RETRIEVAL OF HETEROGENEOUS SERVICES IN C^{2}NIWA REPOSITORY

JULIAN SZYMAŃSKI

Department of Computer Systems Architecture
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology, Poland

(received: 26 May 2015; revised: 29 June 2015;
accepted: 7 July 2015; published online: 1 October 2015)

Abstract: The paper reviews the methods used for retrieval of information and services. The selected approaches presented in the review inspired us to build retrieval mechanisms in a system for searching the resources stored in the C^{2}NIWA repository. We describe the architecture of the system, its functions and the surrounding subsystems to which it is related. For retrieval of C^{2}NIWA services we propos three approaches based on: keyword search, hierarchical catalog and searching with the use of a specification of meta values. The proposed functionality allows us to increase the visibility of the University competences. We propose a module for managing external proposals to carry out cooperative projects using the know-how of the University.

Keywords: C^{2}NIWA, information retrieval, documents categorization, services classification

1. Introduction

In recent times, the development of effective techniques for accessing information has become a hotter and hotter topic. New web browsers are created each and every year, the network bandwidth increases, and the number of digital documents grows. As the amount of the available information is huge, and continuously growing, it requires more and more efficient methods that enable access to the required information. Over the years, different methods have been applied for different domains aiming to add mechanisms that allow us to retrieve the information that fulfills the user's needs. The Holy Grail of retrieval methods is the communication with machines in such a way as humans communicate with each other – using a natural language. As this task involves implementation of mechanisms that allow us to model the understanding of the surrounding world, its realization is hard. Despite very promising advances in that area *e.g.*: IBM Watson[1] [1], in question answering, application of deep neural networks for

1. http://www.ibm.com/smarterplanet/us/en/ibmwatson/

understanding the content of the images [2] or providing the factual answers for the queries made by WolframAlpha[2], the lexical competences of the machines are still far from understanding the real language, thus approximate methods for accessing the information are used.

2. Information Categorization and Retrieval

Different user needs entail different methods enabling access to information which is useful in different areas. The most popular and general approach is retrieval with keywords [3, 4]. The use of keywords is a kind of a bridge in the domain of human-machine communication: the machine can interpret sequences of characters, the human is able to specify his/her information need using fragments of statements. This method is now the dominant trend used for finding the relevant content in the Web. It should be stressed that the approach is a compromise between humans an machines that is reached on an inner communication patch. Keywords are relatively easy for the machines to find in the repositories of the text documents, humans are able to express their information needs in such a form. This method, up till now, has been sufficient for a large number of information retrieval tasks, especially, if the user is looking for resources where he or she can find the information needed, but not obtain a direct answer for the query. The retrieval with keywords allows one to limit the amount of content that is potentially useful for the user and is presented in a so called SERP (Search Engine Results Page). It should be noticed that despite the limitation of the retrieved content to such that contains specified keywords, a SERP set is usually large and requires additional methods for filtering the relevant elements. Basic mechanisms that process SERP sets are based on ranking methods [5] that sort it according to a particular relevance measure. A very successful relevance measure that shows high applicability for information retrieval [6] was proposed by Google in their algorithm named Page Rank [7]. The success of this method has been based on the observation that the content provided to the user should be ordered according to the relevance of pages. In Page Rank it is based on evaluation of the page quality measured with the weighted number of links that reference a given page, where the weight is related to the page quality. The computation of the rankings requires an iterative algorithm that converges to a stable vector of values that describe the quality of nodes in a WWW graph [8]. This initial success of the ranking based on global quality measures shows its applicability while the SERP set constructed with keywords was relatively small. Further development of the ranking methods consists in that that they modify the ways in which the ranking is calculated, *e.g.* using two ranking measures for describing the so called Hubs and Authorities within a WWW graph [9]. The progress in information growth in WWW is the reason why the measure based only on values of nodes quality calculated from simple relations taken from a WWW graph becomes insufficient. Improvements

2. http://www.wolframalpha.com

have been made by adding additional information in the form of so-called signals that describe the information quality. Realizations of particular signals become a patent-protected company secret. Most of them were based on measuring the accuracy of objective information provided by the source, its uniqueness, up-to-dateness and thematic coherence. Other aspects taken into consideration include: the domain reliability, reputation, age, relation of keywords to the domain name, [WHOIS public or hidden declaration of ownership. Some sources[3] report that Google uses over 200 such signals.

It should be noticed that the above-presented retrieval mechanism ranks the search results that have been retrieved by the specified keywords. The extension of this method provides mechanisms that allow expanding the query and adding pages to the SERP that are thematically related, but do not contain the keywords provided by the user [10].

The development of the retrieval methods brings a mechanisms that additionally personalize the retrieved information [11, 12]. In the basic retrieval model, two different users seeking for a particular piece of information using the same keywords will get the same SERP. Current search engine providers, to improve the relevance of the search results, for the ranking also involve mechanisms based on observation of the user behavior. This analysis allows them to create user profiles, thus the content retrieved by the search engine is also dependent on the type of user that retrieves the information. The mechanism of user profiling makes large concerns regarding the user privacy. Despite that, a lot of people use web browsers owned by search engine providers, and surf the web being logged in. It allows the information providers to trace the user and grab a lot of personal information that gives a big advantage over potential competitors. The advantage of large search engines providers comes from possessing the huge amount of data produced by the search users. It allows the monopolists to perform large scale analysis, and construct rankings based on behavior observation [13]. This analysis allows search engines to modify the rankings in a way that is not possible to other companies not having access to such data, *e.g.* they can increase the position of the popular pages, that have been selected by the users as a first checked page.

Identification of the thematic domains of a given web page has been used in systems that aggregate the search results and, instead of presenting a list of ranked results to the user, they provide thematically coherent sets of web pages [14, 15]. The approach uses unsupervised machine learning techniques [16] that are based on particular similarity measures. They are able to combine web pages into so-called thematic clusters that are presented to the user. One of the examples of such a system is http://yippy.com. The site works as a clustering web meta search engine, that presents the results in the form of a structure aggregated into hierarchy and interactive to the user that organizes the SERP.

The concept of providing the user with a structure that organizes the repository of information is not new. Early implementations have been used in

3. http://secureglass.net/200-czynnikow-rankingowych-google

library systems where books were categorized into predefined thematic catalogs. The catalogs in digital libraries offer a meta-description layer for the content stored in the library. Typically it is implemented in the form of an interface that employs both approaches: retrieval based on keywords and navigation over the category structure. One of the well-known examples of a thematic catalog is the Association for Computing Machinery (ACM) [17] directory that provides a system for categorization of educational and research outcomes.

The approaches to formalize the standards of description for cataloging the content have been subject of many studies. One of the first is Dublin Core [18] that offers a small set of vocabulary terms used to describe, as well as web (video, images, web pages, *etc.*), and physical resources such as books, CDs, and objects like artworks. The idea of cataloging states behind the DMOZ – the Open Directory Project [19], that aims at categorizing web resources. The system supported by AOL offers a searchable people-reviewed web directory categorized by language, subject and location. The classification made by DMOZ is a supervised approach to organize the WWW content. Another example of using a dedicated system of categories that has been equipped with a controlled dictionary is the initiative developed by the U.S. National Library of Medicine. The dictionary called Medical Subject Headings (MeSH) [20] has been created for indexing journal articles and books in the life sciences. It has found many applications in searching, especially in the medical database PubMed where it supports the index catalog as well as other databases related to health, *e.g.* ClinicalTrials.gov.

Another approach for structuralisation of resources to improve the search is a Google initiative aimed at building the knowledge base in the form of concepts interconnected into a network. The knowledge graph[4] constructed from integration of a wide variety of sources provides an external database for semantic-search information. Interlinking the Web resources is also behind the initiative called Linked Data [21]. The aim of this approach is to enable data from different sources to be connected so that it can be interlinked and become useful through semantic queries based on a dedicated language [22]

3. Services in Heterogeneous Environments

The Web is not only the textual content. In recent years we have been observing rapid growth of multimedia content as well as increasing deployment of services used for machine to machine communication [23]. For the effective processing and integration of the services a formalized description of their interfaces is required. Thus, approaches such as WSDL (Web Services Description Language) standards have been built to offer unified definitions based on the XML language for describing the functionality provided by the Web Service [24]. Universal Description Discovery and Integration (UDDI) provides a method for publishing

4. https://www.google.com/intl/en/search/about/insidesearch/features/search/knowledge.html

and finding service descriptions [25]. The service description information defined with WSDL is complementary to the information found in a UDDI registry that stands for the mechanism to register and locate web service applications. A UDDI business registration consists of three components:

- White Pages – address, contacts, and known identifiers;
- Yellow Pages – industrial categorizations based on standard taxonomies;
- Green Pages – technical information about services offered by the business.

The description of the services compromise humans and machines: on the one hand the retrieval of services should be based on a human-friendly way, but on the other hand, the description should keep restrictions of formalization for a machine-processable form. Thus, the concept of the catalog has been used for the formal organization of services. Using it, service-oriented architectures allow the companies efficient and effective support of their business processes [26]. Discovery of services whose functionality fulfills user requirements may be supported by enhancing the syntactical information by employing additional semantic knowledge. However, services of different providers are generally described with heterogeneous concepts. Thus, mechanisms for inter-operation of services is needed. If services are located within one corporation, a unified directory with standardized descriptions allows effective retrieval and composition of the services used [27]. The formal approach for describing the services is created using meta descriptions of the resources. These descriptions can be organized into so-called ontologies [28]. These are used as metadata knowledge bases that allow incorporating semantics into analysis of interrelations of stored services. This strongly improves the results of the retrieval as stored resources can be processed in a way similar to human inferences [29].

4. System Description

At the Gdansk University of Technology we have a wide range of services. One of the popular implementations of services in computer science are web services. For their unified deployment, we built a dedicated platform called Wiki_WS [30]. Using the platform one can publish the work and make it externally accessed and executed. The platform allows us to unify the computational services that have been created at the University, as well as to provide a framework for building complex tasks using a combination of execution scenarios. Composing complex services with already existing atomic ones is the subject of our current research [31]. For that purpose we employ a Business Process Execution Language (BPEL) [32] that enables us to specify actions within business processes with the web service interfaces exclusively.

In addition to the web services, other (not technical in terms of IT) services are also offered at the University. They can be in the form of on-line courses, articles, tutorials, but the most valuable form is the know-how carried by the University staff. To integrate the service descriptions and to provide an efficient interface for their retrieval we propose the architecture shown in Figure 1.

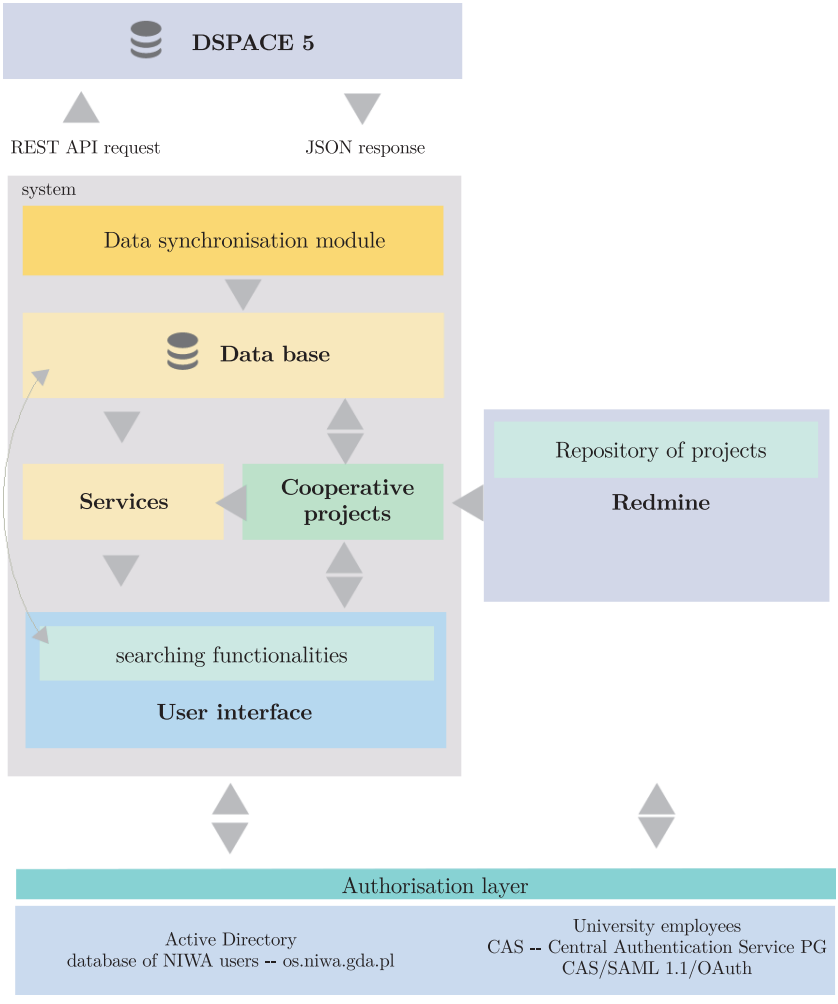

Figure 1. Big picture of the system for retrieval of University services

The descriptions of University services are stored in a repository based on DSpace [33]. The catalog for their organization is started with three super categories that describe general types of the services offered by the C^{2}NIWA:

- Software;
- Publications;
- Services.

Each category groups the specified content of a particular type:

The *Software* category allows storing the source code of libraries, applications, *etc.* that are published as open source and are provided to a wider community. The repository of the C^{2}NIWA data organized by category works as the entrance catalog for the source code management system. Source codes can be stored in a wide range of repositories: such as SourceForge [34], GitHub [35], *etc.*

For storing and hosting the sources within our environment we have used the Redmine [36] system that offers a management system of the IT projects run at the University.

Publications made at the Gdansk University of Technology are categorized under the *Publications* category. It contains sub-directories related to a particular faculty that organizes papers around a particular domain.

The third super category named *Services* aggregates the content related to the deployed resources that are ready to use. It contains sub directories for: Courses, User Guides, Advisory Services, Information Services, Web Services.

The catalog has a hierarchical structure and it can be extended according to the development of the system, introducing more specified granularity of resource groups. The visualization of the catalog enables the functionality for interactive traversing of the resources using the navigation over abstract categories. It gives the opportunity for filtering the thematic domains as well as narrowing the subjects according to the user preferences.

Each of the services in C2 NIWA has a standardized description that unify the way in which it is stored and allows effective retrieval of the resources. The standardized description of the C2 NIWA service is composed of the following fields:

- unique identifier;
- name of service;
- category identifier;
- publication date;
- author;
- description;
- link to resource.

The standardized record of the C2 NIWA service forms meta descriptions of the resources. It allows retrieving entities based on specification of a particular value of the fields. Also, it is possible to combine it with a full-text search within the description of the fields. Combination of these three approaches provides a sufficient mechanism for exploration and retrieval of all resources stored in C2 NIWA repository.

4.1. Authorization and Access Levels

The users of the system are authorized by the central authentication server (CAS) of the Gdansk University of Technology[5]. It guarantees the lack of any redundancy during the registration process as it allows the University staff to use their unique identifiers that they use in other University systems. This integration also allows incorporating the resources from external databases and keep the data coherent with other university systems. The external users that want to use the C2 NIWA system need to register first. In that case the authorization is kept by

5. http://logowanie.pg.gda.pl

local C^{2}NIWA authorization services[6]. After registering, the user gets access to the system functions that allow him/her to make publications of proposals.

The authorization levels assign to roles that are in the system to the user type:

- administrator – has full control over the system;
- moderator – the user that accepts external proposals;
- logged-in user with grants for content publication, we select two types:
  - external – authorized by the C^{2}NIWA services;
  - university employee – authorized by CAS PG;
- ordinary (not logged) user – can only browse the service catalog.

4.2. System user interface

An integral part of the system, that is the most visible, is the user interface. It was composed as a Single Page Application (SPA) that communicates with a server using AJAX (AmplifyJS). This technology fulfills requirements on the application effectiveness, and minimizes usage of the resources during the information exchange using JSON. An example of the interface is shown in Figure 2.

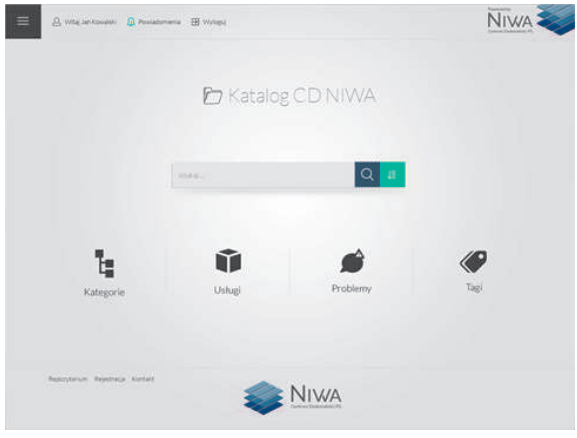

Figure 2. Example of the system interface

The interface allows selecting the way a user interacts with the repository of the services. He or she can go for a look-up to the catalog of the NIWA resources, retrieve them using keyword-based search or use specification of the metadata values. In each of these approaches, the user can introduce other techniques that allow him to narrow and improve the precision of the search results, *e.g.*: after selection of services using the keywords, the user can select the sub-domains of interest using the catalog. Additional retrieval of specific tuples with specification of meta data values allows efficient filtering and narrowing of the results according to the user preferences.

6. http://os.niwa.pg.gda.pl

4.3. Management of Cooperative Projects

The resources stored in the C2 NIWA repository allow us to promote the competences available at the Gdansk University of Technology. They increase the visibility of the work made at the University, and show areas of competence of the people working there. The capabilities of realizing complex tasks using the University know-how, provided by its staff, have been offered to a wider community by the interface for performing the so-called cooperative projects.

To enable this functionality, a dedicated sub-system has been implemented. It allows external entities to introduce a problem to the university community. The problem, having been accepted by the moderator is approved for execution in the University environment. When it becomes visible for people working at the University, it may find a group of people interested in solving it.

From the technical point of view, acceptance of the proposed project, given by the moderator, results in creation of a project in our Redmine system. From now on, the community of people who are interested in this proposal can discuss it, build a team that will work on it, and manage the project development using the Redmine system.

5. Summary

The paper presents a review of the approaches used for retrieval of information as well as methods used for searching for services that contain structuralized descriptions. From the proposed approaches we select three methods which we implement in the C2 NIWA system that integrates and provides University resources. We provide a description of the C2 NIWA interface for retrieval of content, as well we show the satellite systems to which our project is related. We also provide a description of the module used for increasing the visibility of University competences. The module helps the knowledge transfer between University and business environments by providing a functionality of registering the proposals of external entities that they want to make cooperative with University projects.

Acknowledgements

The work was supported as part of the Centre of Competence for Novel Infrastructure of Workable Applications – C2 NIWA.

References

[1] Ferrucci D A 2012 *IBM Journal of Research and Development* **56** (1–1)
[2] Le Q V 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE 8595
[3] Rasolofo Y and Savoy J 2003 *Term proximity scoring for keyword-based retrieval systems*, Springer
[4] Baeza-Yates R, Ribeiro-Neto B, *et al.* 1999 *Modern information retrieval*, ACM press New York, **463**
[5] Liu T Y 2009 *Foundations and Trends in Information Retrieval* **3** 225
[6] Manning C D, Raghavan P and Schütze H 2008 *Introduction to information retrieval*, Cambridge university press Cambridge, **1**

[7] Brin S, Page L 1998 *Computer networks and ISDN systems* **30** 107
[8] Langville A N and Meyer C D 2004 *Internet Mathematics* **1** 335
[9] Kleinberg J M 1999 *Journal of the ACM (JACM)* **46** 604
[10] Xu J, Croft W B 1996 *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM 4
[11] Maleszka M, Mianowska B and Nguyen N T 2013 *Knowl.-Based Syst.* **47** 1
[12] Mulwa C, Lawless S, Sharp M and Wade V 2011 *International Journal of Knowledge and Web Intelligence* **2** 138
[13] Kok A J and Botman A 1988 *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM 343
[14] Leouski A V and Croft W B 2005 *An evaluation of techniques for clustering search results*, Technical report, DTIC Document
[15] Cai D, He X, Li Z, Ma W Y and Wen J R 2004 *Proceedings of the 12th annual ACM international conference on Multimedia*, ACM 952
[16] Gentleman R and Carey V 2008 *Bioconductor Case Studies*, Springer 137
[17] Gold V 2003 *Encyclopedia of Library and Information Science* **1** 174
[18] Weibel S, Kunze J, Lagoze C and Wolf M 1998 *Internet Engineering Task Force RFC* **2413** 132
[19] Sherman C 2000 *Humans do it better: Inside the open directory project* **24** (online)
[20] Lipscomb C E 2000 *Bulletin of the Medical Library Association* **88** 265
[21] Bizer C, Heath T and Berners-Lee T 2009 *Semantic Services, Interoperability and Web Applications: Emerging Concepts* 205
[22] Hartig O, Bizer C and Freytag J C 2009 *Executing SPARQL queries over the web of linked data*, Springer
[23] Chen M, Wan J and Li F 2012 *KSII Transactions on Internet and Information Systems (TIIS)* **6** 480
[24] Christensen E, Curbera F, Meredith G, Weerawarana S, *et al.* 2001 *Web services description language (wsdl) 1.1*
[25] van Steenderen M 2000 *SA Journal of Information Management* **2**
[26] Perrey R and Lycett M 2003 *Proceedings of the Symposium on Applications and the Internet Workshops*, IEEE 116
[27] Rake J, Holschke O and Levina O 2009 *Lecture Notes in Business Information Processing. Business Information Systems*, Abramowicz W, Ed., Springer, **21** 205
[28] Martin D, Burstein M, Hobbs J, Lassila O, McDermott D, McIlraith S, Narayanan S, Paolucci M, Parsia B, Payne T, *et al.* 2004 *W3C member submission* **22** 2007
[29] Klein M and Bernstein A 2001 *The Emerging Semantic Web*
[30] Krawczyk H and Downar M 2012 *Intelligent Tools for Building a Scientific Information Platform*, Springer 251
[31] Budnik L, Dziubich K, Dziubich T and Nasiadka S 2008 *1st International Conference on Information Technology*, IEEE 1
[32] Andrews T, Curbera F, Dholakia H, Goland Y, Klein J, Leymann F, Liu K, Roller D, Smith D, Thatte S, *et al.* 2003 *Business process execution language for web services*
[33] Smith M, Barton M, Bass M, Branschofsky M, McClellan G, Stuve D, Tansley R and Walker J H 2003 *Dspace: An open source dynamic digital repository*
[34] Van Antwerp M and Madey G 2008 *Workshop on Public Data about Software Development (WoPDaSD) at the 4th International Conference on Open Source Systems*, Milan, Italy
[35] Dabbish L, Stuart C, Tsay J and Herbsleb J 2012 *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM 1277
[36] Lesyuk A 2013 *Mastering Redmine*, Packt Publishing Ltd