Proceedings of the Ninth IASTED International Conference
Visualization, Imaging, and Image Processing (VIIP 2009)
July 13 - 15, 2009 Cambridge, UK

SIMILARITY/CLOSENESS-BASED RESOURCE BROWSER

Oleksiy Khriyenko[1], Vagan Terziyan[2]
Industrial Ontologies Group, [1]Agora Center and [2]MIT Department, University of Jyväskylä
P.O. Box 35(Agora), FIN-40014 Jyväskylä
Finland
[1]oleksiy.khriyenko@jyu.fi, [2]vagan.terziyan@jyu.fi

ABSTRACT
Now, when, according to recent Web evolution trends, a human becomes a very dynamic and proactive player in a large highly heterogeneous and distributed environment with a huge amount of different kind of data, services, devices, etc., it is quite necessary to provide a technology and tools for easy and handy human information access and manipulation. Context-awareness and intelligence of user interface brings a new feature that gives a possibility for user to get not just raw data, but required information based on a specified context. A user needs fast and convenient way to specify what she is looking for and get the semantically closest resources to her query. Resource closeness/similarity search is one of the most popular features that users need during resource/information retrieving process. The similarity search has become a fundamental computational task in many applications. Thus, visualization of the resources in a context of their similarity/closeness becomes important functionality of the GUI and browsers. The paper presents an approach for similarity/closeness-based resources search/browsing and visualization, and focused on resource distance measuring functions for resources similarity/closeness calculation.

KEY WORDS
Similarity/closeness search, resource similarity/closeness distance, resource similarity/closeness visualization, similarity/closeness-based resource browsing

1. Introduction

The similarity search has become a fundamental computational task in many applications. Today similarity-based search is used in numerous fields of applications like e-commerce, data mining and knowledge discovery, case-based reasoning, knowledge management, text, image and information retrieval, etc. Current challenge is a distributed nature of information and therefore high heterogeneity of entities to be compared. Heterogeneity here means that information about these entities (resources) is coming from different sources and presented according to different schemas. Even if such information is well-structured and machine-processable still it is quite challenging to automatically compare even obviously similar resources but represented by different properties. Traditional matchmaking techniques (needed e.g. in database querying) may return just a fact “same or not” and they are not usually mature to find if not the same ones but closest instances to the query. Other challenge of heterogeneity is the existence of quite a lot of different definitions of closeness and appropriate mechanisms providing various functions to calculate the similarity [1] as well attempts to smartly combine various similarity functions [2]. Also an important heterogeneity challenge is selection of relevant features of the instances, being compared, to be used for computing similarity. This selection is obviously either subjective or context-dependent [3] and there are quite a lot of techniques for dimensionality reduction or feature selection [4]. Finally, of course, the heterogeneity of resources attributes itself (numerical, nominal, logical, interval, textual, etc.) is a challenging problem and various methods to prepare data [5] for computing similarity as well as computing functions are exist and it is always a challenge to select the right one for a particular task. Recently all above problems has attracted much attention in the database community because of the growing need to deal with large volume of data. Consequently, efficiency has become a matter of concern in design. Although much has been done to develop structures able to perform fast similarity search, results are still not satisfactory, and more research is needed.

Fast development of semantic technology [6][7] and appropriate W3C standards for metadata and ontologies (e.g. RDF,OWL) creates new interesting concept for similarity search, i.e. “semantic similarity”. It assumes measuring similarity between resources based on the likeness of their meaning or semantic content provided by the metadata. Machine processable semantics of data allows automating the process of similarity search even when dealing with heterogeneous and widely distributed resources (documents, services and other capabilities, people, devices, etc). Also ontologies are providing valuable basic domain knowledge (class and property taxonomy) that can be used for more precise similarity measure in addition to the metadata attached to the resources.

Another important issue is how to represent results of a similarity based search to a human, i.e. what kind of visualization technique would be convenient to for a user depending on his tasks. As the current trends develop we expect to experience a future Web which will be media rich, highly interactive and user oriented. The value of this Web will lie not only in the massive amount of information that will be stored within it, but in the ability of Web technologies to organize, interpret and bring this information to the user. Emerging technological trends strongly suggest that it is time to initiate a new stage in multidimensional resource visualization (visualization of resource properties, contexts of inter-resource communication and interaction) and a new stage in semantic metadata based visual browsing across resources [8][9]. The problem with manipulating a huge amount of information is the complexity of search query specification and provisioning of the relevant links for content browsing. The idea of intelligent resource visualization is to simplify the search and browsing processes via multidimensional associative resource visualization means visualization of a resource depending on a context, via association with various aspects of the resource (relations with other resources, domains, areas of interest, etc.). Such visualization can give us a hint, turn us to the right direction, show us related objects and provide links to them. In other words, visualization will utilize context-based filtering and enrichment of the visualized scene with the relevant links.

4I(FOR EYE) technology [8][9] will enable creation of such smart human interfaces through flexible collaboration of an Intelligent GUI Shell, various visualization modules, which we refer to as MetaProvider-services, and the resources of interest. Semantically enhanced context-dependent multidimensional resource visualization provides an opportunity to create intelligent visual interface that presents relevant information in more suitable and personalized for user form. Context-awareness and intelligence of such interface brings a new feature that gives a possibility for user to get not just raw data, but required integrated information based on specified context. Ability of the system to perform semantically enhanced resource search/browsing based on resource semantic description brings a valuable benefit for today Web and for the Web of the future with unlimited amount of resources.

With a purpose to prove the concepts under 4I (FOR EYE) technology, we elaborated and developed functional prototype. We highlighted the most important requirements for such an intelligent human resource browser and implemented initial main parts of it. As one of the results of the ongoing UBIWARE project www.cs.jyu.fi/ai/OntoGroup/U BIWARE _details.htm), we elaborated probably the most important part of 4I vision, which can be called “context provision”. Especially when considering a human, presenting information on a resource of interest alone is not sufficient - information on some “neighboring” objects should be included as well, which form the context of the resource. What is important is that in different decision-making situations, different contexts are relevant: depending on the situation the relevant neighborhood function may be e.g. physical spatial, data-flow connectivity, what-affects-what, similar-type, etc. Resource closeness/similarity search is one of the most popular functions that users need during resource/information retrieving process. Thus, visualization of the resources in a context of their similarity/closeness becomes inherent functionality of the 4I(FOR EYE) Browser.

There are semantic approaches based on ontologies and matching techniques for resource discovery - finding the existing resources that best match the requirements of a given resource request [10][11][12]. But we decided to start from the small steps, and concentrated on several resources’ property types matching.

The paper contains three main sections. Section 2 presents an approach for closeness-based resources search/browsing, Section 3 is focused on resource distance measuring and Section 4 describes an approach for closeness visualization (see Fig.1).

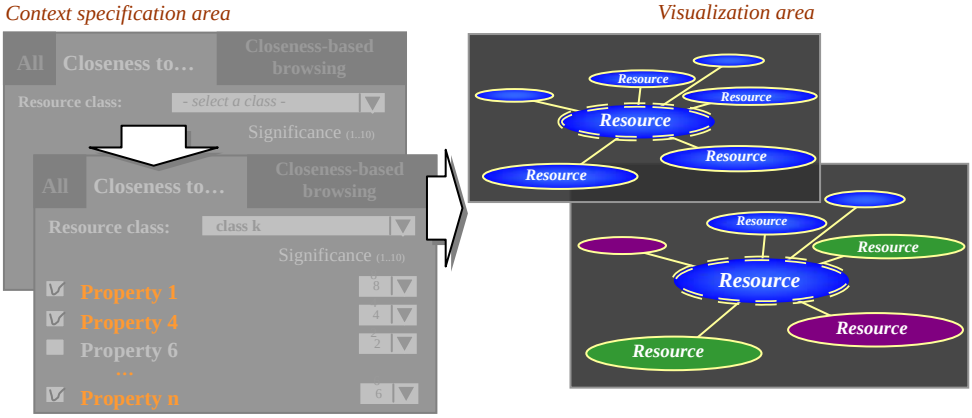

Fig. 1 - Resource visualization in a context of its closeness to other resources.

2. Closeness-based resource search/browsing

To avoid misunderstanding, we have defined two resource relationship types: closeness as *closeness* to resources of another different class and *similarity* as closeness to resources of a same class (can be considered as a particular case of *closeness* type).

Calculation of a distance measure between resources is based on a set of preselected relevant/common properties and significance of those properties for a distance measure. Each pair of resource classes has own set of relevant properties. But still, this attributes are configurable. User may reduce amount of properties (select just subset of them) to be considered by matchmaking algorithm and adjust/customize significance of the properties. If user does not provide detailed specification of the context, then matchmaking bases on all common contextual properties. As a result of a matchmaking algorithm, we have a matrix of distances between subject resource and other resources that helps us clearly show closeness via visual representation in GUI.

Such approach of finding similar/close resources can be applied for "one step" closeness-based resource browsing. The idea is to visualize the resource in a context of its closeness to other resources based on just only one property selected by user. Then user can change the focus and select another resource to visualize it in a context of its closeness to other resources based on any its property. Such closeness can be searched among all relevant resources or among specific class of resources specified by user.

The same technique that we use for resource closeness visualization can be utilized for resource ranking. The only requirement for this is to describe "virtual/abstract" (or chose from existing) etalon resource and calculate the distances of all other resources to that one. For complex resource ranking methods it might be that we have to elaborate appropriate modifications to the existing prototype. We consider a work in this direction as a future one.

In our prototype, 4I GUI Shell uses xml-based resource storage. Such architecture requires converting the date from original format to internal xml representation. Comparison between the resources is performed based on common properties. With the purpose to make a first step towards resource closeness detection we signed up five resource's property types:

Text field types:

Type 1: Just a pure word/sentence. Additional contextual information for this field is its significance.

```
<fieldContext>
   <field_type>textField</field_type>
   <field_significance>…</field_significance>
   <field_calculation_method>…</field_calculation_method>
</fieldContext>
```

Type 2: Text field is presented by list of key words/sentences. Additional contextual information for this field is its significance.

```
<fieldContext>
   <field_type>keyWordsField</field_type>
   <field_significance>…</field_significance>
   <field_calculation_method>…</field_calculation_method>
</fieldContext>
```

Type 3: Text field is divided to the set of attributes and presented by correspondent list of values (words/sentences) of the attributes. In this case, the number of the attributes for certain text field should be defined and lists of possible (defined) values of the attributes should be defined and presented. In another words, it is defined amount of keywords, where each keyword is selected from a correspondent defined set of values. Additional contextual information for this field is the sets of values for each attribute (keyword) and the significance of the attributes, and as for all fields, significance of the field itself.

```
<fieldContext>
   <field_type>complexTextField</field_type>
   <field_significance>…</field_significance>
   <field_calculation_method>…</field_calculation_method>
   <corClasses>
            <corClass>
         <class_significance>…</class_significance>
         <value>…</value>
      …
      </corClass>
      <corClass>
         <class_significance>…</class_significance>
         <value>…</value>
      …
      </corClass>
   …
   </corClasses>
</fieldContext>
```

Number field: Just number that further will be normalized and compared. Additional contextual information for this field is its significance.

```
<fieldContext>
   <field_type>numberField</field_type>
   <field_significance>…</field_significance>
   <field_calculation_method>…</field_calculation_method>
</fieldContext>
```

Interval field: Field presented by start and end point on a numerical axis. Distance measuring function for such interval field is based on a distance between the centers of the intervals and the lengths of them. Additional contextual information for this field is the significance of these two main parameters, and as for all fields, significance of the field itself.

```
<fieldContext>
   <field_type>intervalField</field_type>
   <field_significance>…</field_significance>
   <field_calculation_method>…</field_calculation_method>
   <subField_significances>
      <value>…</value>
      <value>…</value>
   </subField_significances>
</fieldContext>
```

In current prototype we are concentrated on resource closeness visualization. Such visualization context implies

user specification of the resource properties significance and existence of additional contextual information for the resources properties (depending on their types). Such contextual information is stored in separate xml file.

```
<?xml version="1.0" encoding="UTF-8"?>
<closenessContexts>
   <closenessContext>
      <closenessContext_id>…</closenessContext_id>
      <closenessContext_name>…</closenessContext_name>
      <calculation_method>…</calculation_method>
      <fieldContext>
       …
      </fieldContext>
      <fieldContext>
       …
      </fieldContext>
       …
   </closenessContext>
</closenessContexts>
```

To simplify the process of a context specification user can use a tool for visual creation and modification of visualization context provided by the 4I GUI Shell.

3. Resource Distance Measuring

Resources compared based on selected set of properties that belongs to specified five field types. General distance (closeness) between to resources based on a list of attributes (properties) is a value from 0 to 1, calculated as a weighed value of all the distances of separate attributes.

$$D(X,Y)=\sqrt{\sum_{\forall i, x_i\in X, y_i\in Y}\omega_i\cdot d(x_i,y_i)^2}, \qquad (1)$$

where $d(x_i,y_i)$ is a distance by certain attribute and ω_i is weight for attributes. The requirement for the weights is:

$$\sum_{i=1}^{n}\omega_i=1, \qquad (2)$$

where n is a number of attributes.

Now we have a problem, we have distances between the resources based on certain attributes separately. The centers of masses those sets are not balanced and differently influence on result. We have to modify the function with a correction part that balances centers of masses.

$$D(X,Y)=\sqrt{\sum_{\forall i, x_i\in X, y_i\in Y}\omega_i\cdot\left(0.5+\frac{0.5\cdot\left(d(x_i,y_i)-d_i'\right)}{\max_j\left|d(x_i,y_i)_j-d_i'\right|}\right)^2}, \qquad (3)$$

where d_i' is *median* or *arithmetic average* of corresponding samples of distances by i -th attribute.

In case of a **Text field (type 1)** we calculates distance based on full matching of string values, and distance takes the values 0 or 1.

In case of a **Text field (type 2)** the distance between two objects based on such a field can be calculated based on following formula:

$$D(X,Y)=\frac{N'}{N}, \qquad (4)$$

where N' is a number of matched/equal instances and N is a general number of all instances in the lists of two comparable objects.

In case of a **Text field (type 3)** we have decided to use well-known PEBLS distance evaluation for nominal values. PEBLS is the Parallel Exemplar-Based Learning System by Cost and Salzberg [13]. This is a nearest-neighbor learning system designed for applications where the instances have symbolic feature values. Let's consider a text field that extended/enriched by set of sub-fields/attributes with values defined on a finite set of possible values. The distance d_j^l between two values v_1 and v_2 for j attribute in a context of attribute l is:

$$d_j^{l\,2}(v_1,v_2)=\sum_{i=1}^{k}\left(\frac{C_{1i}}{C_1}-\frac{C_{2i}}{C_2}\right)^2, \qquad (5)$$

where C_1 and C_2 are the numbers of instances in the training set with selected values v_1 and v_2, C_{1i} and C_{2i} are the numbers of instances from the i-th class, where the values v_1 and v_2 were selected, and k is the number of the classes of instances (from values of attribute l).

Thus, general distance d_j between two values v_1 and v_2 for j attribute is:

$$d_j(v_1,v_2)=\frac{\sum_{l=1,\,l\neq j}^{n}d_j^l(v_1,v_2)}{n-1}, \qquad (6)$$

where n is a number of attributes (sub-fields).

Finally, distance between two objects based on such complex field can be calculated based on following formula:

$$D(V_1,V_2)=\sqrt{\sum_{\forall i, v_1^i\in V_1, v_2^i\in V_2}\omega_i\cdot d_i(v_1^i,v_2^i)^2}, \qquad (7)$$

where $d_i(v_1^i,v_2^i)$ is a distance by certain attribute (sub-field) and ω_i is weight for attributes. The requirement for the weights is:

$$\sum_{i=1}^{n}\omega_i=1, \qquad (8)$$

where n is a number of attributes.

If there is no differences in significance of the attributes, then $\omega_i=1/n$ and

$$D(V_1,V_2)=\sqrt{\frac{\sum_{\forall i, v_1^i\in V_1, v_2^i\in V_2}d_i(v_1^i,v_2^i)^2}{n}}. \qquad (9)$$

If some attributes (sub-fields) are not specified, it means that there is no information in the current text field that

concerns those attributes. The distances between values of such attributes are - "0".

In case of a Number field, distance measuring is bases on normalization all of the values from whole sample of them. The formula of a distance between two values v_i and v_k is:

$$d(v_i, v_k) = \left| \frac{v_i - v_k}{v_{\max} - v_{\min}} \right|, \qquad (10)$$

where $v_{\max}$ and $v_{\min}$ are the maximum and minimum values from the sample.

In case of an **Interval field**, we have decided to focus on main aspects of time periods comparison: durations of the periods and distance between the intervals. And the simplest formula that implicitly take into account these parameters (see Figure 2a) is:

$$d([a_i, b_i], [a_j, b_j]) = \frac{D - r}{D_0}, \qquad (11)$$

where
$$r = \frac{r_1 + r_2}{2} = \frac{(b_1 - a_1) + (b_2 - a_2)}{2},$$
$$D = \max(b_i, b_j) - \min(a_i, a_j),$$
$$D_0 = \max_{\forall p}(b_p) - \min_{\forall q}(a_q).$$

If we are going to explicitly control the influence of interval attributes and to tune their significances, then we can use another more complex formula. For such intervals distance measurement we are going to use formula that takes into account the distance between the intervals' centers and difference between the lengths of the intervals (see Fig.2b):

$$D(t_i, t_k) = \sqrt{k_m \cdot \left(\frac{|m_i - m_k|}{M} \right)^2 + k_l \cdot \left(\frac{|l_i - l_k|}{l_{\max}} \right)^2}, \qquad (12)$$

$$D(t_i, t_k) = \sqrt{k_m \cdot \left(\frac{\Delta m}{M} \right)^2 + k_l \cdot \left(\frac{\Delta l}{l_{\max}} \right)^2}, \qquad (13)$$

where m_i and m_k are the values of the intervals' centers on a time line, l_i and l_k are the durations/lengths of the intervals, M and $l_{\max}$ are maximum distance between the centers of two intervals and maximum duration/length of interval from a sample set. The coefficients k_m and k_l regulate significance of the distance between the intervals and difference between the lengths of the intervals. The condition for those coefficients is:

$$k_m + k_l = 1, \qquad (14)$$

But still, depending on sample, $\frac{\Delta m}{M}$ and $\frac{\Delta l}{l_{\max}}$ can differently influence on a result, even if the coefficients will be equal. We have to modify the function with a correction part that balances centers of masses.

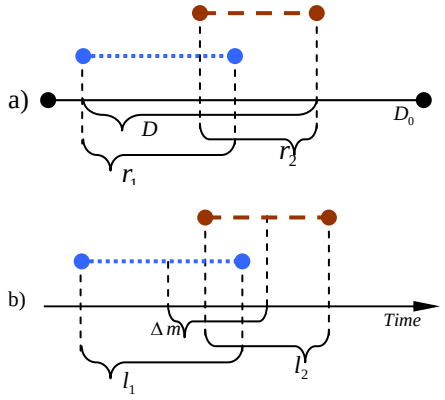

Fig. 2 - The intervals comparison: a) type1 and b) type2.

$$D(t_i, t_k) = \sqrt{k_m \cdot \left(0.5 + \frac{0.5 \cdot \left(\frac{\Delta m}{M} - m' \right)}{\max_j \left| \frac{\Delta m}{M}_j - m' \right|} \right)^2 + k_l \cdot \left(0.5 + \frac{0.5 \cdot \left(\frac{\Delta l}{l_{\max}} - l' \right)}{\max_r \left| \frac{\Delta l}{l_{\max}}_r - l' \right|} \right)^2}, \qquad (15)$$

where m' and l' are medians or arithmetic averages of corresponding samples.

It is important to mention that intervals as values of some attributes of a resource have a lot of meaning in various applications, e.g. in electronic commerce. Very often a customer cannot exactly specify the properties' values of a desired product (e.g. "I want a car with age from 5 to 7 years and mileage from 20 000 to 30 000"). Similarity among such kind of uncertain orders among themselves or against description of available products can be done by utilization of the above formulas.

Other possible distance measuring methods: To find the distance between two terms, it is also possible to utilize a dissimilarity measure, called Normalized Google Distance (NGD), introduced in [14]. NGD takes advantage of the number of hits returned by Google to compute the semantic distance between concepts. The concepts are represented with their labels which are fed to the Google search engine as search terms. Given two search terms x and y, the normalized Google distance between x and y, $NGD(x, y)$, can be obtained as follows

$$NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log M - \min\{\log f(x), \log f(y)\}} \qquad (15)$$

where $f(x)$ is the number of Google hits for the search term x, $f(y)$ is the number of Google hits for the search term y, $f(x, y)$ is the number of Google hits for the topple of search terms x, y and M is the number of web pages indexed by Google (*Currently, the Google search engine indexes approximately ten billion pages - M~10^{10}*).

Intuitively, $NGD(x, y)$ is a measure for the symmetric conditional probability of co-occurrence of the terms x and y: given a web-page containing one of the terms x

or y, $NGD(x, y)$ measures the probability of that web-page also containing the other term.
If we are dealing with nominal attributes that have values defined on an infinite set of values, it is also reasonable to utilize existing remote services that measure Semantic Relatedness of the strings. One of those services is Measures of Semantic Relatedness (MSRs - *http://cwl-projects.cogsci.rpi.edu/msr/msr-about.html*). MSRs are computational means for calculating the association strength between terms. More specifically, MSRs take the form of computer programs that can extract relatedness between any two terms based on large text corpora. MSRs have been used to produce models of human web-browsing behavior [15], augmented search engine technology [16], semantic relevancy maps [17], essay-grading algorithms for ETS [18], and could be useful for any cognitive models or AI agents that have to deal with text.

Lack of Standardization in MSR Services: Although there are multiple MSR services that are readily available to the research community, these services are (1) scattered and (2) inconsistently formatted. All of the available MSR web servers use different input/output standards, making it less than ideal for researchers that may want to compare, contrast, alter, and average these measures. Some MSRs are available to download, but these technologies are even more diverse in protocol, and are much harder to use. To make matters worse, many MSRs are not publicly accessible, and of the available MSRs, very few parameter sets (e.g. different corpora, different sensitivity parameters) are offered. For example, ICAN [19] is a well-founded MSR that one may implement, but no public ICAN service exists. PMI is a popular and easy-to- implement measure, but you would be hard-pressed to find a PMI service based on a news corpus, or an email corpus, etc. The MSR Web Server is an ongoing effort to gather various MSRs and corpora, to make these publicly available, and to give researchers easy standardized access to semantic relatedness scores from all MSR-corpus pairs.

4. Similarity/Closeness resource visualization

Talking about visualization techniques, we have to consider usability issues that bring user-friendly information representation. What is the closeness of the resources? It is a value from 0 to 1. The easiest way to show the distances between resources is to present them on a line. But in case, when we visualize resources by their visual representatives (images / resource logos), we have to take into account the sizes of the images and such line-based representation becomes not so convenient any more. It is also possible to show the compared resources on a circle or sphere with different radiuses (distances to the correspondent resource). Again, in this case, it is quite difficult to see the difference between the distances of two resources to the initial one, especially if they are located on the opposite sites from the center (initial resource).

Taking into account all this nuances, we decided to put the resources on a spiral that lies on a surface of the cone (see Fig. 4). The minimal distance between the resources has been taken as a step on an axis/height of the cone. Just that parameter (distance on the axis/height) shows the closeness of the resources. To avoid an overlap (in case of a viewpoint from the top of the cone) of the images that belong to resources located next to each other, we have calculated the location angle (α) on each (step-based) cone cut (see Fig.3). Additionally, we provided a possibility to rotate the cone to find the best view point (see Fig.4).

We added possibility to create new, delete and modify the similarity contexts. Such visualization context implies user specification of the resource properties significance and existence of additional contextual information for the resources properties (depending on their types). As we can see from the field type's descriptions (section 2), configuration of the resource similarity context is specification of significances of resource properties/fields, subfields (in case of complexTextField and intervalField) and distance calculation method if there are several of them (for the moment we have three calculation methods for intervalField and three methods for the general resource closeness/similarity measurement). Via interface for such context configuration/personalization user has the access to the context parameters that can be changed and has a possibility either to edit currently selected context or to create a new one.

The distance calculation methods utilize coefficients that have the values in diapason [0..1]. To simplify the interface and make it more user friendly, we decided to consider the "*absolute significance*" of the resource fields as percentages from the full influence of the fields. In this case the sum of the fields' significances should be equal 100%. The same approach has been applied for the sub fields if there are any. For the "absolute significance" system supports two modes:

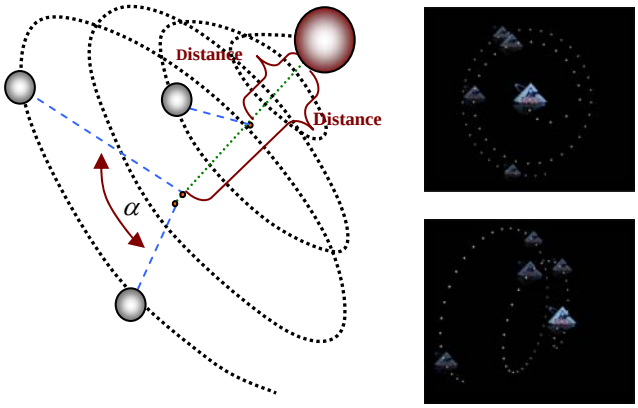

Fig. 3 - Spiral-based distance representation.

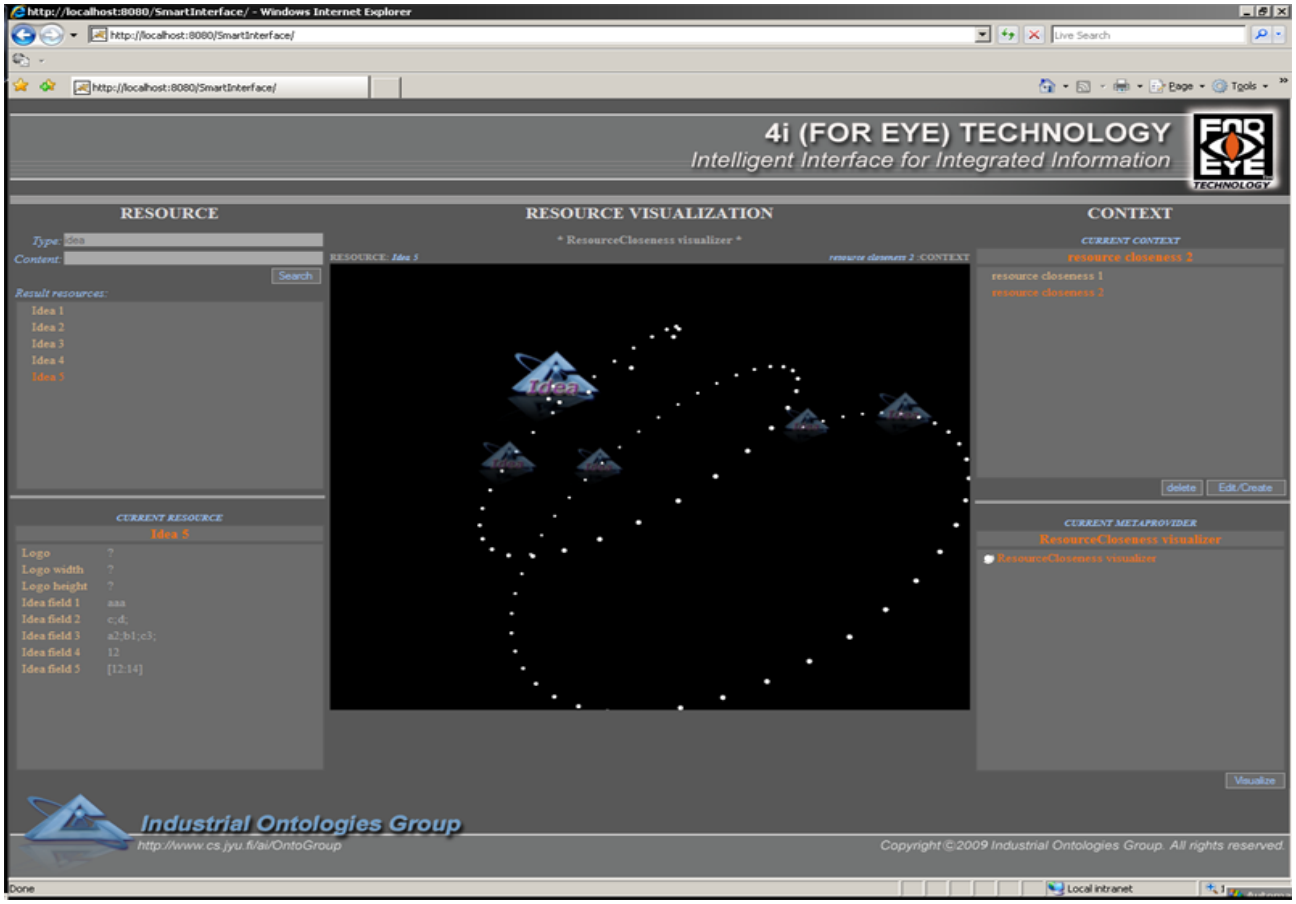

Fig. 4 - Resource closeness visualization in 4i (FOR EYE) Browser.

- *fully user controlled mode*: In this mode user can set any level of significance he/she wants. But at the same time, he/she takes a responsibility to check the condition that sum of the correspondent fields equals 100. To enter this mode user should uncheck the "recalculation" checkbox.
- *mode with automatic recalculation of the significances*: This mode will be useful in case when user change just the value of one field and do not want to change the values of other fields to fit the necessary conditions. In this mode system automatically recalculate the values of other fields proportionally to the previous values.

Sometimes it becomes difficult to define the significance for all the fields in percentages, and user prefers to specify "*relative significance*" for the field/property. In this case user estimates the significance of each field/property by value from 0 to 100 separately. With the "relative significance" the absolute values do not make sense, only comparative differences of the values are taken into account. Further system itself transforms these values to the "absolute significance" and user can play with percentages later on if he/she wishes.

As ever user saved the changes or created a new resource similarity context, MetaProvider that provides visualization module creates and sends a new visualization to the 4I Browser accordingly to the new configuration.

5. Conclusion and future work

Today similarity-based search is used in numerous fields of applications. This paper presents an approach towards resource similarity/closeness distance measuring and similarity/closeness-based resource visual browsing.

We distinguish between similarity search within a particular class of instances and closeness search for instances of different classes. We also argue that there is sense to talk about similarity or closeness only if the context (e.g. why similarity/closeness is measured) is defined. We utilize a context by its influence to the importance of certain attributes of the compared instances. We provide similarity measures for several types of a resource attributes and the way how separately measured component similarities can be normalized and combined to integrated similarity (closeness) measure. We also present an approach on how to present similarity/closeness search results to the user.

We consider that a user have to have an opportunity to utilize a similarity/closeness search tool as a kind of browser for resources in semantic space, i.e. able to easily manipulate the context filter, get, see and access the resources ordered by distance from the selected one in a natural way.

Elaborated prototype of such a browser shows us potential of the approach and can be used in different application areas (e.g. electronic commerce, social Web applications, measuring intangibles, etc.) to simplify a way of information retrieval in human-oriented manner.

In current prototype we have not concentrated our efforts on development of distance measuring functions that depends on remote services. It always makes system more unreliable. But still, these methods can be utilized as well, if the goal can not be achieved in another way. As a future work, we still are going to increase a number of distance measuring methods and types of compared resource description fields. Also we are planning to integrate presented similarity distance measuring functions with ontology-based distance measuring approaches to elaborate more precise similarity function.

Acknowledgements

This research has been performed as part of UBIWARE project in Agora Center (University of Jyvaskyla, Finland) funded by TEKES and industrial consortium of Metso Automation, ABB, Fingrid, Inno-W and Hansa Ecuras. We are very grateful to the members of "Industrial Ontologies Group" for fruitful cooperation and Inno-W Company for useful industrial case.

References

[1] D. R. Wilson & T. R. Martinez, Improved Heterogeneous Distance Function, *Journal of Artificial Intelligence Research*, 6, 1997, 1-34.

[2] T. Yamada, K. Yamashita, & N. Ishii, Text Classification by Combining Different Distance Functions with Weights, *Proceedings of the Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD'06)*, IEEE CS Press, 2006, 85-90.

[3] V. Terziyan, Predictive and Contextual Feature Separation for Bayesian Metanetworks, *Proceedings of KES-2007 / WIRN-2007*, Springer, LNAI 4694, 2007, 634–644.

[4] M. Dash & H. Liu, 1997. Feature Selection for Classification. *Intelligent Data Analysis* 1 (3), 1997.

[5] M. Refaat, *Data Preparation for Data Mining Using SAS*, Academic Press, 2007, 399 pp.

[6] A.P. Sheth & C. Ramakrishnan, Semantic (Web) Technology in Action: Ontology Driven Information Systems For Search, Integration and Analysis. *IEEE Data Engineering Bulletin*, 2003.

[7] M. Davis, 2004. The Business Value of Semantic Technologies, Presentation and Report, Semantic Technologies for E-Government, September 2004.

[8] O. Khriyenko, 4I (FOR EYE) Technology: Intelligent Interface for Integrated Information. In *Proc. of the 9th International Conference on Enterprise Information Systems (ICEIS-2007)*. Funchal, Madeira – Portugal, 2007.

[9] O. Khriyenko, Context-sensitive Multidimensional Resource Visualization. In *Proc. of the 7th IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP 2007)*. Palma de Mallorca, Spain, 2007.

[10] S. Castano, A. Ferrara, S. Montanelli & G. Racca, Matching Techniques for Resource Discovery in Distributed Systems Using Heterogeneous Ontology Descriptions. In *Proc. of the Int. Conference on Coding and Computing (ITCC 2004)*, IEEE Computer Society, Las Vegas, Nevada, USA, April 2004.

[11] H. Chorfi & M. Jemni, PERSO: A System to customize e-training. In *Proc. of 5th International Conference on New Educational Environments*, May 26-28 2003, Lucerne, Switzerland.

[12] H. Jing & E. Tzoukermann, Determining semantic equivalence of terms in information retrieval: an approach based on context distance and morphology. Book chapter *in Recent Advances in Computational Terminology*. John Benjamins Publishing, 2001.

[13] S. Cost & S. Salzberg, A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10(1):57–78, 1993.

[14] R. Cilibrasi & P. Vitanyi, The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3), 2007, 370–383.

[15] P. Pirolli, Rational analyses of information foraging on the Web. *Cognitive Science*, 29(3), 2005, 343-373.

[16] S. Dumais, Data-driven approaches to information access. *Cognitive Science*, 27(3), 2003, 491-524.

[17] V. D. Veksler & W. D.Gray, Mapping semantic relevancy of information displays. *Paper presented at the CHI 2007*, San Jose, CA, 2007.

[18] T. K. Landauer, P. W. Foltz & Laham, D., Introduction to latent semantic analysis. *Discourse Processes*, 25, 1998, 259-284.

[19] B. Lemaire & G. Denhiére, Incremental construction of an associative network from a corpus. In *K. D. Forbus, D. Gentner & T. Regier (Eds.), 26th Annual Meeting of the Cognitive Science Society, CogSci2004*. Hillsdale, NJ: Lawrence Erlbaum Publisher, 2004.