The TSTS Method in Cultural Heritage Search

Mirosław Stawniak and Wojciech Cellary

Department of Information Technology, Poznań University of Economics
{stawniak,cellary}@kti.ue.poznan.pl

Abstract. In cultural heritage content management systems in which cultural objects are described with the use of their semantic, temporal and spatial properties, the search capabilities taking all those properties into consideration are very limited. The difficulty comes from the fact that concepts evolve over time and depend on location. In this paper the TSTS search method is presented based on the TST similarity measure that allows assessing the similarity factor between different resources in a knowledgebase. A ranked search result is generated basing on the semantic distance between the fuzzy set created for the user query and fuzzy sets describing potential results in the time-space continuum.

Keywords: Semantic web, semantic search, cultural objects, cultural heritage.

1 Introduction

In recent years increasing interest in creating virtual museums is observed. Virtual museum objects are stored in heritage content management systems. Detailed descriptions of the objects include e.g. the dates and places of their creation, storage and renovation. In many content management systems objects are also mutually linked according to their semantic conceptual relationships. Current heritage content management systems usually allow users to search for cultural objects, however, the search capabilities provided are quite limited. The problem is that concepts related to cultural objects are dependent on time and geographical location, can evolve over time, and may have different meaning in different locations. As a result, search methods developed so far, both keyword-based and semantically oriented, fail because of the imprecision of data - on the one hand contained in museum knowledgebases, and on the other hand used in the queries by users.

In this paper a TSTS (Theme-Space-Time Search) method of semantic search is proposed that takes into account three factors to answer a user query: relations between concepts in context of their temporal and spatial properties. Contribution of the TSTS method to technological innovation in the domain of semantic search in the cultural heritage field is presented in Section 2. In Section 3, current solutions for cultural objects search engines are discussed. In Section 4, the TSTS method is presented, including the TST concept similarity measure and its application in search. The characteristics of the TSTS method are discussed in Section 5. Finally, Section 6 concludes the paper.

L.M. Camarinha-Matos, P. Pereira, and L. Ribeiro (Eds.): DoCEIS 2010, IFIP AICT 314, pp. 44–51, 2010.

2 Contribution of the TSTS Search Method to Technological Innovation

To present contribution of the TSTS search method to technological innovation consider the following example of a fragment of a cultural heritage domain presented in Figure 1. This domain deals with cultural objects, geographical places, and historical periods. Rectangles represent concept instances, while ovals – their properties. For the sake of clarity, only properties describing time spans (denoted by t) and geographical coordinates (denoted by g) in literal values are presented. In this example some of the concept instances are directly described by the time and space information (e.g. Vase 1 was created in time t_1 and place g_1), while for others time/space information is expressed by references to other concept instances (e.g. Vase 2 was created in time t_2 in Greece). Moreover, some concept instances can be linked with more than one pair (time, place) since they evolved over time or appeared in different places at different times. Greece is an example of an evolving concept instance, because the borders of Greece have changed many times. As a result, location and area of 'Greece' depend on the date considered. As a consequence, when the term 'Greece' is used in a query, time has to be specified to obtain a precise response. Similarly, term 'Archaic Period' is not bound to only one period in time, but depending on place may refer to different time spans. For example the archaic period occurred in Egypt between 3100-2600 BC, but in Greece between 750-480 BC.

A typical query issued by an archaeologist is: *find all the vases originating from Greece created in the archaic period.*

There are two types of search methods that could be used to answer such a query: keyword search and semantic search. Keyword search is the simplest and the most common search method. In the cultural heritage context, where terms are ambiguous, results obtained using this method are usually inaccurate. This method is able to return only those objects for which a given keyword occurs inside their textual description.

Semantic search methods are usually more accurate than keyword based methods, because they are able to analyze relations between concepts and their instances. However, if a purely semantic search method was used in the example above, it

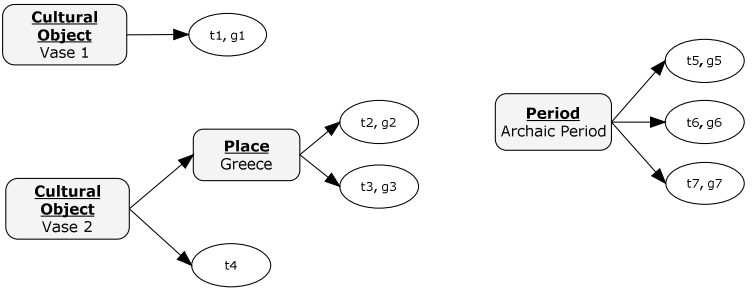

Fig. 1. A fragment of a simplified semantic model instance for a cultural heritage domain

would also return inaccurate results, because some concepts are not explicitly connected to each other, but they relate only through their temporal and geographical coincidence (e.g. 'Archaic period' and 'Vase 1' are related because they both occurred in the same place at the same time).

The TSTS method of semantic search presented in this paper takes into account three factors to answer a user query: relations between concepts together with their temporal and spatial properties. As the majority of users are accustomed to expressing their information needs in terms of keywords, the TSTS method allows users to ask questions using a traditional 'Google like' interface. However, the TSTS uses semantic information concerning the application domain to obtain results that are not possible to be found when traditional search methods are applied.

The TSTS method is based on the TST similarity measure that allows assessing the distance between different concepts in a semantic, spatiotemporal database. During the evaluation of this measure, for each concept a fuzzy set of points in the time-space continuum is constructed. In such a fuzzy set, the more closely a point is related to the analyzed node, the higher degree of membership it receives. In the next step fuzzy set operations such as intersection and union are applied to the fuzzy sets that correspond to the concepts from the user query. As a result, another fuzzy set is obtained that represents the region in space and time that the user is interested in.

In the example above, in a user query three concepts 'Vase', 'Greece' and 'Archaic Period' are mentioned. The concept 'Vase' yields the set containing points in the time-space continuum that correspond to all acts of the creation of a vase. For the concept 'Greece', a set is generated including all the points in space in which Greece has ever existed. Finally, the concept 'Archaic Period' yields the set of points in which that period occurred. Intersection of these three fuzzy sets is a region in time and space matching the user's query. Finally, cultural objects that are 'closest' to that region are found and presented as the query result.

3 Searching Cultural Objects

A number of different approaches to cultural object searching have been developed and presented in the literature.

Fitzwilliam Museum [1] website makes its collection accessible through the standard OPAC (Online Public Access Catalog) search interface. The search is based on keywords. Although, it is possible to issue a query limiting search results to a certain place and time period, search is performed on textual basis. Mainly due to data inconsistency, search results are prone to be inaccurate, which is a common disadvantage of search methods based on keywords.

Heritage 4 [2] is a system for storing and presenting cultural objects used to build the virtual Museum of Copenhagen [3]. In this system users can conduct searches choosing textual, geographical and/or, temporal criteria or through relations between objects. Although in the Heritage 4 system search is made via an appealing interface, it does not support complex queries such as the one presented in Section 1. The spatiotemporal search can only operate on objects that are directly described by space and time properties and those properties cannot be inferred through relations with other objects. Therefore, this system does not indirectly support the evolving concepts.

Semantic search is a process used to improve searching capabilities by using semantic networks to disambiguate queries in order to generate more relevant results. The semantic search methods are divided into two categories: exact methods and approximate methods.

The SPARQL language is devoted to the exact semantic search methods. SPARQL is a language for querying the RDF semantic networks developed by RDF Data Access Working Group (DAWG) [4]. SPARQL is designed for finding patterns in semantic graphs. Although the SPARQL language is very powerful, initially it was not designed to answer the queries that deal with time and space . Some extensions to SPARQL exist [5] that allow querying for time and space, but they make the use of this language even more sophisticated.

Another interesting approach to the exact semantic search is proposed in [6], in which spatial and temporal query operators have been formalized and evaluated. Although the proposed framework was designed to deal with queries containing references to time and space, there is no direct support for the changing or evolving entities.

In general, the exact search methods have two disadvantages. First disadvantage concerns the relationship with a knowledgebase. The exact search methods either require a user to know the structure of a knowledgebase and then the search system may cooperate with different knowledgebases, or the search system is strictly bound to knowledgebases of a single structure fixed a priori. The second disadvantage of the exact search methods is their inability to return the results that do not meet a given search criterion, but which are close to meet it. Although this property is not considered to be a disadvantage in certain applications, in the cultural heritage domain it is a disadvantage. In the cultural heritage domain spatiotemporal data in the knowledgebase and in the user query can be imprecise, so the possibility to find objects that are close to meeting a given criterion and ranking the results is very important.

Approximate semantic search methods are based on similarity measures. These methods usually do not require a user to know the structure of the semantic network and they permit to rank the results. Although there has been a number of such measures proposed [7-8], none of them takes both time and space information into consideration.

The TSTS method presented in this paper can be categorised as an approximate semantic search method, because it is based on a similarity measure, namely the TST (Theme-Space-Time). The main feature of the TSTS method, as opposed to the existing methods, is its ability to seamlessly deal with the concepts evolving or changing in time and space. Also, the TSTS method does not require a user to know the structure of the searched knowledgebase. Finally, the TSTS method is able to operate on imprecise data and rank the search results.

4 The TSTS Method

In this section, the word *node* is used to denote a concept or a concept instance in a semantic network; the word *edge* – to denote a relation instance, and the word *graph* – to denote the semantic network.

4.1 Concept Similarity Measure

The TST (Theme-Space-Time) measure presented in this paper indicates similarity between two concept instances in a semantic network. The computation of this measure in done in three steps: weight mapping, fuzzy set creation, and distance measuring.

4.1.1 Weight Mapping

In a classical semantic network each edge has only a label associated to it. Therefore, it is only possible to indicate the presence or absence of a relation between two concept instances. The computation of the TST measure requires additional numerical value associated with the relation instances expressing their strength. Different ideas can be found in literature for devising a calculation that can generate a formula for the strength of each existing relation instance in the knowledgebase. However, despite the research, a formula that outperforms all the others in a large set of experiments, or proves to be the best for all application domains has not been found. In this paper the technique of calculating a numerical weight value based on analysis of link structure, called *weight mapping,* presented in [8] has been applied. The following specificity measure was used:

$$W_{ij} = \frac{1}{\sqrt{n_j}}. \tag{1}$$

The value n_j is equal to the number of instances of a given relation type that have concept C_j as its destination node. Therefore, the weight of the relation is inversely proportional to the number of relations with the concept C_j. The more concepts with the same type as C_i and related to C there are$_j$, the higher W_{ij} is.

4.1.2 Fuzzy Set Creation

In the fuzzy sets theory each element is given a degree of membership. Mathematically it is described with the aid of a membership function valued in the real unit interval [0,1]. Operations on fuzzy sets are generalizations of crisp set operations such as union or intersection. There is more than one possible generalization. Fuzzy intersection and fuzzy union are the functions that map the degree of membership from given sets into a new degree of membership.

The TSTS method requires that for each concept being measured the fuzzy set of points representing the concept in space-time continuum is created. The main idea is that if a given concept is related to other concepts that are positioned somewhere in time and space, there is a great possibility that the concept itself is also related to the same point.

For concept C_i the membership function expressing the fuzzy set is expressed by the following recursive formula:

$$\mu_i(x) = \bigcup_j (1-k)\alpha_i(x) \oplus_{Wij} \mu_j(x). \tag{2}$$

Function μ_i denotes the fuzzy set for the concept C_i. Function α_i represents the region in the space-time continuum defined by the properties immediately defined for node C_i. Function μ_j denotes the fuzzy set for the concept C_j being the neighbouring node of the node C_i.

Operator $\oplus_{Wij}$ is able to interpolate between the fuzzy union and fuzzy intersection depending on the weight W_{ij} of the relation between nodes C_i and C_j. This operator is defined by the following function:

$$\mu_{i\oplus j}(x) = W_{ij} s_{i\cup j}(x) + (1 - W_{ij}) t_{i\cap j}(x)$$
$$s_{i\cup j}(x) = \mu_i(x) + \mu_j(x) - \mu_i(x)\mu_j(x), \quad t_{i\cap j}(x) = \mu_i(x)\mu_j(x) \tag{3}$$

The *s* and *t* are the functions that were chosen for computing fuzzy union and fuzzy intersection, respectively.

During the process of computing the fuzzy set, for each propagation through an edge the degree of membership is reduced. The process repeats until newly visited nodes do not add any significant information or there are no further nodes to be processed. The *k* value corresponds to the percentage of the degree of the membership value that the membership function is decreased by every time an edge is processed, effectively functioning as an attenuation factor.

4.1.3 Measuring Distance

The TST measure is the distance between two nodes and it is the compound of the semantic distance in the semantic network, and the distance in time and space. It is defined by the following formula:

$$TST(C_i, C_j) = \lambda_s d_s(C_i, C_j) * \lambda_g d_g(\mu_i, \mu_j) * \lambda_t d_t(\mu_i, \mu_j). \tag{4}$$

Function $d_s(C_i, C_j)$ is the semantic similarity between concepts C_i and C_j. Computing the semantic similarity is widely discussed in the literature, where many different formulas were proposed. For the TSTS method, a function that simply counts the intermediate nodes on the shortest path between C_i and C_j was chosen.

The functions $d_g(\mu_i, \mu_j)$ and $d_t(\mu_i, \mu_j)$ are the distances between points in the fuzzy sets computed for concepts C_i and C_j in the geographic coordinates and time dimension respectively, expressed by the formulas:

$$d_g(\mu_i, \mu_j) = \int_0^1 \min_{(t_i, g_i)\in \mu_i^{>\alpha}, (t_j, g_j)\in \mu_j^{>\alpha}} (|g_i - g_j|) d\alpha \tag{4}$$

$$d_t(\mu_i, \mu_j) = \int_0^1 \min_{(t_i, g_i)\in \mu_i^{>\alpha}, (t_j, g_j)\in \mu_j^{>\alpha}} (|t_i - t_j|) d\alpha, \tag{5}$$

where $\mu_i^{>\alpha}$ is a set of points from the fuzzy set described by function μ_i for which the degree of membership is greater then α .Value g_i is the geographical part of an element in the fuzzy set. Value t_i denotes the time part of an element in the fuzzy set. Parameters λ_s , λ_g , λ_t are the scaling factors that allow the end user to weight which part of the measure is the most important.

4.2 Answering User Queries

Searching for cultural objects that satisfy a query specified by a user is performed in three phases: the resolving phase, the fuzzy set building phase, and the measuring phase. In the resolving phase each keyword or a phrase from the user query is mapped to an existing node in the underlying semantic network. In many cases a keyword alone cannot be used to unambiguously identify the node. In such case the user is presented with a list of matching nodes and is asked to indicate the concept or its instance of his/her interest.

In the next phase, a fuzzy set of points in the time-space continuum is created from the nodes contained in the user query. The nodes from the user query are joined through the three most common relations: *and*, *or* and *not*. The resulting fuzzy set is constructed using appropriate fuzzy set operations on the fuzzy sets corresponding to the nodes in the query – fuzzy intersection for *and*, fuzzy union for *or,* and fuzzy complement for *not*.

Finally, for each cultural object in the underlying knowledgebase the TST distance between the node representing the cultural object and the user query is computed. If the user query contains only one node, the TST distance has the form described in Section 4.1.3. If the user query consists of more than one concept, the semantic distance d_s in the presented formula needs to be computed differently. The value d_s is obtained by combining semantic distances between nodes referenced in user query and the currently analysed cultural object node using the same functions that were used to join the fuzzy sets. On the base of the TST measure the results are ranked. Those cultural objects for which the TST distance is the lowest are the best match.

In the presented TSTS method, the results can be shown as a flat list, because the TST measure is a numerical value. However, the final value of TST measure is compounded of three distances – in semantic, temporal and geographical dimension. All those distances can be separately visualized for each object found, giving a user a better understanding of why a particular cultural object was considered. Moreover, the fuzzy set constructed from the user query can also be visualized, indicating the actual area in time and space that was covered by the user query.

5 Discussion

The TSTS method responses to the problem of searching objects in the cultural heritage domain, where concepts are dependent on time and geographical location. As a result, those concepts are imprecise both in the knowledgebases describing cultural heritage subdomains and in the query expression. Imprecision concerns rivers, seashores, islands, countries, periods, as well as artistic, literary and intellectual

movements, etc. The use of fuzzy sets to represent concepts in the time-space continuum permits to deal with this imprecision. The TSTS method is independent of the knowledgebase schema. The search results may be ranked, as they have numerical values of the TST similarity measure attached.

6 Conclusions

The TSTS search method is particularly important in case of multinational and multicultural audience visiting a virtual museum. In such case, queries are imprecise more than ever, as searchers do not have background knowledge of the heritage domain. The TSTS method provides information that contributes to acquiring that knowledge, because it locates the concepts searched in time and space. Thanks to that location, the objects found may be presented in a clear and attractive way on maps and time scales.

Future research on the TSTS search method will be focused on the comparison of different measures. Generalization and specialization of concepts has to be taken into account to improve accuracy of the TSTS method, when applied in closed cultural heritage subdomains.

References

1. Fitzwilliam Museum, http://www.fitzmuseum.cam.ac.uk/
2. Snizek, B., König, T.: Heritage 4 – A 4D Cultural Heritage Management System. In: 15th Int. Conf. on Virtual Systems and Multimedia VSMM 2009, Vienna, Austria (2009)
3. Museum of Copenhagen, http://absalon.nu/
4. SPARQL Query Language for RDF (2008), http://www.w3.org/TR/rdf-sparql-query/
5. Dodds, L.: SPARQL Geo Extensions (2006), http://xmlarmyknife.com/blog/archives/000281.html
6. Perry, M., Hakimpour, F., Sheth, A.: Analyzing theme, space, and time: an ontology-based approach. In: Proc. 14th Annual ACM Int. Symp. on Advances in Geographic Information Systems. GIS 2006, Arlington, Virginia, USA, November 10-11, pp. 147–154. ACM, New York (2006)
7. Rocha, C., Schwabe, D., Aragao, M.P.: A hybrid approach for searching in the semantic web. In: Proceedings of the 13th international Conf. on World Wide Web. WWW 2004, May 17 - 20, pp. 374–383. ACM, New York (2004)
8. Ziegler, C., Simon, K., Lausen, G.: Automatic computation of semantic proximity using taxonomic knowledge. In: Proc. 15th ACM Int. Conf. on Information and Knowledge Management. CIKM 2006, pp. 465–474. ACM, New York (2006)