ISBN 978-953-56303-8-8
UDK 338.48+796+004+504(063)
God. 3, br. 3, 2014.

ZBORNIK RADOVA - TEHNOLOŠKI RAZVOJ U FUNKCIJI ODRŽIVOG TURIZMA I SPORTA

ISBN 978-953-56303-8-8
UDK 338.48+796+004+504(063)
God. 3, br. 3, 2014.

Međimursko veleučilište u Čakovcu

organizira

3. međunarodnu konferenciju Menadžment turizma i sporta

TEHNOLOŠKI RAZVOJ U FUNKCIJI ODRŽIVOG TURIZMA I SPORTA

Zbornik radova

pod pokroviteljstvom

Ministarstva znanosti, obrazovanja i sporta
Ministarstva turizma
Hrvatskog olimpijskog odbora
Hrvatske olimpijske akademije
Međimurske županije
Turističke zajednice Međimurske županije

Sv. Martin na Muri, travanj, 2014.

Međunarodna konferencija Menadžment turizma i sporta

TEHNOLOŠKI RAZVOJ U FUNKCIJI ODRŽIVOG TURIZMA I SPORTA

Sv. Martin na Muri, 10. i 11. travnja 2014.

Partneri:

Sveučilište Pannonia, Kampus Nagykanizsa
Sveučilište Mendel u Brnu, Fakultet za regionalni razvoj i međunarodne studije
Visoka škola za menadžment u turizmu i informatici u Virovitici
Univerza za Primorskem, Fakulteta za turistične študije – Turistica, Portorož
Sveučilište u Rijeci, Fakultet za menadžment u turizmu i ugostiteljstvu Opatija
Sveučilište u Zagrebu, Ekonomski fakultet
Sveučilište u Zagrebu, Kineziološki fakultet
Visoka škola za sportski menadžment Aspira
Gimnastički klub "Marijan Zadravec-Macan"
Međimurski savez sportske rekreacije "Sport za sve"

Pokrovitelji:

Ministarstvo znanosti, obrazovanja i sporta
Hrvatska olimpijska akademija
Hrvatski olimpijski odbor
Međimurska županija
Turistička zajednica Međimurske županije
Ministarstvo turizma

Supokrovitelji:

Općina Sv. Martin na Muri
Grad Čakovec
Spa&Sport Resort Sv. Martin

Izdavač

Međimursko veleučilište u Čakovcu

Glavna i odgovorna urednica

dr. sc. Nevenka Breslauer, prof. v. š.

Tajnica Uredništva

Mirjana Trstenjak

Recenzenti

prof. dr. sc. Mato Bartoluci
prof. dr. sc. Vladimir Pletenac
prof. dr. sc. Vidoje Vujić
Ivana Jadrić, dipl. oec.
izr. prof. dr. Aleksandra Brezovec
izr. prof. dr. Igor Jurinčič
doc. dr. sc. Romina Alkier
doc. dr. sc. Vesna Bedeković
doc. dr. sc. Vatroslav Horvat
doc. dr. sc. Goran Karanović
doc. dr. sc. Mirko Smoljić
doc. dr. sc. Zrinka Zadel
dr. sc. Alan Clarke
dr. sc. Oliver Jukić
dr. sc. Ivana Nikolić
dr. sc. Metod Šuligoj
mr. sc. Ivan Hegeduš
mr. sc. Mihael Kukec
mr. sc. Vesna Sabol
Draženka Birkić, univ. spec. oec.
Mihaela Mesarić, prof..
Ksenija Pletenac, prof.
Damir Mađerić, dipl. ing.
Ratko Matotek, dipl. ing. građ.
Ivica Levanat, dipl. ing.
Rikard Bakan, mag. oec.
Mladena Bedeković, dipl. oec.
Irena Bosnić, mag. oec
Andrijana Kos Kavran, mag. oec.
Đurđa Somođi, dipl. oec.

Programski odbor

prof. dr. sc. Mato Bartoluci
dr. sc. Nevenka Breslauer, prof. v. š.
dr. Nataša Pomazalová
prof. dr. sc. Mirna Andrijašević
dr. sc. Adrienn Skrop
dr. sc. Oliver Jukić
dr. sc. Mario Baić
dr. sc. Aleš Ruda
dr. sc. István Heckl
dr. sc. Blaženka Novak
mr. sc. Duje Bonaci
dr. sc. Saša Ceraj
dr. sc. Zlatko Mateša

Organizacijski odbor

dr. sc. Nevenka Breslauer, prof. v. š. – predsjednica
Mirjana Trstenjak–član
Andrijana Kos Kavran–član
Maja Bregović–član
Željko Hrebak–član
Marijan Vugrinčić–član
mr. sc. Stevo Tkalčec–član
Franjo Makovec–član

Lektori

Maja Bregović
Ivana Bujan

Grafičko oblikovanje i tisak

Graphprint d.o.o.

Naklada

200 komada

Accepted: 12.02.2014.
Reviewed: 13.03.2014.

Original scientific paper
UDC 025.4.036:004

Semantic analysis based search and retrieval system for discovering tourism services

Adrienn Skrop
Department of Computer Science and Systems Technology, University of Pannonia
8200 Veszprém, Egyetem u. 10. Hungary
e-mail: skrop@dcs.uni-pannon.hu

Abstract: *This paper presents research results obtained within the ÉLMÉNY2MAX project by applying semantic analysis based search and retrieval technologies in order to discover tourism services. The ÉLMÉNY2MAX system is designed to find tourism services packages that are close to each other in time and space according to users' preferences. The present paper is a detailed description of the information search and retrieval component of the ÉLMÉNY2MAX system. This part of the system recommends services using information provided by websites, portals and databases. Relevant information is located by semantic analysis based search technology. Tourism services relevant to user preferences are determined using novel information retrieval models that are capable of changing system categoricity at low computational costs.*

Keywords: *information search and retrieval, semantic search, vector space model, hyperbolic geometry, similarity measures*

1. Introduction

As it is well-known, the web has become one of the most popular Internet applications both for users and information providers. The information stored on Web pages can be categorized in several categories being used by a typical target user group. One important category is the category of tourism related services, which includes the Web pages of e.g. hotels, local attractions, festivals etc.. As a result, many people use the Web to prepare for a self-organized holiday. These users visit web sites to find accommodation, restaurants, travel information, programs, maps etc. The primary aim of a user is to organize a holiday that best meets his/her preferences. Preference may include several factors e.g. price, time interval and quality. It is difficult and time consuming to visit many Web pages to compare offers and to take into consideration as many factors as possible.

2. ÉLMÉNY2MAX

To overcome those difficulties which occur in the organization of programs or holidays the ÉLMÉNY2MAX system was designed and implemented. This work was carried out through the project ÉLMÉNY2MAX–Development of a search and browse system to locate and assist the participation of tourist services close to each other in time and space by a project consortium.

The ÉLMÉNY2MAX system provides personalized service for tourists by offering program packages that satisfy users' preferences. The goal of the system is to collect Web based information and services and to associate them using advanced technologies and mathematical models. On the one hand the system uses a user defined profile to identify relevant services. On the other hand social network and web calendar data is used to refine results, if available. The ÉLMÉNY2MAX system integrates many software components. The present paper presents the search and retrieval module of the system.

3. Search and retrieval module

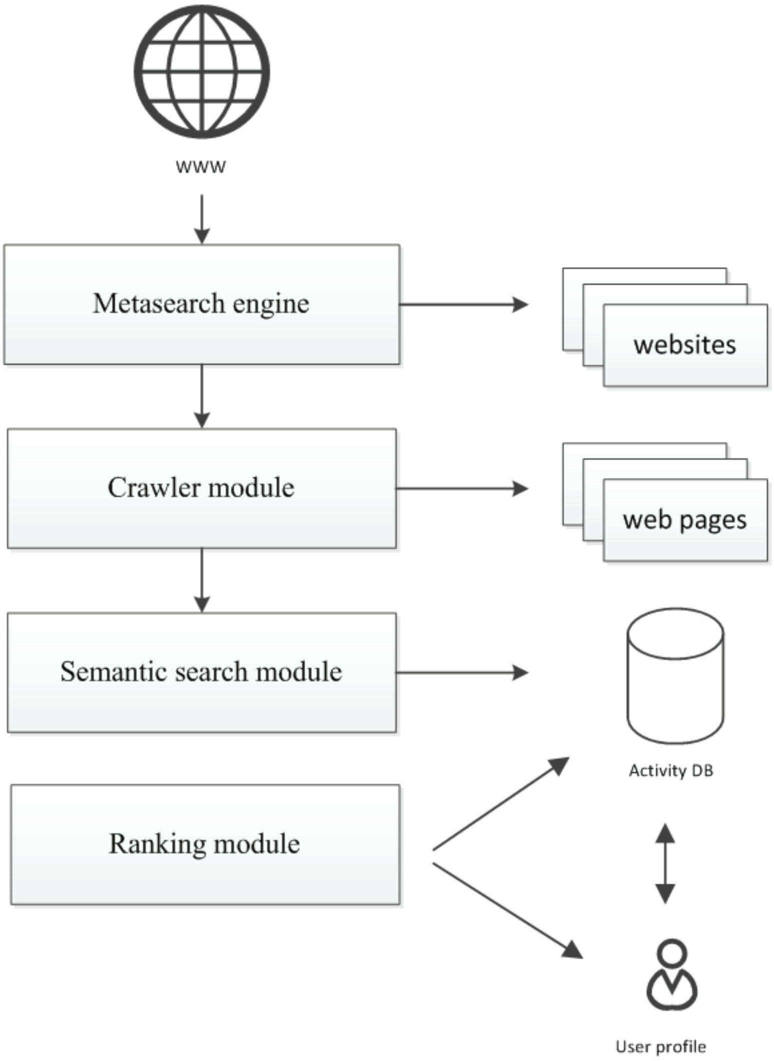
The search and retrieval module is responsible for discovering tourism related Web pages and collecting relevant data. To locate relevant data, semantic analysis based search technology is used. As opposed to keyword based information-retrieval, more precise results can be obtained. Novel crawler technologies are used to collect and classify free time activities.

The application consists of a number of computer program modules written in several languages as well as related documentation. Figure 1 shows the architecture of the module.

The task of the **Metasearch engine** is to discover and process tourism related relevant Web sites. It is implemented as a conventional keyword based Metasearch engine. Metasearch engines are search engines that search other engines (Croft et al, 2010). They submit the search query to several other search engines and return a summary of the results. This strategy gives the search a broader scope than searching a single search engine. The implemented Metasearch engine uses the hit list of Google[1] and Bing[2].

The **Crawler module** investigates the structure of Web sites, determines those pages of Web sites that contain relevant data, and indexes these pages using keywords.

Figure 1. Architecture of the search and retrieval module.



[1] http://www.google.com/
[2] http://www.bing.com/

The **Semantic search module** is responsible for the automated processing of Web pages that were identified by the Crawler module. Using this module relevant data items are located and inserted into the **Activity database**. Missing or inaccurate data is approximated using different methodologies.

The **Ranking module** matches user profiles and tourism services. It uses mathematical models to define similarity. Similarity is defined as a kind of distance. Based on this mathematical distance the system can determine whether user profiles and tourism activities/services are *close enough*. Several mathematical models are used to calculate similarity; thereby the degree of relevance is adjustable.

4. Methods

In this section the methods used by the ranking module are presented.

4.1. Vector space information-retrieval (IR)

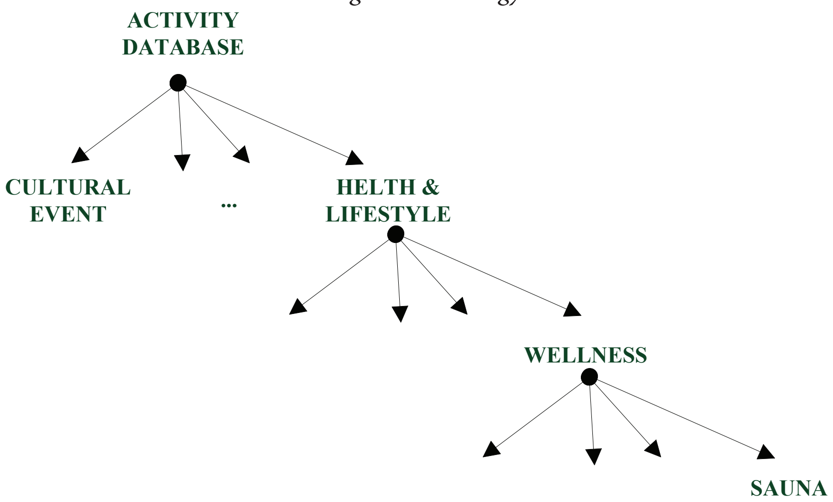
The ÉLMÉNY2MAX system introduces a new concept called *experience index*. The experience index determines the similarity between a user profile–hereinafter Profile -and touristic activities–hereinafter Activity–in the Activity database. Experience index is a real number. It takes values on the 0-1 interval.

In order to determine experience index a mathematical model of Profile and Activity is necessary. The vector space model proved to be applicable (Baeza-Yates and Ribeiro-Neto, 1999).

In this model user Profiles and Activities are represented as vectors. The similarity of Profile vectors and Activity vectors are determined using similarity measures. If the similarity i.e. the experience index is higher than a predefined threshold value, the Activity might interest the user.

As a first step the ÉLMÉNY2MAX vector space is defined. In ÉLMÉNY2MAX ontology is used to index Profiles and Activities. The ontology contains tourism related vocabulary organized into a hierarchy. Figure 2 shows a fraction of the ontology. The implemented ontology has 225 labels.

Figure 2. Ontology.



User Profiles are indexed by ontology labels. Each label takes a user defined value. The value is a real number on the 0-1 interval. Activities are indexed the same way. Based on these considerations, ontology labels are used as the dimensions of the vector space. In the vector space Profiles and Activities are represented as vectors.

Given a finite set L of ontology labels $L = \{l_1, ..., l_i, ..., l_n\}$. Any activity A_j is assigned a vector $\mathbf{v}_j$, and any Profile P_k is assigned a vector $\mathbf{v}_k$ of finite real numbers, as follows:

$$\mathbf{v}_j = (w_{ij})_{i=1,...,n} = (w_{1j}, ..., w_{ij}, ..., w_{nj})$$
$$\mathbf{v}_k = (w_{ik})_{i=1,...,n} = (w_{1k}, ..., w_{ik}, ..., w_{nk})$$

The weight w_{ij} is interpreted as an extent to which the label l_i characterises an Activity or a Profile.

An Activity A_j is represented to a user having Profile P_k if the Activity and the Profile are similar enough, i.e. a similarity measure S_{jk} between the Activity vector $\boldsymbol{v}_j$ and the Profile vector $\boldsymbol{v}_k$ is over some threshold K, i.e.

$$S_{jk} = s(\mathbf{v}_j, \mathbf{v}_k) > K$$

The following similarity measures are implemented in the application to calculate experience index:

- Dot product:
  $s_{jk} = (\mathbf{v}_j, \mathbf{v}_k) = \sum_{i=1}^{n} w_{ij} \cdot w_{ik}$
- Cosine measure:
  $s_{jk} = c_{jk} = \frac{(v_j.v_k)}{(||\boldsymbol{v}_j||\cdot||\boldsymbol{v}_k||)}$
- Dice's coefficient:
  $s_{jk} = d_{jk} = \frac{2(\mathrm{v}_j.\mathrm{v}_k)}{\sum_{i=1}^{n} \mathrm{w}_{ij}+\mathrm{w}_{ik}}$
- Jaccard's coefficient:
  $s_{jk} = J_{jk} = \frac{\frac{(\mathrm{v}_j.\mathrm{v}_k)}{\sum_{i=1}^{n} \mathrm{w}_{ij}+\mathrm{w}_{ik}}}{2^{\mathrm{w}_{ij}\mathrm{w}_{ik}}}$

4.2. Hyperbolic information-retrieval

In the vector space model the feature space is mathematically modelled by the orthonormal Euclidean space. In the hyperbolic information-retrieval model the vector space is defined over the Cayley-Klein Hyperbolic Geometry (Anderson, 1999). In hyperbolic IR the similarity measure is derived from the hyperbolic distance (Góth and Skrop, 2005).

In the ÉLMÉNY2MAX system the hyperbolic model is implemented, too, to calculate experience index. Profile vectors and Activity vectors are same as for the vector space model. The hyperbolic similarity measure is defined as follows:

$$S_{j,k} = \sigma_{j,k} = \left(ln\left(e \cdot \frac{r + \sqrt{\sum_{i=1}^{n} (w_{ij} - w_{ik})^2}}{r - \sqrt{\sum_{i=1}^{n} (w_{ij} - w_{ik})^2}} \right) \right)^{-1}$$

where $r > max_{\boldsymbol{v}_j} d_E(\boldsymbol{v}_j \boldsymbol{v}_k)$ and $d_E(\boldsymbol{v}_j \boldsymbol{v}_k) = \sqrt{\sum_{i=1}^{n} (w_{ij} - w_{ik})^2}$, i.e. the Euclidean distance of the Profile and Activity vector.

In the hyperbolic model the ranking order of the Activities is the same as in the vector model. However, thanks to the distinct categoricity of the measures–the degree of similarity is different. This property can be used to adjust similarity measurement according to preferences.

4.3. Interaction information-retrieval

Besides the vector space models other techniques were investigated to calculate the experience index. Interaction information-retrieval (I^2R) model (Dominich, 2003) proved to be applicable, too. Clustering is a well-known technique applied in IR. It is typically used to group documents to be searched. A special case of clustering is adaptive clustering i.e., a clustering in which the cluster structure is being developed under or is being influenced by an interaction with the user. One way of conceiving adaptive clustering is to adopt a connectionist-based view.

In the ÉLMÉNY2MAX application the experience index calculation using adaptive clustering is implemented as follows. Any Activity is represented by an object. An object o_i, $i = 1, 2, ..., M$, is assigned a set of identifiers. Identifiers are predefined ontology labels l_{ik}, $k = 1, 2, ..., n_i$. There are weighted and directed links between any pair (o_i, o_j), $i \neq j$, of objects.

The one is the relative frequency denoted by w_{ijp} of a term given an Activity, i.e., the ratio between the relevance r_{ijp} of ontology label l_{jp} in Activity object o_i, and the length n_i of o_i, i.e. total number of ontology labels assigned to o_i:

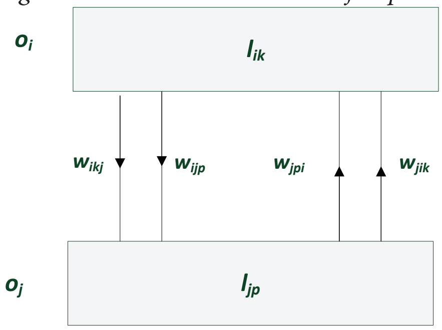
$$w_{ijp} = \frac{r_{ijp}}{n_i}, p = 1, \dots, n_j$$

The other weight is the extent to which a given ontology label reflects the characteristic of an Activity, i.e., the inverse document frequency. If r_{ikj} denotes the relevance of label l_{ik} in o_j, and df_{ik} is the number of Activities that can be indexed by label l_{ik}, then w_{ikj} is given by the inverse document frequency formula, and thus represents the extent to which l_{ik} reflects the characteristic of o_j:

$$w_{ikj} = r_{ijp} log \frac{2M}{df_{ik}}$$

The other two connections–in the opposite direction–have the same meaning as above: w_{jik} corresponds to w_{ijp}, while w_{jpi} corresponds to w_{ikj}. Figure 3 shows the connections between object pairs.

Figure 3. Connections between object pairs.



The Activities are represented as an interconnected network; every Activity is linked to every other Activity. The user Profile is conceived as being an object, too. It is interconnected with the already interconnected Activities causing some of the already existing connections to change because of the change of M and df_{ik}. The objects are conceived as being a network of artificial neurons in which a spreading of activation takes place according to a winner takes all strategy. The activation is initiated at the Profile, and spreads over along the strongest connection thus passing on to another neuron, and so on. The strength of the connection between any pair (o_i, o_j), $i \neq$ i, of objects, and thus between the Profile and another Activity object o_i is defined as follows:

$$K_{ij} = \sum_{p=1}^{n_j} w_{jpi} + \sum_{k=1}^{n_i} w_{jik}$$

After a number of steps the spreading of activation reaches an object that was already affected. This is analogous to a local memory recalled by the Profile. Those Activities are said to be interesting regarding to the user Profile which belongs to the same circle. The corresponding Activities are ranked in the order of maximal activation, i.e. the experience index.

5. Conclusion

The ÉLMÉNY2MAX system was develop to help tourists finding relevant free-time activities, organizing holidays etc.. One of the most important components of the system is the search and retrieval module. It is responsible for discovering and offering tourism related activities that may satisfy users' preferences. The similarity between Activities and user Profiles is determined by using the experience index. The experience index is determined by using different mathematical models and similarity measures. Thanks to the distinct categoricity of different measures the degree of similarity is easily adjustable according to preferences.

The experimental ÉLMÉNY2MAX system is available at http://elmenymax.dcs.uni-pannon.hu:4000/. The system can be improved by using a number of specific mathematical models and by implementing development proposals arising from the use of the ÉLMÉNY2MAX system.

6. Acknowledgements

This research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project GOP-1.1.1-11-2011-0029. Thanks go to B.Sc. students for helping us implementing the computer program.

References

1. Anderson, J.W. (1999). Hyperbolic Geometry. Springer Verlag, New York.
2. Baeza-Yates, R., and Ribeiro-Neto, B. (1999). Modern information retrieval. New York: ACM press.
3. Croft, W. B., Metzler, D., & Strohman, T. (2010). Search engines: Information retrieval in practice (p. 283). Reading: Addison-Wesley.
4. Dominich, S. (2003). „Connectionist interaction information retrieval." *Information processing & management,* vol. 39(2), 167-193.
5. Dominich, S., Skrop, A., and Tuza, ZS. (2006). Formal Theory of Connectionist Web Retrieval. In: Crestani, F. et al. (eds.) Soft Computing in Web Information Retrieval, Springer Verlag, pp. 161-194. ISBN: 3-540-31588-8.
6. Góth, J., Skrop, A. (2005). „Varying retrieval categoricity using hyperbolic geometry". *Information Retrieval*, vol. 8(2), 265-283.