics techniques.

2. **Identification of user relevant service nodes**:

The process of user-tailored information delivery consists of the identification of environmental service nodes in the compiled repository that are relevant to the query of the user, their profile and their context. This is not trivial, given that a user may be moving, be located in an area which is not directly covered by any node, etc.

3. **Extraction and distillation of the data from the webpages of the nodes**:

To distill the data from webpages, advanced natural language processing techniques are needed for webpage parsing, information extraction and text mining. Although these techniques can be tuned to deal with the presentation mode of environmental (i.e., air quality, meteorological, traffic, etc.) data and information, the task of webpage scraping remains a very challenging task. In particular, given a service node, all and only the relevant data (e.g. all the temperature measurements for a city reported in a weather forecast website, but not advertisements) must be extracted.

Given the fact that much information in environmental websites is encoded as images and maps, we also plan to employ image analysis, with the goal to extract information from them.

4. **Converting the data into content**:

In order to guarantee a motivated orchestration of heterogeneous environmental service nodes and offer user-tailored decision support services and environmental information production, we need to convert the data into structured unified content,

which will allow for application of intelligent reasoning algorithms. To this end, the extracted environmental information is integrated into an environmental knowledge base (KB).

Our KB, which is codified in the standard semantic web ontology language OWL [8], covers environmental content such as meteorological conditions and phenomena, air quality, and pollen, as well as other relevant environment-related content essential for the targeted user-tailored service: travel and traffic information, human diseases, geographical data, monitoring station details, user profile details, etc. In addition, the KB is also capable of formally representing the description of the user's inquiry.

The current version of the KB contains around 202 classes, 143 attributes and properties, 463 individuals[1]. Its Description Logic (DL) expressivity is *ALCHOIQ(D)*. The KB has been obtained by (i) including customized version of currently available ontologies (e.g., parts of the SWEET ontology), (ii) automatically extracting key concepts from domain relevant text sources, and (iii) manually adding additional properties and attributes.

5. **Fusion of environmental content**:

Environmental service nodes may provide competing or complementary data on the same or related aspect for the same or the neighboring location. To ensure the availability of a most reliable and comprehensive content as basis for further processing stages, the content proceeding from these nodes must be fused. As already in the case of node orchestration, this implies an assessment of the quality of the contributing services and data.

6. **Assessment of the content with respect to the needs of the user**:

Once the data from the nodes have been incorporated into the KB, they need to be evaluated and reasoned about in order to infer how they affect the addressee, given his/her personal health and life circumstances and the purpose of the request of the information. For instance, a citizen may request information because he/she wants to decide upon a planned action, be aware of extreme episodes or monitor the environmental conditions in a location.

7. **Selection of user-relevant content and its delivery**:

Not all content in the KB is apt to be communicated to the addressee: some of it would sound trivial or irrelevant. Intelligent content selection strategies that take into account the background of the user and the intended use of the information are thus needed to decide which elements of the content are worth and meaningful to be communicated. To deliver the selected content, techniques are required that present the content in a suitable mode (text, graphic and/or table) in the language of the preference of the addressee.

8. **Interaction with the user**:

The interaction between the system and the user is also an important aspect of this work. The user must be able to formulate the problem in a simple and intuitive format – be it based on natural language or on graphical building blocks. The user should also receive the generated information in a suitable form and, as already mentioned above, in the language of his/her preference.

[1] These data refer to the "empty" KB, i.e. without considering any environmental data coming from the nodes.

4 Service-Oriented Infrastructure

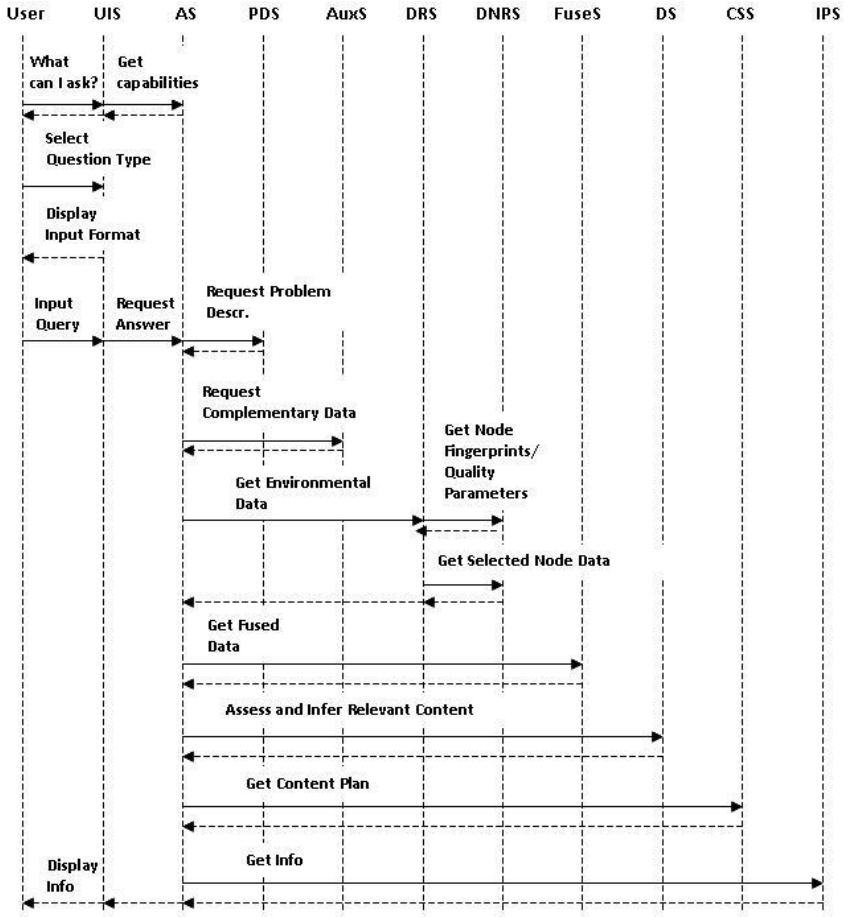

Fig. 2. Sequence diagram for the execution of the services for delivery of environmental information

In order to accommodate for all tasks described above, we opt for a service-based architecture. This architecture is based on a methodology which has been developed in ORCHESTRA [6] for risk management, and which has been extended in SANY [7] to cover the domain of sensor networks and standard-based sensor web enablement. The focus of this methodology is on a platform neutral specification. In other words, it aims to provide the basic concepts and their interrelationships (conceptual models) as abstract specifications. The design is guided by the methodology developed in the ISO/IEC Reference Model for Open Distributed Processing (RM-ODP), which explicitly foresees an engineering step that maps solution types, such as information

models, services and interfaces specified in information and service viewpoints, respectively, to distributed system technologies.

We defined application-specific major tasks and actions as abstract service specifications, which can be implemented as service instances on a specific platform. Web service instances for these services are currently being developed, which can be redefined and substituted as needed.

Figure 2 displays a simplified sample workflow with the major application services in action. Two services are not cited in Figure 2 since they are consulted by nearly all other services: the Knowledge Base Access Service and the User Profile Management Service. The figure also does not include the services related to data discovery (Figure 1) and information distillation from webpages.

A main dispatcher service (called Answer Service, AS) controls the workflow and the execution of the services. First, the user interacts with the system via the User Interaction Service (UIS). In the case that the user is unsure with respect to the types of information they can ask for, he/she can inquire this information by requesting it from the Problem Description Service (PDS).

To ensure a full comprehension of the problem or user generated question, we decided to operate with controlled graphical and natural language input formats. Once the user has chosen what kind of question he wants to submit to the system, the UIS provides the user the corresponding formats. Thereupon, the user can formulate his/her query, which is subsequently translated by the PDS into a formal ontology-based representation understood by the system. After the problem description is generated, this is passed by the UIS to the AS as a “Request Answer” inquiry. Then, the AS assesses what kinds of data beyond environmental data are required to answer the query of the user and solicits these data from the Auxiliary Services (AuxS). For instance, such services can provide travel route information in the case that the user's query concerns the environmental conditions for a bicycle tour from A to B.

After having acquired the complementary data, the AS can request from the Data Retrieval Service (DRS) the environmental data needed to answer the user query. The DRS solicits these data from the environmental nodes that were identified by the Data Node Retrieval Service (DNRS) as relevant to the user's query and the complementary data. The DNRS retrieves this information from the data node repository after the node discovery phase has taken place.

As already mentioned, the retrieved nodes may deliver complementary or competing data of varying quality (to keep the presentation simple, we dispense with the illustration of the orchestration of service nodes). The Fusion Service (FS) applies uncertainty metrics to obtain the optimal and maximally complete data set, which is passed by the AS to the Decision Service (DS). The DS converts the data set into knowledge, or content, in that it relates it to the knowledge in our KB, reasons about it, and assesses it from the perspective of its relevance to the user. From this content, the Content Selection Service (CSS) compiles a content plan, which contains the knowledge to be communicated to the user as the answer. The Information Production Service (IPS) takes the content plan as input and generates information in the language and mode (text, table, or graphic) of the preference of the user, which then is passed to the user.

5 Outlook

We are currently implementing the described service infrastructure – including the environmental node discovery. The first operational prototype of the service will be available for demonstration by early summer 2011.

Acknowledgments. This work is partially funded by the European Commission under the contract number FP7-248594 “Personalized Environmental Service Configuration and Delivery Orchestration” (PESCaDO).

References

1. Karatzas, K.: State-of-the-art in the dissemination of AQ information to the general public. In: Proceedings of EnviroInfo, vol. 2, pp. 41--47. Warsaw (2007)
2. Peinel, G., Rose, T., San José R.: Customized Information Services for Environmental Awareness in Urban Areas. In: Proceedings of the 7th World Congress on Intelligent Transport Systems. Turin (2000)
3. Wanner, L., Bohnet B., Bouayad-Agha, N., Lareau F., Nicklass, D.: MARQUIS: Generation of User-Tailored Multilingual Air Quality Bulletins. J. Applied Artificial Intelligence. 24 (10), 914--952 (2010)
4. Wöber, K.:. Domain Specific Search Engines, In: Fesenmaier, D. R., Werthner, H., Wöber, K. (eds.) Travel Destination Recommendation Systems: Behavioral Foundations and Applications, 205—226. Cambridge, MA: CAB International, (2006)
5. Oyama, S., Kokubo, T., Ishida, T.: Domain-Specific Web Search with Keyword Spices Awareness in Urban Areas. J. IEEE Transactions on Knowledge and Data Engineering. 16 (1), 17--24 (2004)
6. Usländer, T. (ed.): Reference Model for the ORCHESTRA Architecture Version 2.1. OGC Best Practices Document 07-097, http://portal.opengeospatial.org/files/?artifact_id=23286 (2007)
7. Usländer, T.: Specification of the Sensor Service Architecture, Version 3.0 (Rev. 3.1). OGC Discussion Paper 09-132r1. Deliverable D2.3.4 of the European Integrated Project SANY, FP6-IST-033564, http://portal.opengeospatial.org/files/?artifact_id=35888\&version=1 (2009)
8. World Wide Web Consortium: OWL Web Ontology Language Reference, http://www.w3.org/TR/owl-overview/