Targeted Information Retrieval in the Web Space as a Response to User Needs

Kristina Machova
Faculty of Electrical Engineering and Informatics
Technical University of Kosice
Letná 9, 042 00 Košice, Slovakia
Kristina.Machova@tuke.sk

Abstract. *Paper introduces methods for targeted information retrieval in the web space as a response to user needs using semantic technologies, web services and GPS. It focuses on targeted information of various types like, for example, contact information for organizations of a given category, their location on the map using GPS navigation, etc.*

Keywords. Information retrieval, Web space, Semantic agent, Targeted information, GPS

1 Introduction

One of the most known imaginary libraries is the Babylon library from a fiction by J. L. Borges. The Babylon library is infinitive. It contains all knowledge of the world in an infinite number of hexagonal galleries but its content is not ordered and therefore there are no access methods available. We are lords of the treasure of all books ever written, but do not know where is a particular book, we are interested in. The Internet contains a huge number of information. Some of information chunks are very useful but some are not. Similarly to the Babylon library, these information pieces are useless if we do not know where they can be found. From this point of view, the importance of information retrieval is essential [6].

The WWW (Word Wide Web) is nowadays the most quickly growing service on the Internet. The Web content can be divided into two groups: surface web and deep web [3]. The surface web consists of static and dynamic publicly accessible web pages, which contains approximately 2.5 milliards of documents. The hidden web contains special databases and dynamic web pages accessible via the Web, which are not often used by common web users - surfers. The information available in the hidden - deep web is approximately 400 or 550 times larger than information located within the surface layer of the web. So the ability of information retrieval also in the deep web is quite important.

Many web search engines are known. The first one was Wandex created in 1993 for employees of MIT. It was subsequently followed by systems like Aliweb, Yahoo, Lycos, Excite and in 2001 by Google, which changed results and relevancy of information retrieval. Google has introduced sorting of web links according to PageRank [2]. The PageRank is a quite simple criterion taking into account how many pages an actual page has links to and how many pages have links to the actual page. But results retrieved and provided by a search engine can be manipulated employing SEO (Search Engine Optimization) techniques [7]. For example, the page is fulfilled by a great number of key words, which make the page nearly unreadable but increase its rank. Or interesting key words are collocated into meta-tags. The Page Rank measure can be increased with the aid of hidden links, mirroring of pages, cloaking (robot is provided with a different page than user browser) or link farming (many pages have links to each other).

Nowadays the Web has many problems according to [1], for instance: high recall connected with low precision, low or even zero recall, search results being web pages and not searched information, and search results dependent on a used vocabulary. The most popular search engines try to solve the problem with low precision of search results using page ranking and subsequent ordering the retrieved pages according to this rank. This solution supposes that web user is willing to read only some of the first found pages and so it is important to locate the more precise web pages on the top of the resulting page list. The second problem with low or even zero recall requires the refinement of web searching approaches and understanding user needs. A solution to this problem can be represented by semantic search [4]. The semantic search can be an appropriate tool for making search results independent from used vocabulary. Mainly such semantic technologies as metadata (XML, RDF documents) and ontologies (OWL documents) can be used for these purposes. An ontology can represent a vocabulary of all words used in tags of RDF documents. Understanding of user needs can be met in the frame of the semantic web. The semantic web idea contemplate the existence of intelligent software semantic agents, which are able to search for relevant information as the response on complicated web user demands in relatively short time intervals. The goal of these agents is also searching for relations between known knowledge pieces.

Our work introduces an implementation of a system for retrieving information about firms actually living on the Web (i.e. their web presentations are modified continuously). It focuses on such kinds of information like firm name, address, telephone number, web presentation link and GPS position. We used services like: StrikeIron in English, GoldPages in Slovak, web service for obtaining document WSDL, and GPS service.

2 Used techniques

2.1 Metadata

The abortion of the Babylon tower building completion was caused by the disability to make people understood one another. A bafflement of languages started when the tower started to achieve the sky. Some prophets claim that in the case of computers the bafflement of languages started at the very beginning. Nowadays, the ambition is not only the ability of communication between creators of computers but also between computers themselves.

Maybe it will be possible with the aid of metadata, for example XML (Extensible Markup Language). But this language is unable to enrich information with metadata about the semantics of this information. It allows only better processing of the information pieces by computer programs. The XML uses tags for explanation what some parts of information mean or what data type this information has. It enables for computers to understand this information. Tags or marks are identifiers. In our work they serve to identify information pieces like firm, title, street, town, telephone number, and GPS position. These marks are illustrated in following code:

```
<?xml version="1.0"encoding="UTF-8">
<firms>
  <firm idf="1524">
    <title>PIZZERIA PALERMO</title>
    <street>Masarykova 4</street>
    <town>Kosice</town>
    <telephone>055 6233333</telephone>
  </firm>
  ...
  <firm idf="5870">
    <title>ALA ALADIN - PIZZA, GYROS</title>
    <street>Hlavna 40</street>
    <town>Kosice</town>
    <telephone>055 6230823</telephone>
  </firm>
</firms>
```

The tree structure of the firm file is illustrated in Fig. 1.

Our work focuses on software semantic agents [9]. Such agent enables to search for targeted information within web pages and create a pattern of firm file like one illustrated in Fig. 1. A similar approach to information processing with the aid of semantic technologies is introduced in [8]. Semantically enriched web information processing can enable information understanding by computers.

2.2 Web services

A web service is a modular application, which is installed somewhere on the Web. Users can access this service and use it anywhere. Web services are based on standards of W3C world organization (`http://www.w3.org`). This technology

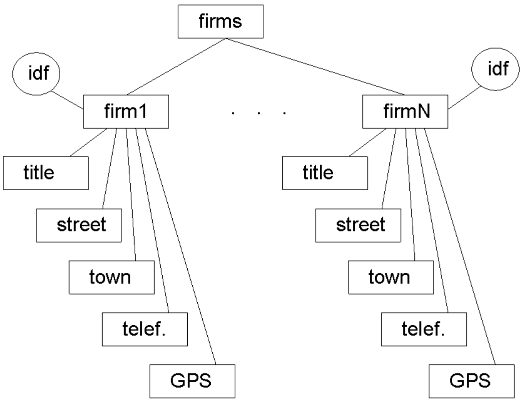


Figure 1: Tree-like structure of the firm file

becomes on the top of a popularity scale. Microsoft has established on this technology its NET architecture with contribution for development and management of web services. Basic architecture of web services according to [5] assumes interaction of three roles:

- *Service Providers* provide web services to clients;
- *Service Registries* enable registration, searching and identification of services;
- *Service Requester* searches for web services and uses them.

Web services (WS) provide basic information about themselves. They can be described in WDSL (Web Service Definition Language) documents. Web services are able to send and accept XML documents using SOAP (Simple Object Access Protocol) over application protocols of the Internet like HTTP, SMTP and so on. Web services can be searched by specialized mediators UDDI (Universal Description Discovery and Integration). In the presented work, the web services for obtaining WSDL document was used and web services "StrikeIron" and "GoldPages" as well.

The retrieved information pieces about firms and organizations are classified into the following categories:

- Car
- Flat, house, garden
- Offices, shops
- Communication and computing technologies
- Culture, education
- Institutions and organizations
- Industry and transport
- Agrarian sector and forests
- Restaurants
- Services for firms and professional services
- Services for citizens and handicrafts
- Civil engineering
- Accommodation, hotel trade and travelling
- Warehouse, consumer commodity
- Health and insurance companies

2.3 Global positioning system

The Global Positioning System (GPS) is a satellite navigation system on exact position acquiring. It has been developed by US army and has been titled NAVSTAR GPS - NAVigation Signal for Timing And Ranging. It is a passive system which is able to provide data about position continuously 24 hours each day in whatever weather. It consists of 24 satellites (plus 3 backup satellites). Also Europe Union is developing a navigation system Galileo. This system will have 30 satellites and so it will provide larger covering than the present one.

Nowadays, the GPS technology is used in many applications. With the aid of GPS some interesting positions in terrain can be memorized and can enable returning to these places. GPS is becoming the navigation standard in civil, air and navy transport industry. In the USA, new application is developing for automatic aircraft landing using GPS navigation. The most dynamic developing area of GPS applications is the application field in car industry. GPS is available as a small specialised gadget or GPS can be found in mobiles, which can be used very successfully in tourism, travel and sports activities.

GPS positions acquiring is closely related with Geographic Information System (GIS). It is a computer system for integration, storing, modification, analysis, sharing and visualization of geographic information. One of the most known methods for GPS position acquiring is GPS tracking, provided in one or two steps. If the GPS tracking is provided in one step, street names are denoted at the same time as GPS logs. This approach can be used in case when the territory is mapped without possibility to return. If the GPS tracking is provided in two steps, in the first step GPS logs are denoted and in the second steps street names are denoted. A GPS logger is needed for GPS logs creating. The GPS logger creates records of GPS positions with time marks. From these data, maps and databases are created and many of them are available on the Internet. Such kind of data was used in the presented work.

3 Implementation

The introduced techniques have been implemented and integrated into an application named PathFinder. The PathFinder contains two basic modules. The first module uses Strike Iron and is intended for targeted information retrieval in the USA. The second one – Golden Pages (Zlate Stranky in Slovak) – is intended for targeted information retrieval in Slovakia. After the PathFinder starts, an introductory page is displayed. The form of this introductory page is illustrated in Fig. 2.

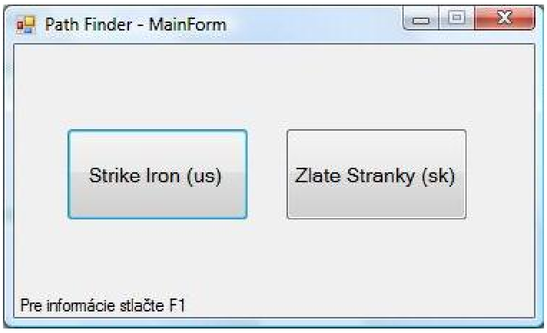
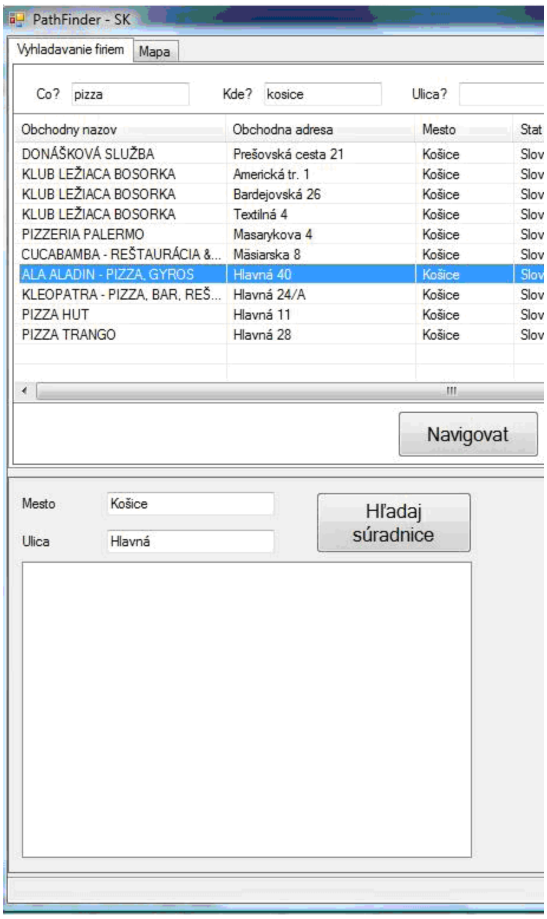


Figure 2: Introductory page of the PathFinder application

The application PathFinder tries to merge targeted information retrieval and identification of the GPS position of retrieved organizations/firms. GPS position enables to visualize the organization or firm site on the map. After clicking on the button Golden Pages (Zlate Stranky), information retrieval starts. If the category Restaurants in Kosice is selected and specialization Pizza is added, then user receives results, which are illustrated in Fig. 3.



Co? pizza Kde? kosice Ulica?

Obchodny nazov	Obchodna adresa	Mesto	Stat
DONÁŠKOVÁ SLUŽBA	Prešovská cesta 21	Košice	Slov
KLUB LEŽIACA BOSORKA	Americká tr. 1	Košice	Slov
KLUB LEŽIACA BOSORKA	Bardejovská 26	Košice	Slov
KLUB LEŽIACA BOSORKA	Textilná 4	Košice	Slov
PIZZERIA PALERMO	Masarykova 4	Košice	Slov
CUCABAMBA - REŠTAURÁCIA &...	Mäsiarska 8	Košice	Slov
ALA ALADIN - PIZZA, GYROS	Hlavná 40	Košice	Slov
KLEOPATRA - PIZZA, BAR, REŠ...	Hlavná 24/A	Košice	Slov
PIZZA HUT	Hlavná 11	Košice	Slov
PIZZA TRANGO	Hlavná 28	Košice	Slov

Mesto Košice

Ulica Hlavná

Figure 3: Results of information retrieval stage of the PathFinder for category Restaurants and specialization Pizza in Kosice

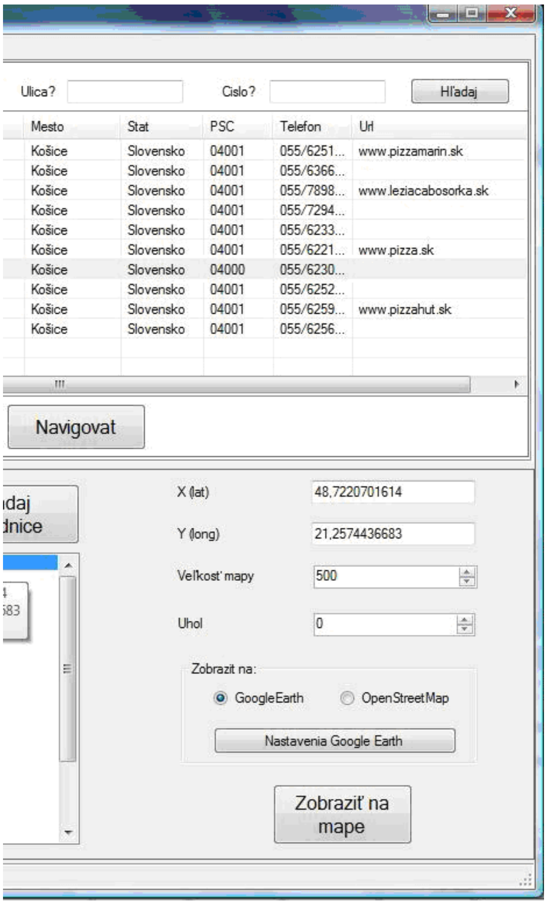
The PathFinder application consists of three parts:

1. Core – library of programms and calculation methods. The core library contains such classes like: *AppProcessing*, *DBConnectionOleDB*, *Global*, *GPSPosition*, *Helper*, *SearchedResults*, *StrikeIronManager*, *WebManager*, *XMLManager*, and so on.

2. PathFinder – main part for user interface. This part contains more specific classes and methods, for example: *GoogleEarthSettings, Publishers, Slovakia, GoogleEarthSettingsForms, MainForm, SlovakLocForm, USLocForm*, and so on.

3. TestApp – additional library for testing methods.

After user selection of the item "Ala Aladin – Pizza, Gyros" (marked by the cursor) and after clicking on the button Navigation (Navigovat), the both keys (town and street) are copied into bottom navigation part of the PathFinder window. After clicking on the button Coordinates Search (Hladaj suradnice), coordinates of the selected "Ala Aladin – Pizza, Gyros" are displayed in the right bottom corner of the application, as you can see in Fig. 4.



Ulica? Cislo?

Mesto	Stat	PSC	Telefon	Url
Košice	Slovensko	04001	055/6251...	www.pizzamarin.sk
Košice	Slovensko	04001	055/6366...	
Košice	Slovensko	04001	055/7898...	www.leziacabosorka.sk
Košice	Slovensko	04001	055/7294...	
Košice	Slovensko	04001	055/6233...	
Košice	Slovensko	04001	055/6221...	www.pizza.sk
Košice	Slovensko	04000	055/6230...	
Košice	Slovensko	04001	055/6252...	
Košice	Slovensko	04001	055/6259...	www.pizzahut.sk
Košice	Slovensko	04001	055/6256...	

X (lat) 48,7220701614

Y (long) 21,2574436683

Veľkosť mapy 500

Uhol 0

Zobrazit na: ● GoogleEarth ○ OpenStreetMap

Figure 4: Navigation information – coordinates of the selected "Ala Aladin – Pizza, Gyros"

The PathFinder user is expected to select type of map for results visualization: Google Earth from the Internet or OpenStreetMap. If he/she selects OpenStreetMap, the results illustrated in Fig. 5 will be available for him/her.

The application was tested in the sequence of experiments for various firms and organizations from Slovakia region. These tests were provided for institutions as the shop "VODAR" in Kosice, firms "YANYC DIESEL", "MOBILE HOUSE s.r.o", "MAJA" – cosmetic salon, "BERYL" guest house, etc. The precision of results of performed tests was sufficiently good for the statement that presented solution of the problem of GPS position searching is suitable for PathFinder application.

4 Conclusions

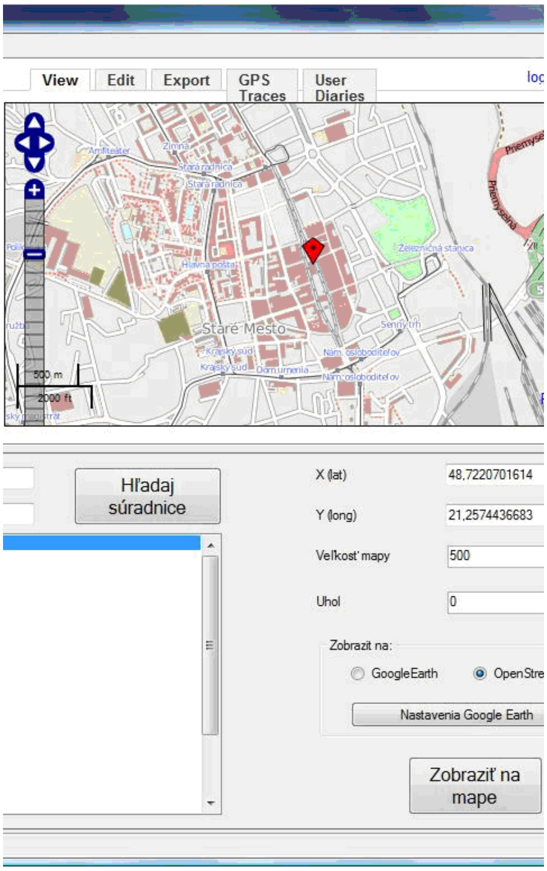
Presented application provides targeted information retrieval (name, address, telephone number, position on the map) within the same interface. It decreases user cognitive load. User does not need to open many various web services and systems. The implementation of the PathFinder does it instead of him/her. The implementation can by a sound basis for next development of semantic software agents.

This application can be used for other domains, for example information retrieval of contact information for a physician specialist (according to patient problems and according to patient's health insurance company) and positions on the town maps.

Another desired domain for PathFinder application can be a domain of estate agents and so on.

5 Acknowledgments

The work presented in the paper was supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic within the 1/0042/10 Project "Methods for identification, annotation, search, access and composition of services using semantic metadata in support of selected process types".



X (lat) 48,7220701614

Y (long) 21,2574436683

Veľkosť mapy 500

Uhol 0

Zobraziť na: ○ GoogleEarth ● OpenStre

Figure 5: Results of the navigation stage of PathFinder application for selected pizza restaurant

References

[1] Antoniu, G., van Harmelen, F.: A Semantic Web Primer. Massachusetts Institute of Technology, USA, 2004, 238 p., ISBN 0-262-01210-3.

[2] Augeri, Ch. J.: On Graph Isomorphism and the PageRank Algorithm. Doctoral thesis, Defense Technical Information Center, SEP 2008, 153 p.

[3] Bergman, M. K.: The Deep Web: Surfacing Hidden Value. Communication Abstracts, Vol.26, No.2, 2003, pp. 155-298.

[4] Fensel, D. at al.: Enabling Semantic Web Services. Springer-Verlag, Berlin, 2007, 188 p., ISBN 3-540-34519-1.

[5] Gála, L., Pour, J., Prokop, T.: Firm informatics. Prague, Grada, 2005, 282 p., ISBN 8024712784.

[6] Grossman, D. A., Frieder, O.: Information retrieval. Algorithms and Heuristics. Springer, Dordrecht, Netherlands, 2004, ISBN 1-4020-3003-7.

[7] Ledford, J. L.: SEO Search Engine Optimization Bible. John Wiley & Sons Inc., 2009, ISBN 9780470496800.

[8] Návrat, P., Bieliková, M., Chudá, D., Rozinajová, V.: Intelligent Information Processing in Semantically Enriched Web. Foundations of Intelligent Systems. Lecture notes in Computer Science (subseries: Lecture Notes in Artificial Intelligence), Springer, Berlin Heidelberg, Vol. 5722, pp. 331-340, 2009, ISSN 1867-8211.

[9] Schmotzer, M.: Agent cooperation in multi agent groups. Kognition and Artificial Life IX. Edition centre FPF of the Slezska University in Opava, Opava, 2009, pp. 287-290, ISBN 978-80-7248-516-1.