AUTOMATYKA • 2009 • Tom 13 • Zeszyt 2

Antoni Ligęza*, Sebastian Ernst*, Grzegorz J. Nalepa*, Marcin Szpyrka*

A Conceptual Model for Web Knowledge Acquisition System with GIS Component**

1. Introduction

Internet has become a single most important resource for instant information sharing. It can also serve as efficient means for data and knowledge acquisition. From the point of view of ordinary users it can be considered to constitute a very flexible and powerful version of the concept of the so-called *blackboard architecture*. The rapid growth of the Internet prompted a rapid development of specific software. This software addresses different aspects of information organization and interchange related issues, such as storing, retrieval, searching, indexing, aggregating, sharing, updating, etc.

Building efficient tools for supporting Knowledge Acquisition and further Knowledge Management is a challenge and hot research area with potentially infinite numbers of practical applications. In modern computer science the web technologies open a completely new chances for massive, distributed knowledge acquisition. Examples of such social phenomena as *Wikipedia* constitute a working proof of high potentials incorporated in the synergy of human and web interaction. Increasing functionalities of web applications and almost unlimited computational power of modern hardware seems to promise that solving this problem is a matter of time.

This paper discusses certain issues concerning the conceptual model for a distributed knowledge acquisition system gathering and organizing knowledge concerning threats of various nature and aimed at improving safety of citizens in urban environments. The system for registering citizen-provided information is a part of the INDECT FP7 Project. Contemporary tools and techniques to be applied, including GIS technologies are presented in brief and future problems to be solve are identified.

The main focus of our project is on tools to process the information provided by citizens via a specialized website. In fact, a Web System software for citizen provided informa-

* Department of Automatics, AGH University of Science and Technology, Krakow

** This work has been partially performed in the framework of the EU ICT Project INDECT (FP7-218086).

tion, automatic knowledge extraction, knowledge management and GIS integration is to be developed. This task is intended to complement other work oriented towards automated information extraction from existing web resources by building an Internet-based, distributed information acquisition and automated knowledge management system. It will combine a CMS (Content Management System) system with a KMS (Knowledge Management System) incorporating intelligent information processing tools based on knowledge engineering. It will allow storing and retrieving of partially analyzed investigations. For spatial information processing and graphical information presentation a GIS (Geographical Information Systems) technology will be used and rule-based systems technology for automated inference will be incorporated (see the HEKATE Project http://hekate.ia.agh.edu.pl).

The core of any larger information management system is a Database Systems, and more specifically a Relational Database Management Systems [RDBMS]. Thanks to relatively simple internal model of data and development of indexing techniques they also assure good scalability. On the other hand, to provide full necessary functionalities, RDBMS must be cooperating with other technologies and systems; below we list some most promising ones for the INDECT application.

Today web technologies constitute an advanced and universal programming framework. This framework has a very heterogeneous structure including: data encoding and structuring languages (such as XHTML and XML), meta-data languages (such as RDF), data transformation languages (such as XSL/T/FO), data presentation languages (such as CSS), server-side programming languages (such as PHP, JSP), and client-side programming languages (such as Javascript). Moreover, the web technologies of CMS evolve towards very promising direction of *Semantic Wikis*.

Another very promising technology is the one of Geographical Information Systems (GIS). Such systems allow to incorporate spatial knowledge representation and spatial knowledge management. They usually contain a multi-level vector and raster 2D (2.5, 3D) data allowing for practical visualization of spatial phenomena.

This paper presents shortly the core ideas of conceptual models of the system for acquisition and management of citizen provided information and provides some overview of selected technologies and tools critical for it. These are mainly GIS and *Semantic Wikis* having in mind potential application for the INDECT Project under development. Below a conceptual model of the system is presented. The overview has been conducted in order to identify main issues related to knowledge acquisition, representation, and management with respect to the project.

2. A conceptual model for the citizen provided information acquisition system

The Citizen Provided Information Acquisition System is a multi-component system of extended functionality, including spatial information acquisition and management. It is based on the assumption that in fact *not all the knowledge can be found over Internet*

(recent, historic, particular, domain-specific, area-specific). On the other hand, *people are willing to cooperate* (see the success of Wikipedia and other collaborative projects). They only require that it is simple and not time-consuming (instant action), they can see their contribution and the impact of it (other people can evaluate and confirm it), and in general their effort must lead to useful results. Further, the knowledge must be processed automatically (cost elimination), and graphical interface (GUI and GIS) is necessary,

The main goal of the system is to serve as a distributed knowledge acquisition system for data, information and knowledge provided by citizens, as well as to enable limited automated knowledge management. In principle, the working scenario for the systems is as follows.

The system offers a web interface (based on thin-client technologies; in practice a standard web viewer is). The interface offers various functionalities for different types of users. The main functionality refers to enabling definition of a new threat (in the form of filling a set of web forms) and positioning the threat on a map (using a GIS interface). The provided knowledge is then checked for syntactic correctness and processed in an automatic way. The ultimate result is stored in the internal knowledge base.

The main functionalities offered by the system include the following ones: (i) Guided definition of the threats, (ii) Relational entry type with predefined categories ('click/select'; rather than 'write'), (iii) GIS-type entry and visualization, (iv) Data and knowledge refinement, (v) Automated data and knowledge aggregation, (vi) Automated knowledge management, (vii) User evaluation of data; evaluation of users, (viii) Data and knowledge evaluation and verification (truth-value, consistency, informative value), (ix) Multi-level signaling of emergency, (x) Automated generation of periodic reports, (xi) Tracing of changes (moves, modifications, reincarnation,...), (xii) Selective access: combined relational and spatial search; area restriction,

The key issue for automated knowledge classification is development of a taxonomy of threats. The location-specific threats are further classified into the following categories of phenomena related to: Air, Animals, Biology, Chemicals, Crime, Environment Pollution, Explosives, Fire, Flooding, Infrastructure (Bridges, Buildings, Installations (Electric, Gas, Waste-Treatment), Parking, Parks, Streets) Military-related, Plants, Traffic-related (with further subcategories: Air, Off-road, Railway, Road, Water). There is also separated taxonomy of general threats (not related to a specific localization).

The overall architecture of the systems is presented in Figure 1. The system is composed of:

- three input channels (textual, including attribute-values for the database, GIS, for spatial information and one for binary objects, such photos),
- three output channels (for ordinary users, for analytical services, and presenting emergency reports),
- internal knowledge refinement, classification and aggregation scheme,
- internal database,
- knowledge and user evaluation system.

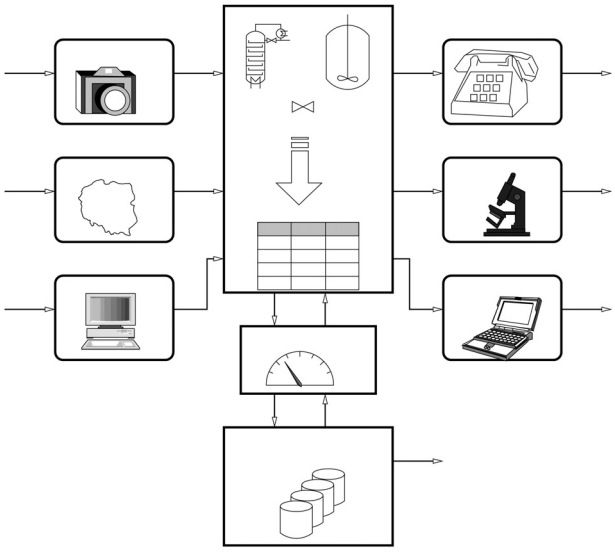

Fig. 1. A schematic outline of the system structure and components

The most interesting components related to novel web and data/knowledge processing technologies are the GIS component for spatial localization and presentation of threats and the sementic wiki component for semantic knowledge acquisition and processing. Below we analyze some critical aspects and state-of-the-art of these two technologies

3. GIS technology and tools

Geographical Information Systems (GIS) is a wide term encompassing various systems aimed at capturing, storing, analysis, management and presentation of geospatial data. The term itself is very wide and includes solutions for different applications, from touristic mapping to strategic military and emergency planning solutions.

3.1. Geospatial data management

Geospatial, or *georeferenced* data refers to data which includes a *spatial* component, describing the location of spatial distribution of geographic phenomena, as well as an *attribute* component describing its properties [Neteler]. The two basic data models used for spatial data are:

- the *raster data model*, where a given area is split into equally-sized *fields* or *pixels*, and each pixel is given a value (a number, a set of numbers or null),
- the *vector data model*, where objects are described using their location properties and represented as geometric primitives, such as points, lines and polygons.

Raster data, which is in fact a matrix of values, is useful for storing features such as elevation, temperature or measured humidity. Also, raster data containing satellite imagery

is often used for visualization purposes. Vector data, on the other hand, uses geometric shapes to represent points, lines and polygons. The Open Geospatial Consortium, in its GML (Geography Markup Language) specification, defines a set of classes for representing geographic (spatial) features,

Raster and vector data forms layers, composing a GIS dataset. For instance, a raster elevation layer can be supplemented with a vector layer containing roads and a binary raster layer representing urban areas. Alternatively, the urban areas could be represented within a vector layer, as a polygon.

3.2. Overview of GIS solutions

Applications of very different characteristics fall into the category of GIS software. Institutions often choose an integrated GIS solution, such as the open-source GRASS GIS [Neteler]. Such solutions integrate both raster and vector data and support various parts of the GIS workflow, such as map digitization, 2D and 3D analysis and visualization.

However, this approach does not fit the needs of the web knowledge acquisition project very well. Furthermore, the integrated architecture was not really created to serve as a backend for a web application, and would require a lot of "layer glueing" to serve that purpose. On the other hand, the advanced analysis tools available in GRASS are not necessary for a public web application.

Therefore, the preferred approach seems to develop a lightweight, modular architecture and use individual GIS components designed precisely to provide the necessary functionalities, while limiting the system growth and unnecessary overhead.

Relational databases are well-suited to storing and indexing many types of data, but lack functionality when it comes to processing spatial data. Even though it may be possible to represent 2D or 3D coordinates using simple data types such as *integer* or *float*, it would be neither convenient nor effective. Lack of dedicated indexes would mean long query processing, and queries selecting objects located within a given circle would be all but straightforward.

On the other hand, relational databases have many properties which are favorable from the project's point of view. Therefore, these database engines need to be supplemented with spatial functionality. This includes:

spatial data types – able to store geometries of records in database tables,
spatial operations – to combine, process and analyze these geometries; for instance, functions to compute an area of the intersection of two polygons,
spatial indexes – for efficient processing of spatial or georeferenced data.

PostGIS a spatial extension for PostgreSQL, is an advanced toolkit providing geometry types according to the OGC data model (see Fig.), spatial predicates for detecting geometry interaction, spatial operators for computing the area, distance and length, geometry modification functions, as well as R-tree-based spatial indexes.

Spatial objects are usually created using the *GeometryFromText* function, which takes a WKT geometry as a parameter. As PostGIS is a separate program, geometry columns are added to existing tables using the *AddGeometryColumn* function.

The crosses function is a so-called geometry relationship function. Other spatial relationship functions supported by PostGIS include [PostGIS]: distance, dwithin – check if distance is within a given range, equals, disjoint, intersects, touches and contains.

3.3. GIS and the web

The evolution of web techniques and technologies has resulted in arrival of many location-based web services. This section describes two most significant groups – solutions used for easy visualization of georeferenced data within a web browser and a new trend, providing a location-based "virtual reality" services.

Early attempts to bring mapping software – usually concerned with vehicle route planning – to the web often involved usage of proprietary technologies. This involved development of heavyweight Java applets which, while providing a rich user interface, imposed additional requirements on the user's machine. This was not a problem for ordinary desktop users, but few mobile devices were able to support Java applets at all. Some modern services, such as Map24, still use the Java technology and provide advanced features such as 3D visualization. Other services (i.e., Zumi) prefer to use the Adobe Flash technology to provide data visualization for web applications. This technology is also rather demanding in terms of processing power and its application for mobile devices is very limited.

In contrast to the two aforementioned technologies, it seems that the future belongs to lightweight DHTML applications which, through extensive use of JavaScript and CSS for visualization and user interaction and AJAX for data exchange, provide an attractive and convenient mapping interface within any moder web browser.

There are many examples of such applications within the Web, with the most notable including Google Maps, Yahoo! Maps, Microsoft Live Search Maps and OpenStreetMap itself. They are all based on a similar principle: GIS data is rendered by the server into image tiles for different zoom levels. A JavaScript-equipped DHTML element downloads these tiles using the HTTP protocol and arranges them using data retrieved using AJAX. Often, it is possible to choose from among various tile types, including road maps, physical maps and satellite imagery. Many of these services provide APIs, allowing the use of maps in other web applications and placement of objects on top of the map tiles.

4. Semantic wikis

A wiki system is a community-driven web-based collaboration tool. It allows users to build content in the form of the so-called wiki pages, as well as uploaded media files. Wikipages are plain text documents containing special wiki markup (wikitext), e.g. for structuring content.

An important feature of wikis is the integrated version control functionality, crucial in a collaborative environment. It allows registering all subsequent versions of every page, thus allowing to see introduced content differences. All wiki edits may be identified by user names and time stamps, so it is possible to recreate any previous state of the wiki at any given time.

From the technical point of view a wiki has a regular web-based client-server architecture. It is run on the webserver and accessed by a regular browser. On the server side wikis require different runtime environment (e.g. PHP), possibly with a relational database system (RDBMS). A comprehensive comparison of different wiki systems can be found on http://www.wikimatrix.org.

An important step in the direction of enriching standard wikis with the semantic information has been provided by the introduction of the so-called *semantic wikis*, such as the IkeWiki, OntoWiki, SemanticMediaWiki, or SweetWiki [GJN]. In such systems the standard wikitext is extended with the semantic annotations. These include relations (represented as RDF triples) and categories (here RDFS is needed). It is possible to query the semantic knowledge, thus providing dynamic wiki pages. Ultimately these extension can also allow for building an ontology of the domain with which the content of the wiki is related. This extension introduces not just new content engineering possibilities, but also semantic search and analysis of the content. However, from the knowledge engineering point of view expressing semantics is not enough. In fact a knowledge-based system should provide effective knowledge representation and processing methods.

Several semantic wiki systems are available, most of them in the development stage providing demo versions, see http://semanticweb.org/wiki Semantic_Wiki_State_Of_The _Art. A recent FP7 project Kiwi (http://www.kiwi-project.eu) aims at providing a collaborative knowledge management based on semantic wikis (it is the continuation of IkeWiki effort).

In the context of the INDECT GIS system, semantic wikis seem to be a perfect tool for supporting the distributed knowledge acquisition process from the citizens, collaborating with the police on building a safer environment. Wikis provide a familiar and very accessible web interface, suitable for both regular and mobile users. Semantic annotations would be related to the threats ontology. The hipertext nature of the system would allow for easy linking the contents of the system with external sources providing supplemental information.

5. Concluding remarks

UE FP7 INDECT Project named *Intelligent information system supporting observation, searching and detection for security of citizens in urban environment* is expected to provide tools for increasing safety of citizens. One part of the project is considered in the paper system for registering citizen-provided information of various kinds (text, photo, etc.) via a specialized website. It seems that contemporary web tools and techniques are good

enough to be treated as the start point for the system development. Examples of such technologies are GIS and Semantic Wikis. The conceptual model of the system and a survey of GIS technologies and Semantic Wikis have been presented in the paper.

The system for web knowledge acquisition is expected to manage the gathered knowledge automatically, i.e. automated information extraction, knowledge management and GIS integration is to be developed. Some problems concerning GIS integration has been also pointed out in the paper. Furthermore, the state-of-the-art of contemporary Semantic Wikis technologies has been presented too. The hipertext nature of the systems and possibility of building an ontology of the domain with which the content of the wiki is related make the tool one of most important components of the system.

References

[1] [Neteler] Neteler M., Mitasova H., *Open Source GIS: A GRASS GIS Approach, Second Edition*. Kluwer Academic Publishers, Springer, Boston 2004.
[2] [NASA-WGS84] *The EGM96 Geoid Undulation with Respect to the WGS84 Ellipsoid.* NASA, http://cddis.nasa.gov/926/egm96/doc/S11.HTML.
[3] [GML-SPEC] *OpenGIS® Geography Markup Language (GML) Implementation Specification, version 2.1.2*, Open Geospatial Consortium, http://www.opengeospatial.org/standards/gml.
[4] [Wikipedia] *Wikipedia, the Free Encyclopedia*, http://en.wikipedia.org.
[5] [Mapcenter] *MapCenter2*, http://mapcenter2.cpsmapper.com.
[6] [OpenStreetMap] *OpenStreetMap*, http://www.openstreetmap.org.
[7] [MySQL-GIS] Karlsson A., *GIS and Spatial Extensions with MySQL*, MySQL DevZone, http://dev.mysql.com/tech-resources/articles/4.1/gis-with-mysql.html.
[8] [PostGIS] *PostGIS 1.3 Documentation*, http://postgis.refractions.net/documentation.
[9] [GPSBabel] *GPSBabel Documentation*, http://www.gpsbabel.org.
[10] [OSM-Wiki] *OpenStreetMap Wiki*, http://wiki.openstreetmap.org.
[11] [Map24] *Map24*, http://www.map24.com.
[12] [GPSvis] *GPS Visualizer*, http://www.gpsvisualizer.com.
[13] [Norc] *Norc*, http://www.norc.pl.
[14] [RDBMS] T. Connolly, C. Begg: *Systemy baz danych. Praktyczne metody projektowania, implementacji i zarządzania*. Wydawnictwo RM, Warszawa 2004, t. 1 i 2.
[15] [INDECT] Intelligent information system supporting observation, searching and detection for security of citizens in urban environment. FP7, 2009–2013, http://www.indect-project.eu/.
[16] [GJN] Krötzsch M., Vrandecic D., Völkel M., Haller H., Studer R.: *Semantic wikipedia*. Web Semantics 5, 2007, 251–261.
[17] [GJN] Schaert, S.: Ikewiki: *A semantic wiki for collaborative knowledge management*. In: WETICE ’06: Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, Washington, DC, USA, IEEE Computer Society,2006, 388–396.
[18] [GJN] Buffa M., Gandon F., Ereteo G., Sander P., Faron C.: *Sweetwiki: A semantic wiki.* Web Semantics: Science, Services and Agents on the World Wide Web, 2008, in press.
[19] [GJN] Baumeister J., Reutelshoefer J., Puppe F.: *Knowwe: community-based knowledge capture with knowledge wikis*. In: K-CAP ’07: Proceedings of the 4th international conference on Knowledge capture, New York, NY, USA, ACM, 2007, 189–190.