Identity: How To Name It, How To Find It

Christo Dichev
Winston-Salem State University
601 Martin Luther King, Jr. Dr.
Winston Salem, N.C. 27110
01-336-750-2477
dichevc@wssu.edu

Darina Dicheva
Winston-Salem State University
601 Martin Luther King, Jr. Dr.
Winston Salem, N.C. 27110
01-336-750-2484
dichevad@wssu.edu

Jan Fischer
University of Karlsruhe
Kaiserstraße 12
Karlsruhe, BW 76131
01-336-750-2484
uatc@stud.uni-karlsruhe.de

ABSTRACT

The main objective of this work is to exploit the relationship between the *information findability problem* and a *subject-based organization* of information. Identification of a subject is involved when one wants to say something about that subject or when he or she tries to comprehend what was said by others about it. An example of this type of duality can be seen in the information world where content creators and content consumers need to communicate. In this paper we discuss the concept of subject identity in learning content authoring, where we view a topic map as supporting the communication between a content author and learners. In this context we address both sides of the dual system and propose some solutions intended to assist both content creators and consumers in dealing with problems typical for e-learning repositories. Concerning the learners who need to identify the subject they are looking information about, we suggest that a set of subjects related to it can be interpreted as a weak form of its identity. This can be used for finding a starting point for content exploration and we propose an algorithm for this task. As to the content authors, they need to use agreed-upon names and possibly subject identifiers to identify the subjects they are talking about. In this relation we suggest using Wikipedia articles as a source for both consensual naming and subject identifiers. We claim that Wikipedia can play a role of a shared context between topic map authors and users and propose an approach for extracting consensual information from Wikipedia. The proposed ideas are implemented in the Topic Maps for e-Learning tool (TM4L).

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation]**: Group and Organization Interfaces, *Web-based interaction Collaborative computing*.

General Terms

Algorithms, Human Factors

Keywords

Subject Identity, E-learning, Topic Maps, Information Retrieval, Semantic Web,

1. INTRODUCTION

Technology impacts our lives in many different ways and the impact of the web specifically is pervasive: it changes the way we find information, we absorb information, and we organize information. By providing an enormous amount of online data, it offers an unlimited repository for searching learning materials however the search results contain an excessive amount of noise. It is recognized that while search engines are very useful in retrieving online information, there are a lot of unsolved problems related to their effectiveness. For example, Google returns 1,050,000 resources matching the keywords "prolog, lists" (Fig. 1). Another common weakness is that search engines don't present the information in the manner people want to receive it – by subject matter. The amount of results of a search engine query like the one above, while impressive is not useful in its full volume, since the presented links are not organized thematically and there are no structural cues.

From the viewpoint of e-learning information support, the learners have typically only some idea of what they need to know. They may or may not know how to articulate it and, if they do, what are the right keywords to use. Additionally, they may not know where to start to look from. This implies that a keyword-based search is not very effective for satisfying learners' information needs. Navigational structures, such as topic maps, where the resources are grouped around particular subjects demonstrate greater potentials. Topic maps (TM) [3] are fundamentally about linking related items together and hence can serve as a navigational structure and interface to a learning repository.

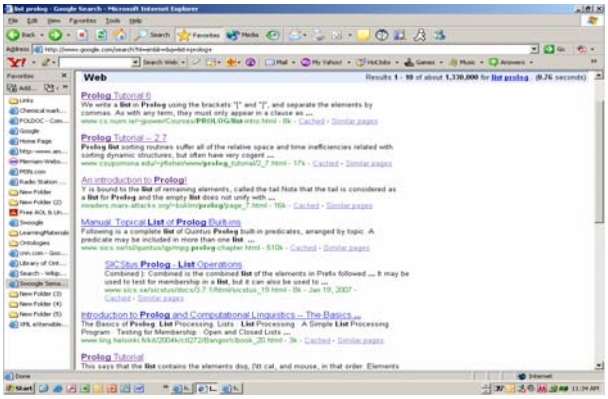

Figure 1. Google search result for 'prolog lists'.

In the information world a subject is typically documented. Therefore, we are interested in finding some documents with learning value about the subjects of interest, which have been published or collected in a retrievable form. In terms of information support for learning tasks, we use a subject to indicate an entry point into a collection for searching documents describing the subject. Gathering documents and data about subjects implies a possibility for identifying and representing

WWW 2007, May 8-12, 2007, Banff, Canada.

them. However, it is not possible in general to describe a subject identity as a single, unambiguous set of characteristics. As we have learned from experience, particularly when designing learning collections, there is no unique way to describe concepts and sometimes there are no convincing reasons to decide whether one particular way should be considered better than another. Practical confirmation of this fact can be found in the existing textbooks, tutorials, and lecture notes, where different ways can be seen of speaking about and referring to the same concepts [2].

The work presented here was driven by an attempt to address the problem of providing an easy and intuitive access to a large volume of heterogeneous learning resources and to find an answer to questions of the type: How can a learner be directed in a principled way from a given tutorial to other relevant tutorials about this subject, to other related subjects and tutorials, articles, and notes linked to them, matching the learners' level of knowledge and current task? The major aspects of this work are centered on enhancing TM4L - an e-learning environment providing editing and browsing support for developing and using Topic Maps-based e-learning repositories [1]. TM4L utilizes topic maps as overlay semantic structures that encode domain knowledge and connect it to learning resources, which are considered relevant to that domain.

The purpose of TM4L is to enable instructors to create topic maps for a given course easily. The topic maps are intended to provide students with an intuitive interface for finding information of interest, related to course tasks such as course projects or nontrivial assignments. The underlying intuition of this work is that the key to solving the *information findability problem* is a *subject-based organization* of information. For e-learning repositories with subject-centered architecture, each concept is a *hub for resources* possibly grouped by additional characteristics. For example, depending on their type, the resources can be classified into online notes, definitions, descriptions, articles, code examples, exercises, PPT presentations, lecture notes, handouts, slides, exam questions, quizzes, etc. However, when the collection of resources grows, it becomes hard to find needed information if the topics are not well organized and properly named. The latter may defeat the advantages of the subject-centered organization of resources. This is especially true for users who are not familiar with the subjects and terminology used (e.g. students).

The task of exploration assumes a starting point of the browsing process; however, users often find it difficult to decide where to start the browsing from. In the case of directories, the adopted approach is to start from the top of the hierarchy, narrowing down the search domain through a number of successive choices. The focus in such applications is on a generic representation that can be reused by a maximum number of users. This type of generality sacrifices the ability to reflect users' or groups' specific viewpoints. The fact that different users can have different perspectives and that these perspectives affect the way they see the world, requires the system to allow for representing and organizing the resources, based on the users' tasks and other possible contextual factors. In this work, perspectives and contextualization are addressed in terms of Topic Maps-based information support. The goal is twofold: enhancing users' navigation support and assisting users to quickly find an appropriate starting point for exploring relevant information. In particular, when browsing an educational topic map, effective support for locating a good starting topic can play a key role in finding the needed resources.

Our view on subject identifiers (and names) is not of absolute and universal identifiers but of an assertion about domain concepts, agreed-upon by a community within a particular period of time and with a particular purpose, that is, within a shared context. We need a shared context to communicate and understand each other. The shared context provides a common conceptual ground, a shared framework and a name space for communication. We came to a conclusion that Wikipedia can play a role of a shared context between topic maps' authors and users.

To support such functionality, we extended TM4L to provide means for harvesting consensus information from Wikipedia in order to assist users in naming, relating or identifying subjects. Here we understand subject in its broad sense, as a carrier/hub of information objects whose purpose is to describe it from different perspectives and contextual assumptions. We believe that utilizing Wikipedia in providing information support for e-learning tasks can bring benefits to both content creators and content users. The key benefit is that Wikipedia can serve as a domain context, which is an important component in communication between TM authors and between TM authors and learners. In addition, it can provide a rich pool of consensual topic names, topic subject indicators and topic subject identifiers that would simplify the development of domain-specific ontologies.

In contrast to the mainstream approach, where the focus is on a machine readable agreement about subject identity, our focus is on a human readable contract on subject identity. It is based on the fact that in the intended applications, a human is the ultimate arbiter of the subject identity. The identity is what the author of a topic map had in mind when the topic was created, and on the other hand, what the learner has in mind when using the topic map.

Identification of a subject is involved when one wants to say something about that subject or when we try to comprehend what was said about it. An example of this type of duality can be seen in the information world where content creators and content consumers need to communicate. In the area of learning content authoring, we view a topic map as a form of communication between a content author and learners. From this viewpoint, we attempt to analyze the different aspects that subject identities and their names in particular can play in organizing e-learning repositories. The focus is on interchange of information between humans through machines. In this context we address both sides of the dual system and propose some solutions intended to assist the content creators as well as content consumers in dealing with problems typical for e-learning repositories.

The paper is organized as follows. In Section 2 we present our view on subject identity and its implication to finding relevant information from human perspective. Section 3 presents our methods for providing users with starting topics for their exploration. In Section 4 we present our approach for extracting consensual information from Wikipedia and Wikibooks and Section 5 discusses relevant work. A concluding discussion is included in Section 6.

2. SUBJECT AND IDENTITY

In the information universe, a subject is traditionally identified by its name (title, code, ID) or if it is an addressable resource - by its

location. There are some other ways to identify a subject that are not sufficiently explored. For example, a subject can be identified by its unique relations to other subjects or properties. If we hear the terms "Horn Clause Logic", "Unification" and "Colmerauer", our guess would be that the discussion is about Prolog. On the other hand, if the terms we heard were "find-all", "bag-of", and "set-of", the guess would be that the talk was about Prolog's extra logical features. Similarly, if we observe the terms "Process Management", "Memory Management" and "File Systems", a reasonable prediction would be that this is a discussion about Operating Systems. But if we perceive the terms "Resident Set", "Page Replacement", "Thrashing", and "Demand Paging", we would guess that the discussion is on Virtual Memory. These and similar examples led us to a hypothesis that can be interpreted as a variant of *Occam's razor:* reflecting a representation of a domain, a perceivable collection of subjects evokes the *minimal structure* containing all perceived subjects. Thus, if for a list of subject names we are able to find the minimal structure containing them, this list of names can be used as a weak form of subject identification (assuming that the algorithm for computing the minimal structure is able to identify "the center" of the structure). An interesting side effect of this form of identification is that the minimal structure identified by the entry list of topic names can be used as a start point for further exploration.

For identifying and referring to subjects we use names (labels, identifiers). Names are used as a means of differentiating between subjects and also as a means of referring to and locating subjects. What these subjects are is not disclosed by the names. Name is a device that facilitates communication. What if we don't know or can not recall a name? Humans are able to deal with such a situation. We can ask for "the blue book with decision tree examples that was used in our AI class", when we can not recall the title "Prolog Programming for Artificial Intelligence" or for "that bolt about 4 inches long, with one of those thingies on the top", when we can not remember "part P734-9". The idea of enabling a similar pattern of interaction between humans and machines is appealing, especially in the area of e-learning where the inability to name a resource is not an exception.

The two concepts, subject and context, are closely related. We can reach an agreement on a subject's identification only in a given context. There is no universal way to assert the identity of something or someone. This implies that any subject identification needs a context. Contexts and therefore subjects can not be defined and bounded in any absolute sense. They can be defined only relatively, with respect to a particular domain (closed world). For example, the subject "*good weather*" may have different interpretation from two individuals - one from Alaska and another from Hawaii. On the other hand, concepts such as "Horn Clause Logic", "Unification", and "find-all" have no meaning and no identity to any Prolog-ignorant individual.

In an e-learning application with information support orientation, the learners need guidance in their seeking for learning resources. If the provided help supports different perspectives, such as: topic-based, knowledge level-based, task-based, etc., this can reduce the complexity of the presented information. The problem is that such a structured approach is never complete. For example, in addition to beginner, intermediate, and advance level materials, some learners might want to view this information structured further by academic terms (Fall 2006, Spring 2007, etc.), others may prefer to see it structured further by universities, yet some other users may want to see it divided in handouts, assignments, code examples, etc. Therefore , it is infeasible to assume that we can build a representation of a context and thus of a subject that is absolute and universal. Contexts have an infinite dimension hence they cannot be described completely.

In practice, when the entries in a collection become too many, we can add additional relationships/dimensions to the entries to differentiate them – a pragmatic approach to differentiation of information. However, this process has pragmatic limits. Since we cannot describe the subjects universally and completely, it is impossible to create a subject reference in a universal way. Within a collection of partially described entries, the user can locate the needed information by browsing.

Knowledge is created based on personal viewpoints and is interpreted in a particular context. This implies a biased view on subjects – one subject can be interpreted/identified differently by two independent agents. Since we can not describe a subject completely, it is impossible to make a subject and its description absolutely identifiable. There will always be a case when subject descriptors have to be interpreted in a particular context for deciding on the subject meaning - this induces the need of browsing. Thus browsing is unavoidable in absolute sense.

3. START POINTS FOR EXPLORATION

Finding a good starting topic is a critical part of browsing a topic map. Obviously, it is worth starting from a topic that allows reaching a large amount of relevant learning content with a few clicks. We propose an algorithm for selecting a starting point. It takes as an input (provided by the user) a set of topics (*entry topics*), and outputs a collection of topics qualified as *starting topics* for topic map exploration. The latter are found through their relationships with the provided entry topics. An intuitive example supporting our approach can be given in terms of Physical maps, Political maps, and Road maps. If members of a group observing a map mention Boston, Dallas and Seattle, we can conclude that most likely they are talking about the US. If the conversation includes also Toronto, then it is likely that they are talking about the US and Canada. If they mention in addition Guadalajara and Monterrey, then most probably the people are talking about North America. If we use these city names as a query mechanism for suggesting a good starting region for geographic exploration, then a query such as "Boston, Dallas, Seattle" should return the US as a starting region for exploration. The starting region alone is not always sufficient to specify the exploration: we need to choose the right type of map as well. When this information cannot be inferred from the query, we have to explicitly request the type of map that matches our task, e.g. Road, Political, Historical map, etc. In some cases the map scale can be also important in identifying the place being sought. Yet another concern can be the year of map publishing, etc.

The TM search can be improved if we switch from keyword search to more semantic driven search, combined with subsequent browsing. This implies returning not just a set of resources containing the keywords but placing the user in a relevant location, i.e. at a starting point for further exploration for resources. We extended TM4L functionality with a Topic Map version of such a *query-initiated-navigation*.

3.1 Minimal Structures

There are different ways to specify a subject. In our case we are interested in information objects (articles, tutorial, handouts, etc.)

describing the subject. When the users are able to describe their exploratory interest in terms of related subjects, it would be helpful to provide them with assistance in the form of a navigational strategy for the area of exploration. If a graphical representation of the topic map is available, showing how the topics are related and grouped, it is easier for a user to get oriented in selecting his or her exploration strategy. From the viewpoint of conventional navigation approaches, the web portals/directories can be interpreted as sites providing a static "main page" as a fixed starting point for the area of exploration and browsing. Pushing this analogy further, we can imagine a portal with a dynamic "main page" adapted to the user by showing only the links to the pages relevant to the user's specified needs. Using this analogy, our goal is to develop an interface that is able to adjust the portal's entry page depending on the user's needs declared in terms of a list of entry topics.

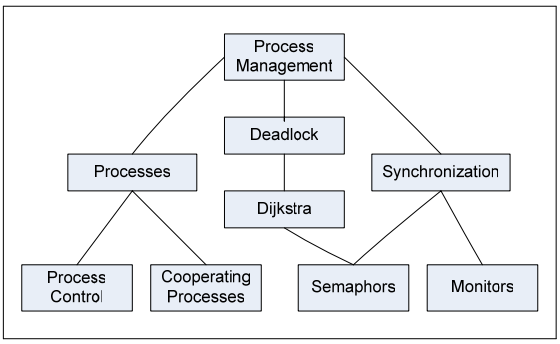

Figure 2. Partial topical structure of *Process management*

The illustrated features of browsing conventional maps can be used as a bridge to analogical features in topic maps. The analogy becomes obvious if we make the cities in conventional maps correspond to topics in topic maps, while geographical regions correspond to sets of topics considered as units. Different types of geographic maps may correspond to topic maps covering different types of learning material, for example, beginner, intermediate, advance level, text notes vs. program code, etc. From an implementation point of view, such functionality would require appropriate interface, where the user can specify (select) the relevant contextual parameters (e.g. beginners, program code, etc.) and then within the proposed context-based region find the needed topic (which is the subject proxy). Based on this analogy, consider the sketchy topic map presented in Fig. 1, assuming that it represents intermediate level learning content. Assume further that in an interactive mode the user submits a sequence (list) of topics, intended as an initial entry for computing the starting point for browsing. Following Fig. 2, for an input list {"Semaphores", "Monitors"} the user will be presented with the topic map segment containing the topics "Synchronization", "Semaphores" and "Monitors" as a starting point for exploration, that is, with the minimal unit (sub-graph containing the topics "Semaphores" and "Monitors"). For the list of topics {"Process Control", "Synchronization", "Monitors"} the user will be presented with the starting list {"Process Control", "Processes", "Process Management", "Synchronization, Monitors"}, etc.

3.2 Identifying Starting Points

Topic maps are essentially interconnected graphs with (potentially) many dimensions of metadata. Therefore, a topic map can be represented as a graph *G(T, A)* of topics *T* and associations *A*. Using this graph representation, the task of finding a starting point for exploration based on a user's entry topic list can be formulated in terms of finding the *minimal sub-graph* containing the list of the entry topics.

More precisely, given a graph *G(T, A)* the aim is to identify a sub-graph G_m of *G* that meets the following conditions:

1. G_m contains all nodes from the *Entry* list (the user input).
2. G_m should be minimal, that is, should contain as less nodes as possible.
3. G_m should be connected (if possible).

In the following description of the algorithm we denote by *Trv = Traversed(T) = (T1, T2, … , Tk)* the set of all topics T_i which are directly associated by any association *A* to topic *T*; that is, each T_i is a neighbor of *T* with respect to an association A.

The algorithm maintains the following data structures: *Path(start_node, end_node, length, path*) is an object that stores a path between two nodes as well as its length, start node, and end node. *Input* is a list that holds the user input and remains unchanged throughout the execution of the algorithm. *Entry* stores a modifiable copy of *Input*. *Open* is a list of topics in the topic map, which are yet to be examined. *Closed* is a list of topics that have been already examined. Similar to Breath First Search (BFS), the *Open* list acts like a queue. This has a danger that the search space may be too large. Thus a depth limit is placed to prevent this. Accordingly, *d(j)* denotes the depth of node *j* and *p(j)* denotes the predecessor of node *j* within the BFS search tree.

1. **FOR** each topic t_i in *Input* (*1 <= i <= #Input*) **DO** BFS
   1. **Initialization.** Copy all entries $\{t_i\}$ from *Input* to *Entry*. Set *Open* = $\{t_i\}$ and *Closed* = $\{t_i\}$. Initialize $p(t_i) = t_i$ and $d(t_i) = 0$. **FOR** all other topics (nodes) t_j set their depths and predecessors to undefined $p(t_j) = d(t_j) = undefined$.
   2. **Algorithm Body**. **WHILE** *Entry* and *Open* contain at least one element, **DO** :
      1. Take the head t_h of *Entry* and delete t_h from *Entry*.
      2. **IF** $d(t_h) > depth\ limit$ **THEN EXIT**.
      3. **FOR** all topics t_j in *Traversed*(t_h) and not in *Closed*, set $d(t_j) = d(t_h) + 1$, $p(t_j) = t_h$ and add t_j to *Open* and *Closed*. If t_j is in *Entry*, delete t_j from *Entry* and create a *Path* object $Path_{i,j}$ that contains all nodes on the *Path* from topic t_i to topic t_j. The path can be determined by following the predecessors of t_j
2. Out of all paths $Path_{i,j}$, create a weighted graph *G=(V, E)* with *V=Input* and $e_{i,j} \in E$ iff $Path_{i,j}$ $(i,j \in V)$. The weight of edge $e_{i,j} \in E$ corresponds to the length given by the number of nodes, of $Path_{i,j}$. Therefore, an edge between two nodes represents the shortest path between these two nodes and the weight of an edge is its length within the original graph (topic map). If there is no edge, then no path has been found between these two nodes.
3. Finally, calculate the minimal spanning tree G_{min} out of G with Kruskal's Algorithm and replace the edges of G_{min} with the path given by its corresponding $Path_{i,j}$ object. G_{min} is the output of the algorithm that represents the minimal topic map structure based on the user's input.

Note that Kruskal's Algorithm is an example of a greedy algorithm [4]. The basic idea it is that we start with a new, empty graph and add the edges in order of increasing cost.

Notice also that the output of the algorithm is a set of nodes. If the goal is to present a single node as an output, the algorithm requires one more phase, where one of the nodes from the output set is selected as a *starting node* based on some criteria. The adopted approach is based on the notion *center* of a graph.

Definition. A *center* of a graph is a node *C* such that the maximum distance from *C* to any other node T_i is minimized.

When the center is unique, the algorithm terminates by returning the center of the output graph as a starting point for exploration.

In general, a center of a graph is not unique; moreover, the set of the output nodes may not present a single graph. In such a case we have to apply additional selection criteria. Our implementation is based on the following algorithm:

1. **If** G_{min} is a forest with *k* trees, then select the tree tr_j *(0 < j <= k)* with maximum topics $t_i \in Input$. The tree tr_j connects a maximal subset of topics out of all trees within G_{min} generated from the user's input. The intuition is that tr_j captures the topic that the user had in mind. The identification of tr_j is based on the following strategy:
2. **FOR** all topics t_k in tr_j **DO**
    1. Accumulate the hop distances from t_k to all other nodes within tr_j.
    2. Determine the node(s) t_{min} with the shortest accumulated distance to all other nodes
    3. **IF** t_{min} is unambiguous, return t_{min} **ELSE**
    4. **IF** t_{min} contains more than one node return the node(s) from t_{min} that maximizes the number of nodes reachable with two hops.
    5. **IF**, after applying this criteria, the center is still not unique return the node t_{min} that maximizes the number of nodes reachable with one hop.
    6. **IF** the center is still unique, return the first node from t_{min}

Thus the algorithm ends with identifying the center topic of the *Output* structure of topics.

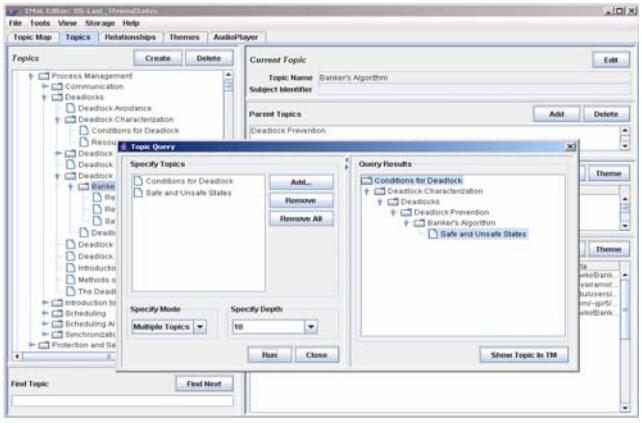

Figure 3. For the Operating System TM and for list of topics {*Conditions for Deadlock, Safe and Unsafe State*} the user is presented the starting list: {*Deadlocks, Deadlock Characterization, Conditions for Deadlock, Deadlock Prevention, Banker's Algorithm, Safe and Unsafe States*}

From an implementation viewpoint, the algorithm terminates based on the user preference: *set of starting nodes* or a *single starting node*. If the user selects the first option, the algorithm terminates by returning the *Output* set. If the user selects the second option, the algorithm completes the second phase and terminates by returning a single node as an output. If the user does not have any other preferences, this node will be suggested as a starting topic for browsing (see Fig. 3).

When we cannot evoke directly the identity of a particular thing, we often recall that thing by recalling some other things related to it or by listing some of its properties. For example, in the course of conversation you may fail to remember the name ML but you may still remember that it is a functional programming language, developed by Robin Milner and that it influenced some newer languages such as Haskell. This information might be enough to convey the subject identity to the other participants in the conversation. Similar facts suggest the possibility of using a set of subjects as a means of determining the identity of another subject uniquely related to them. This intuitive observation about the subject identity can be expressed in terms of the minimal graph used in the above algorithm and satisfying additional constraints.

Let $T_1,T_2,..,T_k$ be a list of entry topics. If the minimal graph containing the entry topics $T_1,T_2,..,T_k$ contains unique topic *T* such that it is related to all the topics in the entry list, then the entry list $T_1,T_2,..,T_k$ can be used as conferring the identity of topic *T*.

This approach allows multiple identifiers, viz. constriction of different sets for a topic's identification. For example, the subject "ML" can be identified by two different lists of topics: {"functional programming", "year of creation 1973"} and {"functional programming", "creator Robin Milner"}.

4. WIKIPEDIA AS A SOURCE OF CONSENSUAL INFORMATION

In general, subject identity enables us to establish which subject is reified by a particular topic. This somewhat problematic task can be simplified when narrowing it to a specific domain such as e-learning. In the area of e-learning, the common sets of subjects are subsets of the conventional academic disciplines. The insight is that in such domains we have an established agreement about the meanings of the topics and associations within the domain. In other words, there is an implied common context aiding the communication between the learning material authors and learners. This kind of agreement is recently promoted by the emerging social and collaborative knowledge construction. This fact encouraged us to use Wikipedia in a topic map construction process. In the context of e-learning repositories' creation, Wikipedia can be used as a source for ontology construction and agreement. Our claim that Wikipedia can be used as a source of shared context is supported by the following observations: Wikipedia is becoming a recognized knowledge repository [8]; more and more pervasive; more and more useful and influential every day; more and more a "trustworthy" source of information to students and the general public.

As we mentioned before, we propose to use Wikipedia articles as a source for both agreed upon names and subject identifiers. The aspect of agreed upon naming is crucial for reliable identification and interchange of information. Therefore, we view names as labels accompanied by an agreement to use them for identifying certain subjects, i.e.

$$name = label + agreement$$

The motivating insight coming from this equation is that Wikipedia can provide both the names plus the agreement grounded on its increasingly widespread use.

4.1 What to harvest from Wikipedia

Marking up existing Web resources with human-readable annotations is a task with its own challenges originating from problems, such as lack of widely accepted names for concepts and relationships. Dublin Core provides a limited vocabulary for expressing some metadata, but doesn't offer support for semantic annotation. The primary components that authors need in creating their topic maps are topic names, intended for humans to grasp the intention of the concepts and relationships. The authors also need relationships, a pool of occurrences, and occurrence types. Finally, they need a source of Public Subject Identifiers (PSIs) if targeting machine-interpretable annotations.

The identification of the key concepts includes assigning names to them. The problems with naming concepts are well known and in e-learning these problems exhibit specific aspects, derived from the vocabularies originating from various textbooks or course syllabus. A multiplicity of naming is observed even in high level concepts, such as "Computer Architecture" vs. "Computer Organization" vs. "Computer Organization and Design". To provide successful mapping we need a pool of consensual topic names that can play a role of primary name (corresponding to the generally accepted term in the field). As to the relationships between concepts, the available pool is limited – mostly, *whole-part* and *class-subclass* relationships. The lack of a sufficient number of created topic maps makes any claims of topic identifiers (especially for human-readable topics) very difficult.

The driving factor in our approach implemented in TM4L is that Wikipedia can provide assistance in the topic name selection, subject indication, subject identity, and partially in relationship selection. It adopts the following strategy: When creating a topic map, for each topic name entered by the author, TM4L indicates if the same name is used by Wikipedia for naming an article or there is no article with such a name. When the entered topic name matches an article, the corresponding Wikipedia page is displayed in a separate window. If the displayed page is in agreement with the intended meaning of the topic, the author can select it as a subject indicator (for human interpretation) and optionally its URI can be used as a subject identifier for machine interpretation.

The user can further ask for a list of concepts relevant to the current Wikipedia article and thus to the entered topic. In response to this request, TM4L displays a hierarchical structure of concepts used in the current Wikipedia page. From this structure the user can select either separate concept names or a substructure of concept names for automatic insertion into the current topic map. When a structure of concepts and their subordinate concepts is selected, they are added to the current TM together with the corresponding *whole-part* relationships (topics as part of their parent topics). Additional relationships can be captured from the summary tables that list some key facts about the subjects. The relations from these tables are transformed into TM format by modifying the original Wikipedia pattern (e.g. *C++ Designed by Bjarne Stroustrup***)** to match the TM standard (e.g. *Bjarne Stroustrup* $\leftarrow$ *subject* $\leftarrow$ *Designed by* $\rightarrow$ *object* $\rightarrow$ *C++*).

The aim is not to translate the Wikipedia content into a Topic Map format. It is rather to reduce the effort involved in TMs construction by reusing the consensual information from Wikipedia, so that the TM author can focus on harvesting and supplying appropriate learning resources.

Technically, the adopted strategy aims at showing in an un-intrusive way the relevant articles from Wikipedia and Wikibooks, along with a hierarchy of related concepts, in response to a request from a user who tries to define a new topic. Given a user input as a sequence of keywords, the idea is to extract from Wikipedia a topic including the set of topics related to it, matching the user input in order to display them and make available for subsequent selection from the user (See. Fig.3). Once the suggested topic is selected, its URL is automatically stored as a subject identifier for that topic.

4.2 Extracting Topics

In order to extract topics, the TM4L crawler crawls Wikipedia and Wikibooks, starting with the name entered by the user. Thus the result consists of two collections of topics – one from Wikibooks and one from Wikipedia. When the crawling process is completed both collections are merged. In general, the Wikibooks site provides better tables of contents than Wikipedia. On the other hand, some topics are not covered in Wikibooks since its content is limited. To fill such gaps we analyze Wikipedia articles, since it contains more complete information. The combined result from Wikipedia and Wikibooks provides a better coverage of the topics.

The idea is to propose to the user a pool of topics (with consensual topic names) semantically related to the topic currently considered by him. For e-learning, occurrences should be split based on the learners' level of knowledge.

4.3 Implementation

The TM4L Topic Extraction plug-in is written in Java and uses WebSphinx, an open source web crawler.

Technically, Wikipedia and Wikibooks articles are XHTML documents allowing processing by XML-based user agents. Therefore, in implementing the extraction engine we employed Java API for XML Processing (JAXP) to extract topics. JAXP supports the Document Object Model (DOM) API. The adopted strategy exploits the DOM object tree structure where each node contains one of the components from the corresponding XML structure. For each topic found, the extraction engine creates a node in a memory-resident tree. The names (labels) of the tree nodes are checked for occurrence in the current topic map. The constructed tree is then presented to the user for selecting the topics he or she wants to be included in the opened topic map.

Choosing the web pages to be searched. The problem of using directly the standard URLs of Wikipedia and Wikibooks articles is that an article with a required title may not exits or a different format of addressing may be used. For example, the article 'Computer_Architecture_Lab' is located at http://en.wikibooks.org/ wiki/Wikiversity:Computer_Architecture_Lab instead of the expected http://en.wikibooks.org/wiki/ Computer_Architecture_Lab. We have solved this problem by using the Wikipedia search engine that computes the relevance rate of the searched documents (result links). The relevance rate helps also in the case of misspelling a word: the search engine returns the link of highest relevance rate.

The majority of articles contain a table of contents (TOC) aimed at supporting easy page navigation. Although Wikipedia and Wikibooks apply different type of page structuring, several tag and stylesheet rules can be found. We also had to figure out and utilize the rules defining TOC in the web pages. Another consideration is that in Wikipedia and Wikibooks, the significant concepts are typically linked to some other pages. It is hard to extract all interesting concepts from a plain text but we can collect the keywords, which have links and put the linked concepts appropriately nested in the topic structure. Thus the Topic Extractor first tries to find out the TOC of the currently

processed page. In case it is challenging to locate it by parsing, it searches for section titles and links (anchor tags). When the tree structure is created, the concepts of each node are collected. Figure 1 shows the process of the hierarchical topic structure generation.

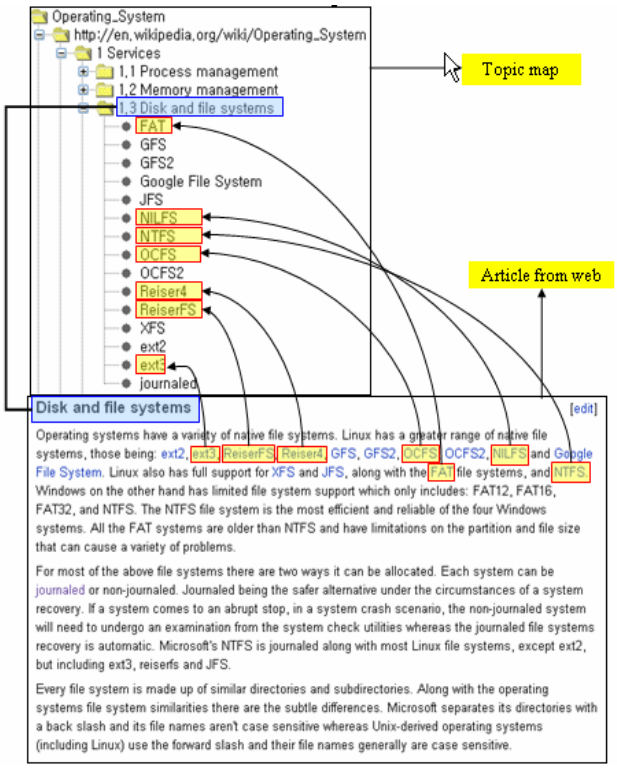

Figure 4. Table of contents collected by crawling a Wikipedia web page

5. RELATED WORK

Studies of concepts such as subjects and subject identity are with a long history starting in philosophy through logic, to linguistics, to computer science. The problem of subject identity has been recognized as more difficult than previously thought by proponents of the Semantic Web [12, 16]. The complexity of the subject identity problem originates from the fact that it is not related only to a particular syntax, or to a particular data model. Different users have different notions of subject identity. This fact was recognized in the original text of ISO 13250 [20], in the definition of subjects as: "...any things whatsoever, regardless of whether they exist or have any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever." Obviously, this definition opens potentially unlimited interpretations.

According to the Topic Maps current standard [20], subject identity is the idea that a topic represents a real world object, which can be identified by means of a subject indicator intended by the topic map author to provide a positive, unambiguous indication of the identity of a subject. So if two topics share the same subject indicator, they both deal with the same subject. Using this concept, it is possible to merge two independently created topic maps. On the other hand, the Topic Map Reference Model (TMRM) is the information model of the semantic integration standard for Topic Maps [15]. Interesting evolution demonstrated in this model is that more emphasis is placed on comparable subject properties than on precise URIs.

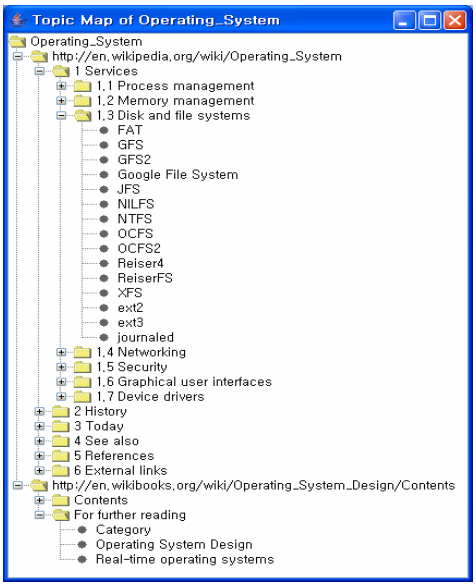

Figure 5. Generated topic map

The problems addressed in this paper are closely related to the idea of interpreting subjects from different perspectives and views [3]. We share the viewpoint expressed by Michele Biezunski that the origin of the problem of locating relevant resource lies in the subjects and in the fact that they are not computable. What are accessible to computers are their proxies, where the relation between subjects and proxies is not one-to-one. Another interesting discussion on the relation between names and subject identities and their role in KMS is presented by Moore [14]

In a similar line Steve Newcomb introduces the *Versavant* Project [21], which provides a Topic Map Application bus acting as "subject addressing engine". The bus allows aligning between different Subject Map Disclosure ontologies. To address the contextual aspect of the "sameness", Vatant introduces the concept of 'Hubject' (Hub + Subject) [19]. A Hubject is a hub connecting context specific representations of a subject. In another related work Maicher introduces a Structuralist Subject Equality decision approach called SIM [13]. The proposed approach allows an exchange of topic maps in the absence of a shared Subject Map ontology and Subject Map vocabulary. Steve Pepper is also addressing the problem of identifying subjects [17]. In particular, he advocates the concept of Public Resource Identifier - an URI that resolves to a Public Resource Descriptor, describing the subject (resource) it identifies. The emphasis in our work is not so much on general methods or specific steps towards solving the subject identity crisis. Our goal was to apply a pragmatic strategy to subject identity and utilize it for identifying subject representations in a particular domain such as e-learning.

The idea of exploiting Wikipedia semantics is not new; several studies on this topic have been carried out in the last few years. One of the first attempts to automatically extract semantic information from Wikipedia [9] aims at building an ontology from the Wikipedia collection. The authors of [10] discuss semantic relationships in Wikipedia and the use of link types for search and reasoning. Recent research has shown that Wikipedia can be successfully employed for NLP tasks, for example, in question answering [1], text classification [7], or semantic

relatedness [18]. A major difference between the above approaches and our approach is in the intended users of the extracted data: in our case it is intended for *human* consumption. Thus the novelty is in the collaborative building of the subject structure by the human author and the agent. In this scenario, the user interface and the presentation of the extracted topical structure are of equal importance compared to the accuracy of concept extraction. All this sets a new research perspective.

6. CONCLUSION

The potential and the challenge of the subject-centric knowledge organization lies in the relationship between subjects, their names, and the resources associated with them. In this paper we address this type of relationship where the focus is on how to use subjects and their names in order to provide humans with relevant data to decide whether or not information about a subject meets their needs. We address this aspect by viewing subjects as context dependant concepts, aiming at differentiation such as: this is the subject and these are the knowledge sources as viewed by individual A for the task B.

Gathering documents with learning value and data about subjects implies a possibility for identifying and representing them. The paper presents our view on subject identity and discusses some implications of recognizing the identity as an abstraction, capturing subject's relationships with other subjects. One implication of this view is that related subjects can be interpreted as a weak form of identity and from a practical perspective can be used as a starting point for browsing exploration. Part of the work presented in this paper addresses this issue from a practical perspective by suggesting a step towards solving the task.

Subjects are addressed through their representations; therefore agreement about subject identity should be grounded on some sort of representation, such as name, addressable resource, or URI. In this paper we propose using Wikipedia articles as a source for both consensual naming and subject identifiers. A shared context provides common conceptual ground and shared framework for communication. We claim that Wikipedia can play a role of a shared context between Topic Maps authors and users. We succeeded to make use of Wikipedia without deep language understanding. This was made possible by applying standard techniques to match the structures with relevant properties. While using Wikipedia alone yields a performance with variable reliability, in combination with Wikibooks it provides a better structured collection of topics, suitable for learning repositories.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. DUE-0442702 "CCLI-EMD: Topic Maps-based courseware to Support Undergraduate Computer Science Courses."

8. REFERENCES

[1] Ahn D, Jijkoun V, Mishne G, Müller K, Rijke M, Schlobach S. Using Wikipedia at the TREC QA Track. In *Proceedings of TREC 2004.*

[2] Biezunski, M., Bryan, M., Newcomb, S.: ISO/IEC 13250:2000 Topic Maps: Information Technology, www.y12.doe.gov/sgml/sc34/document/0129.pdf

[3] Biezunski, M.: A Matter of Perspectives: Talking About Talking About Topic Maps. In: *Proceedings of Extreme Markup Languages*, Montreal, 2005.

[4] Dicheva, D. & Dichev, C.: TM4L: Creating and Browsing Educational Topic Maps, *British Journal of Educational Technology* - BJET (2006) 37(3): 391-404

[5] Dicheva D., Dichev C.: Authoring Educational Topic Maps: Can We Make It Easier? *5th IEEE Inl Conf on Advanced Learning Technologies, ICALT 2005*, July 5-8, Kaohsiung, Taiwan, (2005) 216-219.

[6] Document Description and Processing Languages: http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0446.htm.

[7] Gabrilovich E., Markovitch S.Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *21st Conf on Artificial Intelligence (AAAI 06).* Boston, Mass., July 16-20, 2006.

[8] Hepp M., D. Bachlechner, and K. Siorpaes. Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements. *1st Workshop SemWiki2006 - From Wiki to Semantics* at ESWC 2006, Budva, Montenegro, 2006.

[9] Kozlova N. Automatic Ontology Extraction for Document Classification. *Master's thesis, Saarland University, Germany*, February 2005.

[10] Krotzsch M, Vrandecic D., and Volkel M. Wikipedia and the Semantic Web -The Missing Links. In Proc. of Wikimania 2005, The *First International Wikimedia Conference*, Wikimedia Foundation, 2005.

[11] Kruskal J. B. *On the shortest spanning subtree and the traveling salesman problem*. In: Proceedings of the American Mathematical Society. *7 (1956), pp. 48–50.*

[12] Lacher M. S. and Decker S. On the Integration of Topic Maps and RDF Data. In *Proc. of Semantic Web Working Symposium.* Palo Alto. California. August 2001

[13] Maicher L.Topic Map Exchange in the Absence of Shared Vocabularies, *Proc. of Intern. Workshop on Topic Map Research and Applications (TMRA'05)*, Leipzig, Germany, 2005; LNCS , Springer, http://www.springeronline.com/lncs.

[14] Moore G. Identities & Names in Knowledge Management, In: *XML Europe*, 2002.

[15] Newcomb, S. R.; Durusau, P.: Multiple Subject Map Patterns for Relationships and TMDM Information Items. In: *Proceedings of Extreme Markup Language, Montreal*, 2005

[16] Passin T. *Explorer's Guide to the Semantic Web* Manning Publications, 2004.

[17] Pepper S. and Schwab S. Curing the Web's Identity Crisis: Subject Indicators for RDF. TR, Ontopia, 2003. http://www.ontopia.net/topicmaps/materials/identitycrisis.html

[18] Strube M. Ponzetto S. WikiRelate! Computing Semantic Relatedness Using Wikipedia, Proc. the *21st National Conference on Artificial Intelligence (AAAI 06)*, Boston, Mass., July 16-20, 2006

[19] Vatant, B.: Tools for semantic interoperability: hubjects. http://www.mondeca.com/lab/bernard/hubjects.pdf, (2005).

[20] XML Topic Maps (XTM) http://www.topicmaps.org/xtm.

[21] http://www.versavant.org