TopicMaps: Unified Access to Everyday Data

Jens Heider
(Fraunhofer Institute Secure Information Technology,
64295 Darmstadt, Germany
jens.heider@sit.fraunhofer.de)

Johannes Bergmann
(Innovations Softwaretechnologie GmbH,
88090 Immenstaad, Germany
johannes.bergmann@innovations.de)

Abstract: Daily work with information spread across multiple data sources is still a time consuming task when it comes to managing, searching and securely distributing to dedicated recipients. The paper describes the generation of a homogeneous knowledge representation extracted from heterogeneous personal data sources. Used for unified navigation through personal knowledge it assists the user in retrieving any information even with limited devices such as smartphones through a single interface.
Key Words: knowledge representation, unified navigation, topic maps, design pattern
Category: H.2.5, H.3.3, H.5.2, H.5.4

1 Introduction

These days a huge amount of information has to be organized by individuals. The tasks range from storage of data in memorizabel structures to the secure distribution of information to dedicated recipients. The recurrent tasks contained in this range consists of searching, accessing data and keeping it synchronized across distributed storages, sometimes even across different types such as email attachments and document management systems.

In this context the paper focuses on the automated collection of knowledge from distributed data sources. The described approach will assist the user in his daily tasks by offering a unified interface to any information of his personal knowledge base. This is one core aspect of the MIDMAY-Project (Mobile Information Distribution Management and Access for You) [Heider2004] which builds the framework for mobile devices such as smartphones to evolve to an everyday tool for the secure distribution, management and access of information. On the one hand MIDMAY economically uses wireless communication to keep track of and to control the information and on the other uses wired communication to perform the actual information hosting, transmission and synchronization.

The process of collecting information from various sources is described to show the approach of autonomously generating the personal knowledge base offering a homogeneous navigation interface to distributed heterogeneous sources.

With this knowledge representation, which links together all information sources, the problems of daily information work should be addressed by providing position and access transparency for a mobile knowledge management as discussed in [Grimm et al.2005].

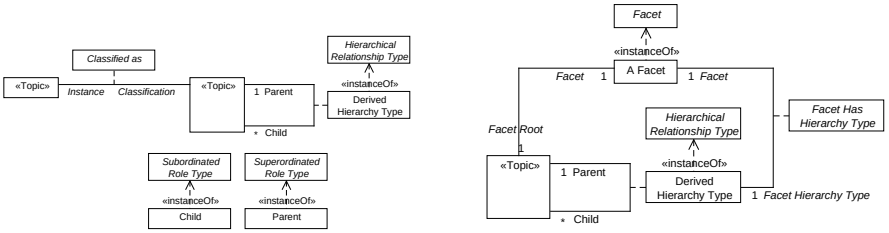
The proposed model is based on TopicMaps as universal backbone for the knowledge representation. This technology is used to build a unified and self-defining representation of all information extracted from personal data sources acting as a linked network to the original data after creating a homogeneous structure of arbitrary knowledge. It is able to rebuild the original structures but doesn't limit the system to specific sources. Based on this knowledge representation the navigation can be optimized for mobile phones as the user is assisted to follow his own associations in mind providing an intuitive interface.

2 Designing homogeneous structures

To unify the access to information a generic model is required which offers typing and structuring to index data sources. The model has to be generic to be used for any data sources and should provide a homogeneous structure to ease the access. This is the precondition for a unified representation which doesn't require specific visualization functions for every information type extracted from arbitrary data sources. The homogeneously structured representation also should offer the possibility to find related pieces of information, so generic patterns have to be defined also for an interconnection between them. This is performed on the top level by using the *Hierarchical Classification Pattern* and the *Faceted Classification Pattern* (see [Ahmed2003]) for constructing association, type and role topics. Using the published subject indicators (PSI) of these topics across different extractors unifies the hierarchical structures contained in the data sources.

The Hierarchical Classification Pattern (see fig. 1(a)) is used to define multiple hierarchies with different associations. For example the two relations *parent-child* and *container-containee* are mapped to different associations but represent the same hierarchical semantic. Therefore the *Hierarchical Relationship Type* is used as type for these associations and the *Superordinate Role Type* and *Subordinate Role Type* are used for instances playing a hierarchical role to unify the semantic for hierarchical relations. A topic is then made an instance of a class inside the hierarchy by using the *Classified As* association. This way all hierarchical associations can be explicitly marked and therefore treated homogeneous as their semantic is preserved by the used types. This is the first step for the unified representation.

A further step is the usage of the *Faceted Classification Pattern* (see fig. 1(b)) which enhances the model to enable an efficient finding of all hierarchies and their root elements. Every hierarchy is represented in this pattern through



(a) Hierarchical Classification Pattern (b) Faceted Classification Pattern

Figure 1: Topic Map Patterns

a topic of the type *Facet*. This topic is associated with the top level root element via the *Facet Has Root* association. The used association is of the type *Facet Has Hierarchy*. The result is a single root element for all represented hierarchies, building a hierarchy of multiple hierarchies. Now a designated entrance point is available to the networked information. This way all included hierarchies can be visualized equally without any changes in the browsing component even if they are expanded dynamically.

Now that all kinds of hierarchical relations can be represented in the generic model, the next step is to define rules for the representation of the information to be accessible itself. The primary goal is to build a representation that offers an easy way to access all linked information. Therefore only the primary data has to be used inside the topic map which interconnects them to a knowledge base of accessible information. This should be the information the user most likely will recall to access the linked data. The key issue here is the reusability of information types across different data sources like e.g. person, subject, time, location and information type. The more general topics are identified and used as topic types the more interconnections by associations can be established across the sources and the easier the user can follow his own mind paths in a natural way to find the desired piece of information he is looking for.

Of course this metadata can be enriched also from other sources leveraging context information automatically generated by preexisting knowledge or ambient sensor devices (see [Badii et al.2006]), but in this paper we will focus on information directly extracted from data sources found in typical working environments. Any information related to the existing source then become the referencing metadata for the information it links to.

3 Specifying generic rules

Rules for the extraction can be generalized for instances of one-, two- and n-dimensional sources. In the case of one dimensional source like lists, the repre-

sentation is unified by associating every entry with the root element of the list source. The order of the entries is preserved using the *variant name* property of every entry's topic which acts as key for the sorting. If these simple data sources are chosen to be accessible by the user, we presume the information contained should be directly accessible inside the representation like a list of appointments. This offers the interconnection with data retrieved from other sources. If the user only wants to access the list as a whole without any further integration of contained data, the system has to be told so, like through decision by file prefixes. This is useful if the list e.g. contains numeric measuring data which isn't useful to be associated to other sources in the context of information retrieval.

3.1 Two-dimensional sources

Two dimensional sources like any kind of tables will be mapped also as hierarchical structure. Tables are processed row by row resulting in topics generated from keyed columns. These may be associated also to other relevant information types contained in the same row. This again depends on the definition and creation of topic types most important in the context the framework should be used in. As a general rule for the usage of topics, any information type should be used that

- is found in more then one used data source (interconnection of data sources),
- will offer memorable attributes (offering access via facts recalled by the user)
- or provides additional context information (classifying information to enrich the representation allowing the system to facilitate the user in finding related information).

Any other information doesn't have to be included in the topic map as it is accessible via the link contained in the metadata of the considered source. A special consideration is required if tables do not contain 1:1 relations only. In the case of 1:N relations also the related tables have to be inspected for relevant information, judged against the same rules as defined above and associated to the keyed topic if appropriate. But associations made in this process sometimes can not be typed for the representation, as they may not be specified in the data source and their meanings are only known to the application using the source. In this case a generic *is related to* association should be used to preserve the relation and to prevent innumerable useless associations. The same is true for N:N relations which may also be contained in some data sources.

3.2 N-dimensional sources

Of course one will encounter also n-dimensional sources like e.g. in the field of data warehousing. The challenge in projecting multidimensional data to a hierarchy for the described purposes is choosing the dimension which offers the most

valuable information for the user qualified by the rules. Whereas simple structured data source often benefits from the modeling of properties and metadata through additional views onto the information objects, this is not the case for complex data sources. By choosing a dimension, loss of possibilities for queries to this data have to be taken against the advantage of interconnection with other sources and a fast access. Otherwise the complete data source has to be represented inside the topic map, which would pervert the idea of a linked access to stored data.

4 Navigating inside unified representation

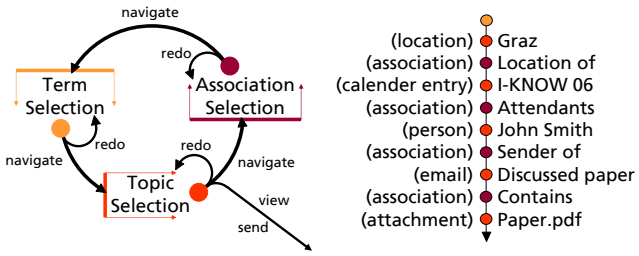
After building the unified representation, a key aspect for the intended usage is the development of a useful navigation interface. As topic maps represents a network of topics interconnected by associations it is obvious to use these paths for navigating through the node-centric network (see [Dave et al.2003]), as it can be implemented on small displays. The interesting question is, which topic should be chosen as the starting node. At this point no information about the search interests is provided by the user but the user has to be offered an entrance to the representation. This problem also appears when accessing information stored at databases where it is often solved by query languages. A similar approach has also been developed for TopicMaps with the query language Tolog[1], which can be used also for mobile clients [Seedorf et al.2005]. But the power of those query languages with their precise description capabilities of the desired result do have also some drawbacks for the mobile user. He has to know the correct syntax and at least some attributes of the data structure to produce meaningful queries. Especially on smaller mobile devices the input of many characters is a disadvantage that should be avoided to speed up the search procedure.

4.1 The Approach

An alternative approach therefore is the usage of parts of the information the user will probably search for. It is performed by splitting up every topic into their sub parts, called terms here. The title of this paper will result in the terms `TOPICMAPS`, `UNIFIED`, `ACCESS`, `TO`, `EVERYDAY` and `DATA` by removing any non-alphabetic characters from the string. As every topic is treated this way, regardless of its type, a complete personal vocabulary of terms is created by the framework and each unique term links to all topics it is found in.

The user then only have to recall a single term to enter the representation at a useful position. For example this could be a term contained in the authors

[1] Implementation by Ontopia of the requirements stated by ISO topic map query language TMQL standardization effort



(a) Navigation Graph (b) History of Navigation

Figure 2: Navigating through the representation by cycling the interface

name, the document title or the folder the file is stored in. If the user can't remember those direct associated terms, he may chooses first a term related to the person which emailed the document and then follows the associations inside the representation to the emails which were send by this person, to the one which contains the desired attachment. If even that isn't known, the user might recall the town where he attended the conference and meet the person who emailed the document. Provided that the personal organizer contains this meeting with location and attendants, the extractors have already included this knowledge to the representation and the user can navigate along the associations to the desired information.

The general idea behind this is to help the user in following his own paths in mind rather than to force him to think in given rules. The second advantage of this concept is the implicit generation of linguistic relations which can help to find related information that isn't explicitly linked by the topic map. Of course these relations sometimes aren't correct in the sense of the actual search, like in the case of homonyms or acronyms, but as the framework is designed for humans, it can be assumed that they are used to deal with those multiple meanings.

4.2 The Interface

Applying this concept of navigation, an easy to use unified access across all information sources can be created by cycling through three stages (see fig. 2(a)). At first the term selection phase is used to choose one search term. The result is the topic selection phase offering a set of topics containing the term. They can be accessed directly or will be used as navigation nodes to other topics. In the later case the association selection phase is entered which provides the associations connected to the selected topic. After choosing one association a set of available terms is created and the user again enters the term selection phase, starting the next cycle.

Navigating inside the topic map is now independent from the contained data types which enables a universal interface without the need for changes in the programming if new sources and data types are included in the representation. Simplifying the navigation the user is allowed to navigate backwards and forth through the cycle. This is offered directly by navigation options presented by each node and more flexible through a navigation history. It lists previous topics and associations the user passed through showing the navigation path. This provides an additional aid to the user locating his position in the representation and offers contextual meanings for the actual node (see [Park and Kim2000]). After finding the desired document, the history of the example described in section 4.1 would look like fig. 2(b).

A further improvement is the ability to tell the interface multiple terms that are relevant and also such that are not relevant. This reduces the number of cycles as more information is provided at the same time in the beginning. Of course this highly depends on the description capabilities of the used terms and the suggestion algorithm to assist the user with an automatic context-based topic search [Maguitman et al.2004]. Further research is going on to tell if more complex search options like search expressions with binary operators and automated path calculation between selected topics offers a significant improvement for the overall search time compared to the more complex setup in the user interface.

Especially the visualization of linear paths between two or more selected topics looks promising as it provides the context between them and offerers navigational options for directly jumping to a desired topic, building a trade-off between graph-centric and node-centric navigation towards a path-centric approach [Dave et al.2003].

Because all possible terms are known to the system, the user only has to enter a few characters to select the complete term from the generated list offered by the interface. In topic selection and association selection phase the interaction is also performed by choosing one list entry to control the navigation. So the complete process can be seen as an instant travel through the personal information universe with an easy to use remote control.

5 Summary and Conclusion

The proposed approach shows the benefits of designing homogeneous structures from existing distributed knowledge to offer a single interface even on limited devices. This is achieved on the conceptual level by unifying hierarchical structures with the help of TopicMap design patterns preserving existing hierarchical structures. The next step in our approach was to specify generic rules for the extraction of information the user may recall, which interconnects the data or which classifies it to enrich the representation allowing the system to facilitate the user in finding related information.

To utilize the design, the resulting framework autonomously generates a unified representation of personal knowledge providing access to the information found in arbitrary sources. Files, emails and appointments are interconnected in this representation managed by the framework. These connections between entries are created by leveraging the redundancy of information arising from combining all provided data sources. The representation therefore contains all structures found in the original sources. Additionally new paths between entries resulting from the unified processing of all information are included. Examples for those interconnecting topics are information about persons, dates, locations and data types contained in multiple sources.

On top of the TopicMap knowledge base, an interface is created using paths to traverse from any entrance point to the information to interact with. To ease the entering of search criteria for the entrance points, the interface uses precalculated data to offer available search items. After finding the desired information through few user interactions it can be viewed, manipulated or send to other users.

The approach leverages existing knowledge helping the user to find desired information through a unified navigation. Considering the flat learning curve for the usage of the described approach it can also be used by non-scientific workers as no query languages has to be learned. The benefits increases the more data sources and context information are interconnected, so the user can use his own associations in mind to access, manage and distribute his personal knowledge.

References

[Ahmed2003] Ahmed, K. (2003). Beyond PSIs : Topic map design patterns. In *Extreme Markup Languages*.

[Badii et al.2006] Badii, A., Hoffmann, M., and Heider, J. (2006). MobiPETS-GRID - context-aware mobile service provisioning framework deploying enhanced personalisation and privacy and security technologies. In *Proceedings SOFTPLATFORMS*.

[Dave et al.2003] Dave, P., Karadkar, U. P., Furuta, R., Francisco-Revilla, L., Shipman, F., Dash, S., and Dalal, Z. (2003). Browsing intricately interconnected paths. In *HYPERTEXT '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 95–103, New York, NY, USA. ACM Press.

[Grimm et al.2005] Grimm, M., Tzari, M.-R., and Balfanz, D. (2005). A reference model for mobile knowledge management. In *Proceedings of I-KNOW*.

[Heider2004] Heider, J. (2004). Vision und Realisierung einer sicheren mobilen Informations-Verteilung, Verwaltung und Abfrage. In *Multikonferenz Wirtschaftsinformatik (MKWI) 2004. Bd 3. Mobile Business Systems*.

[Maguitman et al.2004] Maguitman, A., Leake, D., Reichherzer, T., and Menczer, F. (2004). Dynamic extraction topic descriptors and discriminators: towards automatic context-based topic search. In *CIKM '04: Proc. of the 13th ACM international conference on Information and knowledge management*, pages 463–472. ACM Press.

[Park and Kim2000] Park, J. and Kim, J. (2000). Effects of contextual navigation aids on browsing diverse web systems. In *CHI '00: Proc. of the SIGCHI conference on Human factors in computing systems*. ACM Press.

[Seedorf et al.2005] Seedorf, S., Korthaus, A., and Aleksy, M. (2005). Creating a topic map query tool for mobile devices using J2ME and XML. In *Proceedings of the 4th international symposium on Information and communication technologies WISICT*.