INFORMATION SEARCHING FOR AN EXPERIENCE MANAGEMENT PLATFORM OF THE EU PELLUCID PROJECT

MARTA MAJEWSKA[1,2], KRZYSZTOF KRAWCZYK[2],
RENATA SŁOTA[1], JACEK KITOWSKI[1,2], LADISLAV HLUCHÝ[3]
AND SIMON LAMBERT[4]

[1] *Institute of Computer Science,*
AGH University of Science and Technology,
Al. Mickiewicza 30, Cracow, Poland
{*kito, rena*}*@agh.edu.pl*

[2] *ACK Cyfronet AGH, Nawojki 11, Cracow, Poland*
{*mmajew, krafcoo*}*@icsr.agh.edu.pl*

[3] *Institute of Informatics, Slovak Academy of Sciences,*
Dubravska cesta 9, Bratislava 84237, Slovakia
hluchy.ui@savba.sk

[4] *CCLRC Rutherford Appleton Laboratory,*
Chilton, Didcot, OX11 0QX, UK
S.C.Lambert@rl.ac.uk

(Received 6 June 2004; revised manuscript received 13 July 2004)

Abstract: The EU Pellucid project is developing an experience management system for public organizations with staff mobility. The paper presents an activity whitin the project focused on searching for information in repositories of documents. The project's background and the process of information searching are described. Ontological methods such as semantic annotation and similarity searching, as well as ontology- and full-text-based searching are presented. Monitoring of organizational repositories is discussed.

Keywords: experience management, ontologies, information retrieval, sematic annotation

1. Introduction

Organizational mobility, which enables employees to move from one department to another, is becoming increasingly common and brings its own problems and opportunities. Many organizations make efforts to train new workers, but do very little to adapt those who move from one position to another. In particular, there is very modest support from computer-based information systems. For employees in new positions assistance in performing their new tasks is a key issue.

The aim of the Pellucid project [1] is to develop a flexible and adaptable platform to assist these employees. The Pellucid platform has been conceived as a remedy for the loss of experience caused by employees' mobility. The platform supports the employees during their integration in a new department by giving access to specific experience accumulated in the past. It thus helps to improve the organization's effectiveness and efficiency by capturing and reusing of employees' experience and storing it for further use.

The paper is organised in two parts. First, an overview of the Pellucid platform is given. The main public organizations' requirements and employee profiles are described in Section 2. The Pellucid platform's architecture and a scheme of experience management are presented in Section 3. The second part is focused on aspects of information searching. Search and Access Layer (SAL) tasks are presented in Section 4. SAL functionality, which includes full-text indexing, ontology-based searching, semantic annotation and repository monitoring, is described in subsections. The paper is concluded with a short summary.

2. Public organizations and mobile employees

Analyses of organizational environments and user profiles are essential elements to support the needs of public organizations. Employee expectations strongly depend on the possessed experience and skills. Several types of employees are distinguished:

- Novice – a totally inexperienced employee (*e.g.* a recent graduate);
- New-hired or Relocated Employee – an employee who has already got some practical skills needed at the given position acquired during earlier work for another employer or in another position whitin the same organization;
- Experienced Employee – an employee who has been working at the given position in the organization for a long period of time (he is considered to be the experience provider).

An experienced employee knows well what he wants to obtain. He knows the organizational repositories and domain-specific issues. However, in many cases he is not able to encompass all information and its location. For this type of user a good searching tool is required that would give precise results and understand the organization's domain. Methods that implement ontology-based searching supported by full-text searching are appropriate in this context.

A medium-level employee could be characterized as a user who is sometimes unable to define precise the result he wants to obtain. The system should be able to search very precisely and, at the same time, provide a possibility to extend the results or range of a query. This is achieved by searching for similar documents and allowing the user to define an ontology-based query, respectively.

Finally, a novice knows very little about work in the organization and about organizational repositories and documents. In the majority of cases such a user will encounter difficulties in specifying a query. He might find it helpful to be actively informed about some documents and obtain abstracts presenting essential information from documents that consist of highly domain-specific information like: abbreviations, codes and other non-standard expressions. These requirements are met by notifications and semantic annotation.

2.1. Information characteristics

A very important factor of a public organization's description is a characterization of information gathered in the organization. The information is mainly stored in the form of several document types. These documents are text documents, spreadsheets, electronic messages, *etc.* The documents are gathered in organizational repositories, such as file systems, portals, e-mail repositories and databases. The semantics of a document is considered to be a key aspect of the searching process, though not easily achieved. The most valuable information is domain- or organization-specific. Each document can be regarded as a collection of words or symbols. An organizational document may include some specific expressions, which are impossible to interpret or understand without domain-specific knowledge. They usually consist of different types of semantic codes, which are strings of characters meaningful in the domain context, as well as specific document structures making a single document a semantically rich database. Moreover, in certain domains some well-known words are used in a disparate sense, whilst most of the others are used in their common connotations. Word meaning can also be determined by a specific document's structure.

3. Conceptual foundations of the Pellucid platform

Pellucid – a Platform for Organizationally Mobile Public Employees is a research and development project funded by the European Union [1]. It is realized within the 5th Framework Programme of the Information Society Technologies (IST-2001-34519). Its purpose is to design, develop and validate a flexible software platform for an important kind of knowledge management: to assist organizationally mobile workers in public sector organizations.

3.1. The Pellucid consortium

The Pellucid consortium is composed of eight organizations from five European countries representing industry, the academia and end-users. Apart from participants from ACK Cyfronet AGH and the AGH-UST Institute of Computer Science, the following are scientific partners engaged in the project, responsible for conceptual design and core implementation:

- the coordinator – Council for the Central Laboratory of the Research Councils (CCRC), UK,
- Institute of Informatics, Slovak Academy of Sciences (II SAS), Slovakia.

The industrial partners, involved in customization, deployment and application, are:

- SADIEL, S.A., Spain,
- Softeco Sismat SpA, Italy.

The Pellucid end-users, responsible for evaluating the platform, are:

- Municipality of Genoa, Mobility and Transport Plan Directorate (CdG), Italy,
- Mancomunidad de Municipios del Bajo Guadalquivir (MMBG), an association of local governments in Spain,
- Consejería de la Presidencia, a body of the regional government of Andalusia (Junta de Andalucía), together with SADESI, a company that operates its call centre for telephony problems.

3.2. Structure of the Pellucid platform

An inherent characteristic of any experience management system is its strong dependency on specific requirements of the target installation site. Therefore, the Pellucid system is designed as a framework which needs some customization efforts to be applied to a particular domain.

The architecture of the agent-based Pellucid platform is presented in Figure 1. Its functionality is divided into three layers: an interaction layer, a process layer and a search-and-access layer. The agents of all these layers cooperate with an Organization Memory (OM) [2, 3], which is responsible for storing experience. Knowledge and experience are represented with ontologies [4–6] kept in the OM. An organization's generic ontology contains a description of concepts and relations common to all organizations and necessary to describe structure, assets, activity, as well as the knowledge and experience of the organization. For these purposes it includes descriptions of workflows, business contacts, formal documents and organizational repositories. According to the needs of a particular organization, the generic ontology is extended with a domain-specific part. This part is created in the phase of platform customization by introducing specific features used in particular domains of the organization's activity. Pellucid ontologies are disscussed in more detail in [7].

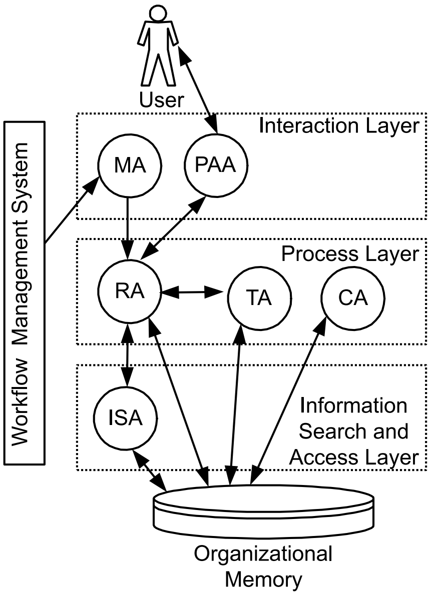

Figure 1. The Pellucid platform architecture

For experience management purposes, Pellucid obtains an employee's work context from external modules, a Workflow Tracking System (WfTS) or a Workflow Management System (WfMS) [8].

3.3. Experience management

The Pellucid framework also implements the main ideas of experience management. The experience management model exploits experience sharing concepts discussed in [9] and comprises three phases: Capture and Store, Analysis and Presentation, and Experience Evolution.

The Capture-and-Store phase is concerned with observing and storing experience in a particular context. There are three ways of capturing experience: analysing employees' actions and workflow events, analysing documents entered into the system, and by direct input from employees. Capturing experience from working actions and events is particularly beneficial for repetitive tasks; the experience is used to create common patterns that can be retrieved in the future in order to assist other employees. The direct input of experience is carried out through free-text notes written by the employees themselves.

The purpose of Pellucid is to support employees by providing them spontaneously with the experience required by the activity they are just performing. The Analysis-and-Presentation phase accomplishes that task with the help of the concept of an active hint, defined as a representation of experience within the organization. An active hint is triggered in a context and includes an action, a knowledge resource and a justification for the hint. The context is determined by the particular activity being carried out by an employee according to the workflow system. The action corresponds to an atomic act on a knowledge resource, such as 'use' a document template, 'read' a document or a note, or 'consider' a contact list. The justification gives a reason for the hint to the employee.

The aim of Experience Evolution is updating the available experience. Due to the changing environment, experience has a limited lifetime. Out-of-date or invalid experience has to be identified and updated or removed.

4. The Search-and-Access Layer

The Search-and-Access Layer is intended to be an intermediary between the agents of the upper layers of the Pellucid platform and external organizational document repositories. The functionality of this is accomplished by an ISA agent (*cf.* Section 4.1). One of the most important agent functionalities is to allow access to and searching for organizational information. Searching is performed on demand and in cooperation with other layers. This functionality is designed to satisfy miscellaneous requirements of flexible user assistance.

To operate and serve external demands the ISA agent is able to search for information located in documents within organizational repositories. The ISA agent needs to know the way to access the repository and the documents, as well as the structure of repositories, documents and information in order to consider advanced search strategies within the document. The method of representing that knowledge, for the purposes of the Pellucid platform, are ontologies, which are universal knowledge representations used by layer components and other agents.

SAL uses document ontology, repository ontology, information ontology and domain-specific ontologies. A document ontology defines the types of documents or information resources and its relationship with the repositories and information. A repository ontology describes the document repositories, the way of accessing them and their structure. An information ontology describes how to retrieve information contained in documents.

4.1. The information Search and Access agent

As mentioned before, an ISA agent is the main component of SAL. It has been designed to support other agents with the information stored in organizational documents (as shown in Figure 2). Its main tasks are providing searching abilities, management of document repositories and cooperation with other components of the platform.

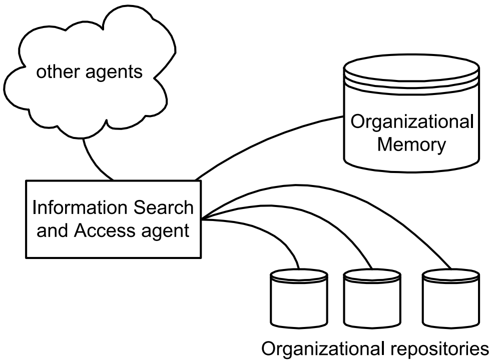
other agents
Organizational Memory
Information Search and Access agent
Organizational repositories

Figure 2. The information Search and Access agent

SAL functionality is mostly event-driven and concerns handling of events occurring in organizational repositories. The events represent changes of the repositories' content, such as creation, modification and removal of documents. The occurrence of an event entails a sequence of actions to be performed to handle it. All repositories are monitored for event occurrences. When an event is detected appropriately formatted information is sent to the component responsible for the events' handling. The purpose of this component is to order an update of index files stored in the system, which often requires re-running of the indexing process. Then, semantic annotation is performed (see Section 4.4), which updates the semantic description of the document. The final action is to store the ontology instance for the event in the OM.

The ISA agent handles the whole document life-cycle. When a new document is created it is indexed and analysed in terms of domain-specific ontology. Consequently, the OM is updated with domain-specific descriptions of the document. When the document is modified, its description is updated, the indices are replaced and ontological instances concerning modifications are added. The removal of a document entails removal of its indices and recording this information in the OM.

The functionality of the ISA agent includes the following:

- repository monitoring – the agent is able to communicate with the repository, translate such native communication to the ontology-based language and store it in the OM;
- access to different document formats – the agent is able to understand several document formats and use their content for the purpose of future searching;
- full-text indexing – the agent is responsible for building the indices;
- semantic annotation – the process of document analysis to enrich its semantic description;

- searching abilities – the agent is able to perform an advanced search based on ontology and full-text indices;
- cooperation with other internal components of the Pellucid platform (Role Agent, Organizational Memory, Monitoring Agent).

The main methods of information searching are discussed below.

4.2. Full-text indexing

Full-text indexing functionality is build into the system to perform statistical, word-based searching and finding similar documents. The core functionality of the component responsible for full-text indexing is as follows:

- managing the document index, including addition of documents to the index and their removal;
- assuring coherence between the repository and the index content to avoid discrepancy due to addition, removal or modification of documents;
- finding documents in the index by means of text queries.

The component responsible for full-text indexing should provide indexing in a few languages, obtaining the theme of a given word, removal of insignificant words, and sorting of search results by relevance.

The component responsible for full-text indexing creates and manages a forward index. A forward index (as opposed to an inverted index) allows one to extract the most significant words from a document. Such data structure is used to search for similar documents. Words extracted from a document form a query directed to the forward index. Each document from the resulting set is considered similar. Creating and updating indices requires accessing the repository where documents are stored and accessing the content of documents in the plain-text form.

Most often, the plain-text indexing functionality is provided by the Lucene searching and indexing engine [10], incorporated into the Pellucid system.

4.3. Ontology-based searching

The most important part of the ISA agent functionality is searching for documents. Two main types of ontological searches are distinguished:

1. searching for documents containing a specific ontology class – concept (*i.e.* a concept which is an element of the domain-specific ontology);
2. searching for documents containing a specific individual (*i.e.* an instance of a concept from the domain-specific ontology).

An integrated search is possible with the use of all available methods, also based on the full-text index.

The algorithm of searching for documents by a specific concept is as follows:

1. extract the concepts to be searched for from the user query;
2. build an RDQL query [11] to search for documents containing the concept;
3. execute a query on the OM;
4. send the response to the appropriate agent.

Similarly, the algorithm of searching for documents by a specific individual consists of the following steps:

1. extract the ontology individual from the OM using the user query;
2. build an RDQL query to search for documents containing the individual;
3. execute a query on the OM;
4. send the response to the appropriate agent.

4.4. Semantic annotation

The idea of semantic annotation stems from the need to make the content of documents meaningful for the system. The easiest way to process the content of documents is to employ indexing techniques. However, an indexing technique offers a statistical, text-based description only. On the basis of the ontology, the ISA agent builds a fragment of a semantic description. Semantic annotation has originated from the observation that some elements of document description or content can be connected with some ontological concepts.

The way of document annotation by domain concepts is presented in Figure 3. A part of the information ontology shows the exploited relation between a domain concept and a document content. A domain concept can have a representation assigned (in a form of textual representation), for example a regular expression. The ways of using semantic annotation are generally directly dependent on the document structure in the organization. Semantic annotation is not only applied to document content; it can be used for finding an ontological description of documents gathered in the OM according to their properties. The document description obtained by semantic annotation is stored in the OM and used for further information searching, such as document similarity.

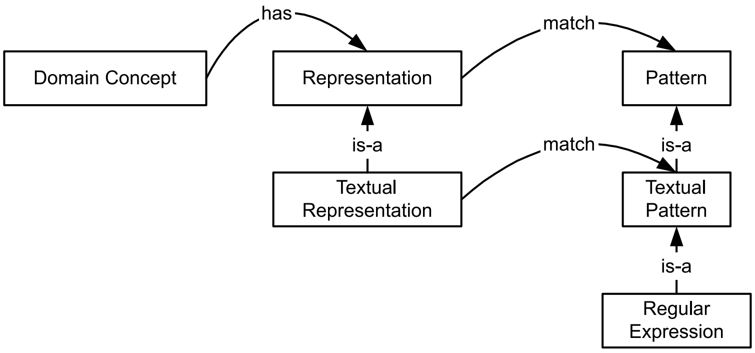

Figure 3. The idea of semantic annotation

In the case of the CdG pilot site, document similarity is used for finding ontological descriptions of documents gathered in the organizational repository according to their file names. The characteristic feature is that some documents, which are required by standard workflow activities, are given names, which encode some information about them. For example, a Timing Cycle document has a standard name described by the pattern presented in Table 1. In this example, the Timing Cycle document P103Mas.wk4 concerns zone Ponente (P), Albaro district (1), Massimo street (Mas) and is created in the 03 dossier. The Pellucid system can constructs a semantic description of the document based on the file name.

Table 1. Semantics of Timing Cycle document name

P	1	03	Mas	.wk4
Zone name ID	District name ID	Dossier ID	Street name ID	
1 or 2 letters	1 digit	2 digits	some letters	extension

4.5. Repository monitoring

The process of repository monitoring requires handling events occurring in the repository such as creation, modification and removal of documents. There are different types of monitors required for different types of repositories. Separate monitors handle particular types of repositories such as file systems (*e.g.* Windows 2000), mail servers or clients (*e.g.* Exchange 4.0, Outlook XP), databases (*e.g.* Oracle Database Server), *etc.*

Let us discuss a file system repository as an example. It is monitored for changes in the content of a given directory tree. The task of the monitor is responding to events. The component responsible for repository monitoring tracks repository events and sends messages about events to another component, responsible for further event handling. The messages contain an events' description defined according to the ontology. The repository monitoring functionality should allow easy maintenance, scalability and customization of the Pellucid platform to the different types of repositories owned by an organization.

The main objective of repository monitoring is to inform about events occurring within a repository. According to the type of the repository, various types of events may happen to repository elements and thus to documents (*i.e.* a different set of events will be assigned to emails and to textual documents stored as files). Components responsible for repository monitoring communicate with repositories and translate all information associated with the repositories' status to a unified format acceptable to the ISA agent.

4.6. Notification

The notification functionality is realized as a part of handling document events and collecting information from organizational repositories. The ISA agent stores information within the OM and informs the Monitoring Agent about the occurrence of new information. The process layer agents analyse the information according to guidelines concerning event notifications. If the information needs to be shown to the user, the appropriate active hint is fired and the user is notified.

4.7. Document access

Document access is strongly connected with the indexing process. The indexing process needs to access the plain-text content of any particular types of documents to perform its tasks. The Pellucid platform allows handling any document formats by dedicated modules called document drivers. A document driver interacts with the external boundary of the Pellucid system. It is developed as a part of the customization activity to deploy within a particular organization. Appropriate drivers are required for each native format of documents used in the organization.

5. Summary

Methods of searching for organizational information have been described in the paper and details of their practical usage in the Pellucid project have been shown. The methods take into account the characteristic structural features of public organizations' information and are consistent with the characteristics of mobile employees, users of the Pellucid platform. The documents gathered in organizational repositories are searched with the help of index-based and ontology-based methods. The usage of ontology allows one to extend the searching process by abilities resulting from domain-specific concepts. Moreover, semantic description of documents can be achieved for the purposes of document similarity. The proposed methods allow one to exploit thoroughly the domain-specific aspects in the searching process.

In 2003, a pre-evaluation was performed of the system's prototype. The end users assessed the functionality of the system as a whole, with special regard to its experience management features. The pre-evaluation resulted in further work focused on the free text notes for experience representation. At the moment of writing of this paper the first version of the Pellucid system has been released. A full user evaluation is performed at three Pellucid pilot sites. The most important objective of the evaluation is to assess the usefulness, completeness and accuracy of experience-based hints (*i.e.* Active Hints and free-text notes). The quality and performance of the Pellucid system are tested. The results of this evaluation will form the basis for further improvements and enhancements of the platform. The project is to be completed in October 2004.

Further information about the Pellucid project can be found at the official website [1]. Information concerning the Polish Pellucid Dissemination Group is also available [12]. The various aspects of design and implementation of the SAL layer have been described in papers [3, 6, 7, 13–15] during the project's development.

Acknowledgements

Thanks are due to the European Commission and participants of the Pellucid project, IST-2001-34519. An AGH-UST grant is also gratefully acknowledged.

References

[1] http://www.sadiel.es/Europa/pellucid/

[2] Abecker A, Bernardi A, Hinkelmann K, Kuhn O and Sintek M 1998 *IEEE Intelligent Systems* **13** (3) 40

[3] Kitowski J, Lambert S, Słota R, Krawczyk K and Dziewierz M 2002 *Proc. of PIONIER – Polish Optical Internet*, Poznan, Poland, pp. 221–233

[4] Uschold M and Gruninger M 1996 *Knowledge Engineering Review* **11** (2) 93

[5] Staab S, Studer R, Schnurr H-P and Sure Y 2001 *IEEE Intelligent Systems* **16** (1) 26

[6] Słota R, Majewska M, Dziewierz M, Krawczyk K, Laclavik M, Balogh Z, Hluchy L, Kitowski J and Lambert S 2004 *Lect. Notes in Comput. Sci.*, **3019** 700, Springer Verlag

[7] Kitowski J, Krawczyk K, Majewska M, Dziewierz M, Słota R, Lambert S, Miles A, Arenas A, Hluchy L, Balogh Z, Laclavik M, Delaitre S, Viano G, Stringa S and Ferrentino P 2004 *Lect. Notes in Comput. Sci. LNAI*, **3035** 75, Springer Verlag

[8] DiCaterino A, Larsen K, Tang M-H and Wang W-L 1997 *An Introduction to Workflow Management Systems*, Center for Technology in Government,
http://www.ctg.albany.edu/publications/reports/workflow_mgmt

[9] Bergmann R 2002 *Experience Management: Foundations, Development Methodology, and Internet-based Applications*, LNAI **2432**, Springer Verlag

[10] The Project Jakarta Lucene Homepage, http://jakarta.apache.org/lucene

[11] W3C Member Submission RDQL-RDF Data Query Language, http://www.w3.org/Submission/2004/SUBM-RDQL-20040109

[12] Local Website of the Pellucid Project, http://www.cyf-kr.edu.pl/5PR/pellucid.htm

[13] Słota R, Krawczyk K, Dziewierz M, Majewska M, Kitowski J and Lambert S 2003 *Proc. of PIONIER – Polish Optical Internet*, Poznan, Poland, pp. 167–177 (in Polish)

[14] Lambert S, Stringa S, Viano G, Kitowski J, Słota R, Krawczyk K, Dziewierz M, Delaitre S, Oroz M B, Gomez A C, Hluchy L, Balogh Z, Laclavik M, Caparros S F, Fassone M and Contursi V 2003 *Lect. Notes in Comput. Sci. LNAI*, **2645** 203, Springer Verlag

[15] Krawczyk K, Majewska M, Dziewierz M, Słota R, Balogh Z, Kitowski J and Lambert S 2004 *Lect. Notes in Comput. Sci.*, **3038** 601, Springer Verlag