




Centrum voor Wiskunde en Informatica

View metadata, citation and similar papers at core.ac.uk

brought to you by  **CORE**

provided by CWI's Institutional Repository

REPORT*RAPPORT*

INS

Information Systems



Information Systems

Clustering semantics for hypermedia presentation

M. Alberink, L.W. Rutledge, L. Hardman, M. Veenstra

REPORT INS-E0409 NOVEMBER 2004

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).

CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2004, Stichting Centrum voor Wiskunde en Informatica

P.O. Box 94079, 1090 GB Amsterdam (NL)

Kruislaan 413, 1098 SJ Amsterdam (NL)

Telephone +31 20 592 9333

Telefax +31 20 592 4199

ISSN 1386-3681

Clustering Semantics for Hypermedia Presentation

ABSTRACT

Semantic annotations of media repositories make relationships among the stored media and relevant concepts explicit. However, these relationships and the media they join are not directly presentable as hypermedia. Previous work shows how clustering over the annotations in the repositories can determine hypermedia presentation structure. Here we explore the application of different clustering techniques to generating hypermedia interfaces to media archives. This paper also describes the effect of each type of clustering on the end user's experience. We then generalize and unify these techniques with the use of proximity measures in further improving generated presentation structure.

1998 ACM Computing Classification System: H.5.4, H.5.1, Categories and Subject DescriptorsH.5.4, H.5.1

Keywords and Phrases: Hypermedia, Presentation Generation, Clustering, Semantics, Style, Document Structure

Clustering Semantics for Hypermedia Presentation

Martin Alberink, Lloyd Rutledge*, Lynda Hardman* and Mettina Veenstra

Telematica Instituut

P.O. Box 589

NL-7500 AN Enschede, The Netherlands

Tel: +31 53 485 04 85

E-mail: FirstName.LastName@telin.nl

*CWI

P.O. Box 94079

NL-1090 GB Amsterdam, The Netherlands

Tel: +31 20 592 40 93

E-mail: FirstName.LastName@cwi.nl

ABSTRACT

Semantic annotations of media repositories make relationships among the stored media and relevant concepts explicit. However, these relationships and the media they join are not directly presentable as hypermedia. Previous work shows how clustering over the annotations in the repositories can determine hypermedia presentation structure. Here we explore the application of different clustering techniques to generating hypermedia interfaces to media archives. This paper also describes the effect of each type of clustering on the end user's experience. We then generalize and unify these techniques with the use of proximity measures in further improving generated presentation structure.

Categories and Subject Descriptors

H.5.4, H.5.1 [Information Interfaces and Presentation (e.g., HCI)]: Hypertext/Hypermedia – *architectures, navigation*; Multimedia Information Systems – *Hypertext navigation and maps, Evaluation/methodology*; I.7.2 [Document and Text Processing]: Document Preparation – *Hypertext/hypermedia, Markup languages, Multi/mixed media, standards*.

General Terms

Algorithms, Documentation, Design, Experimentation, Human Factors, Standardization, Languages, Theory.

Keywords

Hypermedia, Presentation Generation, Clustering, Semantics, Style, Document Structure

1. INTRODUCTION

Facilitating the archiving of media and knowledge for eventual presentation to end users motivates our work within the Topia project. The related research fields of data mining, media retrieval, semantics and presentation generation provide us with useful insights and tools. Our project work up to now demonstrates the feasibility of cluster processing for converting semantics to presentation structure as part of this facilitation. This section discusses this goal, the work we and others have done in pursuing it, and what this paper hopes to contribute.

1.1 Motivation

More and more aspects of the human creation of hypermedia presentations are being modeled and implemented on computers. We observe three steps in this human process: that of collecting, organizing and editing together the media content. The application of style to structured content processes the last step by taking a hierarchical and sequential organization, such as that defined by HTML and XML, and allowing the means of presenting to adapt to different users and environments.

Media indexing and retrieval technology, with an additional boost from semantic annotations and querying, helps provide the first step: that of searching for and assembling media appropriate for a chosen topic. But this technology typically generates an unorganized list of matches, presenting each match as a pre-existing media item or document meant to completely fulfil the user's request. The human author, on the other hand, uses not one but most or all of these items. Furthermore, the author creates a new hypermedia presentation from their combination.

The middle step, that of building a document structure around a collection of retrieved media, remains mostly unassisted by computers. Computing this step, in combination with the other two, would complete the chain of generating hypermedia presentations of any topic upon request from a media repository. One human approach to this step is taking physical, visual media found in an archive, spreading it out on a table, and moving the objects around, searching for patterns and connections, and eventually a story threading them together. This helps the author build a document structure around the media as the basis for final presentation. We feel clustering technology offers some help in automating the finding of patterns between media components selected from a large, interlinked repository and transforming these patterns to document structure.

1.2 Previous Work

In previous work within the Topia project, we started approaching the computation of all three of these steps by developing a prototype system testing its feasibility [9]. We encoded this system by combining Semantic Web and Web style technology [10]. The technology this implementation applied to each of these three steps is shown in Figure 1. Our demonstration generates hypermedia presentations from the user's request for a topic regarding the

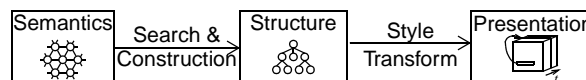


Figure 1. Semantics to Presentation Path [10]

collection of the Rijksmuseum Amsterdam [8]. Figure 2 shows an example of a presentation from our demo.

In this system, the user enters a text string specifying the desired topic for a presentation. A query on the RDFS-defined digital repository returns an RDFS subset of the objects with fields containing this string. Rather than return these objects as a single list from which the user selects one, as typical Web search engines do, this system builds all of these objects into a single, organized presentation. The system's key feature is determining a good overall hierarchical and sequential structure in which to present these returned objects. To demonstrate that this is possible, we implemented one technique for generating this structure, that of concept lattices. However, there are other means available for building presentation structure from an underlying repository.

1.3 Related Work

The generation of coherent presentations from repositories is the goal of several bodies of ongoing research. In previous work, we discuss some related research, including processing semantic annotations to retrieve media components and to derive discourse structures [9][10]. Here, we discuss more recent work, and work that applies to this paper's focus on clustering methodologies.

Different techniques exist for finding *clusters* relating objects to each other. *Flamenco* organizes images retrieved from a metadata-based search into a matrix structure. *Spectacle* expresses proximity of objects by visualizing groups and their overlaps in a two-dimensional Cluster Map. *mSpace* organizes an ontology into a multi-column user interface based on the concept of slices through the ontology's multidimensional structure. *Disc* performs domain-specific queries on RDF repositories to fill in specific discourse-based templates in building a richer broad narrative within generated presentations. Hera takes a model-based approach in generating user-tailored hypermedia presentation from multiple, specific repositories. We describe each of these in turn.

Clustering. Data mining and other areas of research have developed various clustering techniques that we can apply to our system [1]. We identify three basic categories of clustering for applying to our work: those based on *properties*, *links* and *axes*. Each of these categories forms the basis for its own main section in this paper describing its integration into our system.

Flamenco. The Flamenco search interface uses a fixed number of metadata types of images for organizing search results in a square matrix structure [6]. The organization, done at presentation time, allows users to interactively search for collections of images, revising their query after each step according to their latest insights about their own information needs. Usability requirements determined Flamenco's design, making it consistent and recognizable, but also less flexible than our Topia architecture with respect to dealing with other metadata types.

Spectacle. Proximity between objects in presentations can visualize their conceptual proximity. Existing visualization techniques convey groups by placing objects close to each other in presentations, and use the presentation space to show overlaps of groups by placing the overlaps between the overlapping groups [3]. Hierarchical grouping, which we describe in this paper, decomposes the multiple sublevels in grouped semantics. We use the single dimension that distance has in a two-dimensional presentation space for expressing proximity at each hierarchical level in accordance with its dimension.

mSpace. Applying grouping methods to semantic annotations can elicit many patterns in the characteristics of a set of objects. The unconditional inclusion of all groups with a pattern might overwhelm users. The mSpace interaction model prunes a hierarchy of grouped properties, such as the structured progression in this paper [5]. It shows users the cross-section of the description space that user-specified properties. Users themselves then find the patterns concerning the occurrence or absence of objects for combinations of properties in the selected description space. The columns in the mSpace slices show the cross sections of an increasing number of attributes from left to right, and could apply to the attributes of groups generated by axial and relation clustering discussed in this paper.

Disc. Automatic generation of presentations that convey relations between information objects is the focus of several research projects. Some fill story templates in with database contents in order to deliver stories of pre-specified type. The Disc approach is to develop rule structures that retrieve available information for generating specific genres of presentations such as biographies and a curriculum vitae [4]. This approach searches relevant information in order to generate complete stories for well-defined users' needs. Generation rules include all relevant and available information in a story such that it is as complete as possible and still cohesive if certain information is not available. This approach works very well in closed application environments, such as virtual museum environments. Such an environment can classify the type of user information need, such as a virtual tour along highlights and painters' biographies and curricula vitae. The cohesion focuses on inclusion of proper pieces of information in standard communication structures, whereas Topia's invariably has

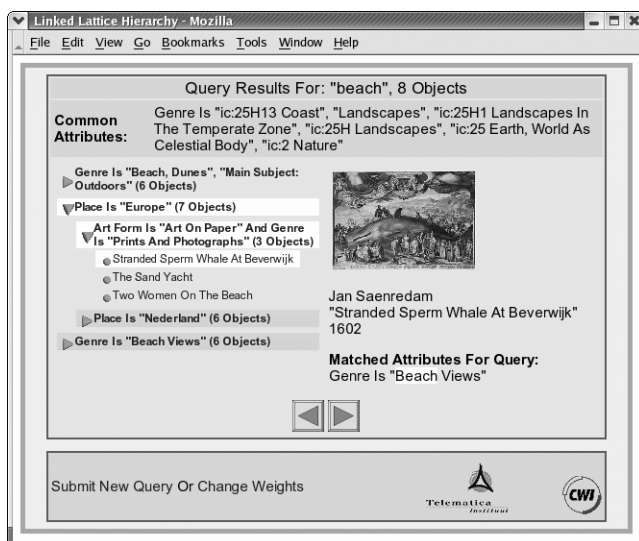


image © Rijksmuseum Amsterdam, used with permission

Figure 2. A Generated Presentation

sequenced hierarchy as standard communication structure, and arranges it such that transitions between objects and groups are cohesive.

Hera. The Hera methodology uses specifications of objects in documents, such as artifacts and artists, for generating conceptual documents based on semantic graphs [11]. These document objects, called slices, use models to specify adaptation to user characteristics, domains and document genres. In this paper, proximity metrics and clustering methods generate groups of a simpler nature. When the meaning of the grouping is understood, these general cluster constructs make sense in many domains.

1.4 Paper Overview

We start by presenting our use of *proximity* between concepts and its expression with general presentation structure. This principle guides the subsequent discussion of how structure derives from three types of clustering: those based on *properties*, *links* and measured *axes*, each of which is the focus of an upcoming section. We then present a model that unifies and combines of these clustering techniques. We end by describing a search algorithm for improving general document structure that incorporates and combines this clustering, proximity and other considerations. Each section also describes the impact of its component of presentation generation on the experience of the end user.

2. PROXIMITY AND STRUCTURE

Our architecture generates presentation *hierarchy* and *sequence* to communicate relations between the presentation's concepts that exist in the underlying repository. The utility of this depends on two principles. One is that hierarchy and sequence each communicate that *proximities exist* between concepts. The other is that the resulting presentations are able to communicate the *nature* of the conceptual proximities that generate each group and sequence. This principle enables a rating of how well a presentation structure communicates its underlying concepts. Such measurement not only provides evaluation of clustering techniques throughout this paper, its processing can guide the improvement of presentation structure during its construction, as we describe toward the end of this paper.

Proximity. Discussion about clustering techniques alternately refers to the concept's *distance*, *similarity* and *proximity* in describing how to examine objects in choosing whether to join them in a cluster [1]. We prefer the term *proximity* for our work. It is more appropriate here than distance because presentation structure typically joins document objects rather than separating them. Furthermore, objects do not have to be similar to be bound in the presentation structure — they simply should have a close relation. Something about two objects near each other in presentation structure should strike human authors and users as tightly, conceptually connected.

Numeric Proximity. We assume that varying degrees of proximity exist between presented concepts. Furthermore, we assume these degrees of proximity are relative to each other, enabling quantitative comparison. Such numeric proximity assignments are only useful if human perception of conceptual proximity and its resulting impact on generated presentations is consistent with this

measure. That is, if the presentation shows one group of objects as more proximate to each other than those on another group, the users' understanding of the underlying conceptual closeness of the objects should compel them to agree that the first group is more tightly related than the second. Since automation requires that these measures are computable, human designers for such systems must create formulas for calculating proximity whose processing remains consistent with user perception.

Proximity Tables. Given a proximity measure that applies to object pairs, one can create a diagonal table mapping each possible pair of objects to the proximity measure between them. Such tables of pairwise proximity represent the overall conceptual proximity between the document objects that the final presentation structure aims to convey. Since a presentation's hierarchical and sequential structure aims to convey overall conceptual proximity, the proximity table formalizes what this structure needs to match.

Hierarchy Conveying Proximity. Hierarchy communicates proximity by having close concepts in the same groups, recursively across multiple levels. This is similar to determining how close relatives are in a family tree. Being direct siblings conveys the closest proximity. More generally, proximity between two objects is inversely proportional to the level of depth they lay below their closest common ancestor.

Hierarchy Conveying Proximity Rationale. The rationale used for each group should be *title-able*, and preferably *presentable*, to the user. We have shown in earlier work the importance for making titles for groups [9]. Figure 1 shows group titles in the outline along the left side of the screen display. More recently, we demonstrated the utility of generating sub-presentations with more information about what a selected group of objects has in common [10]. Here, our system finds additional media objects representing grouping commonalities and makes a screen display from them.

Sequence Conveying Proximity. Sequence communicates not just order but also proximity by having closer concept pairs tend to be close to each other as measured by number of steps along the sequence, with the very closest objects being adjacent. While a single sequence is effectively an axis, the objects in the underlying structure do not necessarily fall along an axis. This axial representation of the proximity table is thus approximate.

Sequence Conveying Proximity Rationale. All the issues for conveying hierarchy rationale apply to sequencing as well. The user needs to understand why objects appear in a sequence in order to benefit from that sequential placement. There are many types of underlying structures that cause sequencing. These different types of rationale result in different patterns of explanation to the user.

3. PROPERTY-BASED CLUSTERING

We have shown in previous work how clustering based on properties can generate outlines, or topic hierarchies, for presentations [9]. Groups in such hierarchies mean that their members have shared property assignments. The weighted concept lattice algorithm from this previous work makes hierarchies that are efficient in maximizing this property-based grouping. Resulting presentations communicate to the user what the topics

have in common with each other in terms of these properties. We use simple titles to convey the nature of these groupings for each group. This section presents new work on our application of concept lattices. This includes making titles for conveying the nature of presentation structures, the derivation of sequence from properties, the use of semantic subsumption to enrich the property set and deriving proximity tables from concept lattices.

Generating Presentation Sub-topic Titles from Property Assignments. Previous work presents the use of titles in the outline display for groups generated by concept lattice clustering [9]. This system generates a title for a group from the property assignment that concept lattice processing determines the group has in common. Simply put, such a title consists of a string representing the property type followed by the words “is a” followed a string representing the property value. Section 4 discusses how we improve upon this property-based title generation by treating properties as graph nodes from the repository with specific text objects designated for the node’s title.

Conveying Sequence Basis with Titles. Our demo processes *numeric properties* to determine sequences within established hierarchies [9]. Specifically, it derives sequence from the artifacts’ years of creation. To convey the nature of this sequence, the demo puts these years in each artifact’s display and title. The use of year is easy because as a simple, readily recognizable, number, its sequencing is immediately apparent in displayed lists. However, we found that this provided too much information, cluttering the outline display, so we removed them. This is a matter of personal taste, of course, and individual style specification could request that, in this case, the years be include in the outline title displays.

Sequence Segues from Shared Properties. Not all sequences derived from numeric properties. Clustering of non-numeric properties potentially apply not only to grouping but also to sequencing. Two objects in a larger group can become adjacent if they have property assignments in common. Applied repeatedly, this can produce chains of objects within a group, with each adjacent pair potential linked by different common properties than the others. This enables an expression of proximity not possible with hierarchy, since, for example, putting the first and second objects in such a generated sequence in one subgroup and the third and fourth in another neglects communicating the commonality between the second and third. A communication device we call the *segue* can convey such chains of chain common properties. A segue is a distinct sub-presentation between the sub-presentations of two adjacent objects displaying the property assignment that joins them.

Semantic Subsumption to Extend Properties. Inferencing that deploys Semantic Web technology provides additional properties for these tables. Our system arranges genres of artwork in a *subclass* hierarchy, allowing subsumption to infer the whole ancestry of genres for each concept node referring directly to any genre. This gives the concept lattice a larger property table to process, enriching the possibilities for the resulting structure.

Total Shared Property Weights as Proximity Measure. The more properties in common that objects have, and the higher the

significance weights the user assigns these properties in our system [9], the larger their conceptual proximity is. Each shared property brings objects closer together. The significance weight assigned to that property determines how much closer that shared property brings the objects. Mathematically, thus, one can measure the proximity between two objects as the total of the weights of the properties the objects share. If they share no properties, their proximity is zero.

4. LINK-BASED CLUSTERING

Clustering techniques exist not just for properties but also for node-edge graphs, such as those representing our semantic relational structure. Such link-based clustering can emulate property-based clustering, as we describe below, but it goes beyond property-based clustering as well. One essential distinction of link-based clustering is that the nature of its groups is nodes sharing links with the originally selected nodes. Like the original nodes, these cluster nodes can be presented, enabling sub-presentations that are more elaborate and can thus potentially more accurately convey the nature of these groups. In addition to better presentations of group rational, link-based clustering finds a larger category of both groups and sequences than property-based clustering, and more means of combining hierarchy and sequence generation. This section progresses by showing how link-based clustering emulates property-based clustering, conveys more about the nature of group selections and goes beyond property processing in generating presentation structure.

Emulating Property-based Clustering. Our implementation of property-based clustering actually uses components of link-based clustering. Some property values, such as specific genres, are actually links to concept objects in the repository. Our concept lattice processing handles these URI values as strings to match, thus not distinguishing links from any other property value. In terms of link-based processing, however, the system effectively recognizes cluster hubs that are a *single link traversal*, along the *same link type*, from each cluster member. As described ahead, this recognition of clusters as presentable graph nodes, instead of simply properties with titles, enables introduction displays of the clusters that enrich the presentation and its informativeness.

Emulating Subsumption in Property-based Clustering. As described in Section 3, our implementation of property-based clustering incorporates semantic inferencing of additional property assignments. This also emulates a particular case of link-based clustering: that of cluster hubs joined to each cluster member by *uniformly typed link chains*. That is, the hub connects to each member with a chain of links of the same type, which in our case is “is-subclass-of”.

Group Introductions from Cluster Nodes. Concepts that bridge a group of selected concepts can themselves be first-class topics in the presentation as introduction displays for the group. We demonstrate this with concept lattices because our implementation always returns at least an attribute assignment whose strings could form a title, but frequently also a *first class concept* with an official title and media of its own. The previous version of our system only presented leaf nodes of the structured progression hierarchy [9]. The leaf nodes represent RDF resources matching the user’s

information request. The composite nodes represent the properties their descendant nodes have in common. However, while not matching the user's request, composite node concepts are often useful in understanding the directly matching resources, such as by providing insight into what causes these returns to have their matching property. This paper introduces to our architecture the presentation of composite nodes of the structured progression in the main display. This makes the topics they represent first-class concepts along with the information request returns.

Hierarchy-first with Cluster Hubs. The graph relations in property-based clusters treat the properties as distinct nodes connected to all cluster members. Not only can the interconnection structure of relations in an edge-clustered node have such hubs, but it can also have chains of subsequent relations and branches sprouting from a main branch [7]. Hubs are themselves presentable, as are the nodes directly connected to them. Thus, branches hierarchically decompose the clustered node.

Sequence-first with Traveling Salesman. One example of a relation-derived sequence is the use of is-influenced-by chains for sequencing painters. Any chain of relations joining a group can sequence that group. Of course, some types of chains cause more meaningful sequential presentations than others do. Weighted relations types can help in choosing the best sequence base out of a variety of possibilities. This would affect the application of the traveling salesman problem in an appropriate manner. Sequence derived from the underlying semantic relation graph is more complex to convey than sequence derived from numeric properties. Our example of influence chains would probably need an explicit statement to convey that sequence's nature.

5. AXIS-BASED CLUSTERING

Clustering from properties or graphs uses concepts specified by the author, which are, respectively, the property assignments and underlying concept nodes in the semantic graph. This section introduces our application to hypermedia generation of clustering along measured axes. This places the leaf-node document objects within a hierarchy of spans along such an axis. Each *span* is a group that does not necessarily correspond to an author-created concept — the system typically creates it specifically for the current presentation. While this adds great flexibility to hierarchy generation and facilitates the user's knowledge discovery, its limitations over property and link-based clustering are the small subset of properties it uses and the simpler presentations of clustering rationale it is restricted to. Figure 3 shows an example of a presentation generated by our demo using clustering along measured axes. This section progresses by defining measured axes in this context and then describing how they apply to presentation structure choices, how systems can convey them and what the ramifications for the user are.

5.1 Generating Presentation Structure from Axes

Use of an axis to generate presentation structure necessitates first deriving the axis from the underlying structure of the repository. Axes are inherently sequential, making the derivation from them of presentation sequence relatively straightforward. There are also clustering techniques for deriving hierarchies from axes. We

discuss here, in turn, the determination of axes, their sequencing and the deriving of hierarchies from them.

Determining Measured Axes. In previous work, our demo ensured that the field for sequence basis is included in each main display, which in our case was the year of creation. The use of year is easy because as a simple, readily recognizable, number, its sequencing is immediately apparent. Such measures can involve complex calculations, such as, in our sample domain, calculating painting surface area from width and height fields, but they remain simple to convey. Applying the principle above, we could put the surface area with each object display. Thus, each formula generating a number to be the sequence index could be included with a descriptive field name, in the object display. However, this only applies when the sequence basis is along a single dimension.

Measured Axes as Sequence Basis. A measured axis provides a clear and direct basis for computing sequence since objects appear along it with numbers. Whether the nature of the sequence is clear to the user is another matter. We have found with our system that year of creation is a conceptual axis whose nature is easy to communicate to the user. An axis based on artifact aspect ratio or area is also communicable, but arguably less so. Other axes may be

image © Rijksmuseum Amsterdam, used with permission

Figure 3. Left Portion of a Presentation from Axial Clustering

computable but meaningless to the user, such as artifact catalog number within the museum archives.

Placing Hierarchies over Sequences. Clustering of objects placed along an axis produces hierarchies of these objects. The outline of Figure 3 shows an example of a hierarchy derived along the “year of creation” axis. Objects that are close together along the axis tend to appear in the same groups. Relatively large gaps along the axis split the hierarchy along relative high levels. Our implementation of axis-derived hierarchies, which generated the presentation in Figure 3, uses the greatest distance algorithm.

5.2 Axis Spans as Virtual Concepts

Axis spans in an axially generated hierarchy correspond to neither specific property assignments nor concept nodes, which are the means for representing and communicating groups in, respectively, property- and link-based clustering. Thus, axis spans are *virtual concepts* from the perspective of the repository. Since the repository has no means of representing or conveying such virtual concepts, the system must generate them from scratch for each presentation. We discuss here the importance of such virtual concepts and how the system can better convey them.

Serendipitous Knowledge Discovery. However, deriving such virtual concepts is not entirely problematic because they provide the user with a potential for knowledge discovery. Sometimes the user can recognize the basis for a virtual cluster even when the system has no explicit object for it. The presentation facilitates this recognition by the user by communicating the *existence* of the cluster, leading the user to contemplate what brought it about. For example, a cluster’s year span may coincide with a period of history about which the user has some knowledge. The user then can recognize from the paintings place in this span some characteristics relating to this period in history.

Generating Titles for Axis Spans. The axis-based clustering on year can return strings for use in titles. These are simple year spans such as “1901-1911”. Generating more media conveying this groups, however, is more complex. As a virtual concept, such a year span has no corresponding conceptual object in the repository. Thus, the repository cannot directly provide media on the year span to present.

Representative Concepts for Axis Spans. While axis spans have no concept resources from which to build introductions, it is interesting to find substitutes. One is to find a *representative* first class topic from all objects in this span. Examples for choosing this include the middle-most in terms of the numbers used, or the most linked-to from other objects, thus denoting importance or popularity. Additional compensating techniques could include a search of all objects matching the group’s rationale, even those not in the original document request query, and finding properties they all, or mostly, share in common. Then the main display for the grouping would be a presentation of these properties.

5.3 Axial Distance as Proximity Measure

Of the proximities measures discussed for our three clustering categories, axial distance applies most readily to the generalized clustering we present in the next section. Specifically, the

proximity between each object pair is the inverse of the distance measure between them along the axis. As we will see, combining this type of proximity with others enables a broader and more powerful measure of proximity, which in turn expands the space of presentation structure the system can explore for improving communication of the underlying conceptual proximity.

6. GENERALIZED PROXIMITY-BASED CLUSTERING

Our architecture generates presentation hierarchy and sequence to maximize the matching between the proximity this structure communicates with that of the proximity table derived from the underlying repository. All grouping and sequence techniques boil down to proximity measures. With our application of concept lattices [9], for example, the number of shared property assignments for a pair of objects is the proximity measure. You can pump this number into a general-purpose proximity matrix to structured progression converter and get the same results. This general converter could also accept output from other types of clustering, such as axis-based clustering. This allows the generation of presentation structure with a closer proximity match than any single type of clustering alone, thus enabling increasingly effective presentations for the user. This section shows how generalized proximity generates hierarchy and sequence, both independently and together, using potentially complex measures of proximity.

Abstracted Proximity. Having a presentation structure-building mechanism that is independent of either property-, link- or axis-based clustering enables the proximity measure to also be independent of these clustering types. One result is that anything at all can be the proximity computation: it does not have to be any technique we describe in the context of the three clustering types. The human designer of the system can create any proximity computable measure. Consequently, this proximity calculation can be any combination of the means described with clustering, along with any means determined beyond clustering.

Hierarchy Derived from Generalized Proximity. One algorithm for maximizing the correspondence between the hierarchically communicated proximity and the repository proximity is the *quartet search* [2]. This generates a binary tree with the original nodes as the leaf node level. One example application is taking MIDI-encoded files of a very broad collection of well-known songs, compressing the concatenation of each possible pair, taking the efficiency of each compression as the proximity measure between the pair, and then converting the diagonal table of pair proximities into a binary tree. The resulting hierarchy was a taxonomy of genres and then artists that experts recognize and agree with. Quartet search can also process a proximity table matching concepts returned by our system’s initial query to generate a hierarchy over them. This clustering accepts any proximity table, regardless of what determined the proximities.

Proximity Tables from Shortest Paths. Quartet search can determine hierarchies from link graphs by using proximity tables derived from shortest path searches. The proximity entry for each object pair is an inverse of the distance of the shortest path found between those objects. This distance is the sum of the weights of

the links traversed along a shortest path. Each link weight is a conceptual distance, which is inversely proportional to the conceptual proximity between the nodes that link joins. A quartet search then optimizes the global proximity the hierarchy represents based on a graph of conceptual distances. This provides a generalized means of utilizing a wider variety of chained conceptual connections between objects when placing them in a hierarchy. Then challenge then becomes determining from these shortest paths meaningful means for conveying the natures of the groupings in the final presentation.

Conveying Rationale for Hierarchies Derived from Generalized Proximity. This generalized processing of proximity tables only determines the existence of groups. It cannot state reasons for why the groups formed that go beyond the abstracted numeric proximity measures. Here, techniques such as finding representative objects may help give some meaning to the group.

Emulating Property-based Clustering. Weighted concept lattices measure proximity between two concepts as the sum of the weights of the property types in the property assignments that the concepts share. A proximity table generated by this calculation processed using quartet search will generate roughly the same results as our earlier implementation of weighted concept lattices.

Sequence Derived from Generalized Proximity. A different algorithm may apply to sequence than was used for grouping in the same document. With distance-based sequencing, for any group, even if it was chosen by distance metrics, we can consider its fully connected weighted graph from pairwise distances and apply a shortest path algorithm to find a sequence. This makes sequences in which each presented sibling is the most similar to the siblings before and after it. We introduce here an algorithm for applying sequence to a tree and optimizing this sequence in terms of sequence derived from the repository proximities. We call this *polarizing* each binary pair in the tree, at all levels. As with forming the tree, we apply a greedy search algorithm to approach the best sequence for the tree. The three measures for sequence proximity correspondence are *leaf node*, *depth-first* and *recursive*.

Simultaneous Derivation of Hierarchy and Sequence. Either a hierarchy-first or a sequence-first presentation structure-building algorithm threatens to focus too much on the quality of either the hierarchy or sequence at the expense of the other. Treating both hierarchy and sequence as equal partners opens up more possibilities for further improvement of overall presentation structure. Since developed clustering techniques typically do not account for sequence, we propose the need for a variation of cluster-based search that builds sequence as the same points it builds hierarchy. Such an algorithm would process each pair of objects in inverse order of their proximity in adding that pair to the overall structure. Each step would change the hierarchy, sequence or both with equal consideration.

7. BEYOND PROXIMITY FOR FURTHER PRESENTATION IMPROVEMENT

There are quality measures other than conveyed conceptual proximity to apply to the structured progression generated. These include consistency and balance. Several *pruning* techniques

modify the tree structure to improve these measures. By both allowing any measure of the quality of presentation structure and providing means to incrementally modify this structure to improve the resulting measurements, this section presents a general algorithm for maximizing a presentation structure's conveying of underlying proximity that can incorporate any combination of any computable measure of both pairwise conceptual proximity and overall document structure.

Generalized Greedy Climb Search. We present here our modeling of the improvement of the structured progression as a best-of search. Our architecture has formulas for clustering into the hierarchy, potentially with separate proximity measures for sequencing with it. This section describes measuring formulas that help alter the discourse tree generated. These formulas may take the importance of following the original discourse tree and be able to weigh this against other considerations, including balance and consistency, and perhaps other things as well. Humans make these considerations when designing structured documents. However, because these algorithms combine complex clustering with best-of search, they may take very long to compute.

Consistency. We consider consistency to be the grouping of all siblings in one group based on the same concept. For example, it is consistent to give a certain section subsections that are all based on different genres, rather than have one be a genre, another an artist and a third a time period. You might need to change the default document tree to get consistency, so your choice comes again from measuring alternatives. Figure 4 shows an example of a presentation generated by our demo with universal consistency, which in our implementation is having all groups form around the same property type. In comparing presentations with and without universal consistency, we find the advantage of clarified hierarchical basis often comes at the cost of a less-balanced, sometimes bulky hierarchical form.

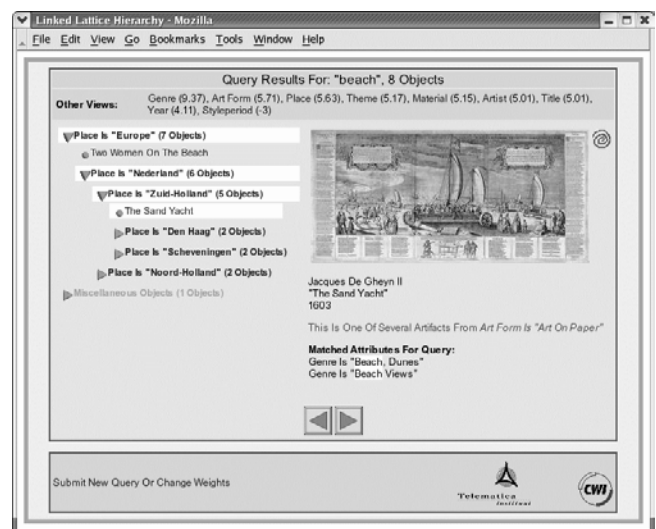


image © Rijksmuseum Amsterdam, used with permission

Figure 4. A Presentation Generated with Consistency

Balance. One factor is how balanced the resulting tree is. Other things being equal, balanced is better because it makes the user's traversal and hyperlinked access through the document more efficient. In a given point in a document, it may make a better document to apply localized pruning to improve balance than to follow the original grouping provided. On the other hand, the original grouping provided can often be more clear to the user than pruning applied to it, so you have to weigh the alternatives: does the balancing provide more clarity than the pruning removes?

8. CONCLUSION

This architecture for generalized proximity and structure metrics enables much extension and experimentation. It also provides insights for deriving coherent presentations from semantic annotations in general. We wrap up here by presenting summary of this paper, potential extensions and some lessons learned.

Summary. We extend our use of clustering for generating general presentation structure from semantics. Link-based and axis-based clustering join property clustering as means for deriving hierarchy and sequence. A general architecture based on proximity expresses and enables the combination of all this clustering for both hierarchy and sequence. Incorporating search into this architecture enables further improvement of presentation generation. This search model improves structure further still by combining proximity with other considerations such as presentation structure balance and consistency. Each of these clustering types, proximity measures and structural considerations has distinct impacts on the user's experience with the presentations generated. This paper ends with an algorithm for combining any measures for both conceptual proximity and presentation structure goodness, which forms the basis for specifying "style" for deriving general presentation structure from underlying semantic hyperlink graphs.

Future Work. While our system now has extended means of generating broader-scale document structure, much of what makes human-crafted presentations informative and coherent is the discourse conveyed at lower levels of detail. Since other research builds systems that approach computing richer discourse in small-scale presentation components, combining our system with theirs would approach automatic generation of larger presentations that are coherent and information on the large and small scales. Hera [11] and Disc [4] are examples of such systems with potential for combination with ours.

Insights Gained. By making use of explicit semantics of annotations of media objects we have been able to generate simple hierarchical and sequential discourse structures. It is clear that these structures are inadequate to emulate the complex discourse of human communication, but these early results lead us to believe that more useful presentations can be generated than one-dimensional relevance ranking lists.

ACKNOWLEDGMENTS

This work was funded by the Topia project of the Telematica Instituut, sponsored in part by IBM. William van Dieten of the Telematica Instituut programmed many of the demo interface components introduced here. Rudi Cilibrasi at CWI for improved our insight into clustering from proximity matrices. Geert-Jan Hoeben and Flavius Frasinicar of the Technische Universiteit Eindhoven helped understand Hera and its relationship with Topia. We thank the Rijksmuseum Amsterdam for their permission to use their Website's database and media content [8].

REFERENCES

- [1] Berkhin, P. Survey of Clustering Data Mining Techniques, Technical Report, Accrue Software (San Jose, CA, USA, 2002).
- [2] Cilibrasi, R. and Vitanyi, P. Clustering by Compression. arXiv.org e-Print cs.CV/0312044 (2003).
- [3] Fluit, C., Sabou, M. and Van Harmelen, F. Supporting User Tasks through Visualisation of Light-weight Ontologies. S. Staab and R. Studer (eds.) Handbook on Ontologies in Information Systems (2003, Springer-Verlag), 415-434.
- [4] Geurts, J., Bocconi, S., Van Ossenbruggen, J. and Hardman, L. Towards Ontology-driven Discourse: From Semantic Graphs to Multimedia Presentations. In Proceedings of the Second International Semantic Web Conference (ISWC2003) (Sanibel Island, Florida, USA, October 20-23, 2003), 597-612.
- [5] Gibbins, N., Harris, S. and schraefel, m. Applying mSpace Interfaces to the Semantic Web. University of Southampton Electronics and Computer Science EPrint 8639 (November 30, 2003).
- [6] Hearst, M., English, J., Sinha, R., Swearingen, K. & Yee, P. Finding the Flow in Web Site Search. Communications of the ACM 45, 9 (2002), 42-49.
- [7] Modha, D. and Spangler, W. Clustering Hypertext with Applications to Web Searching. In Proceedings of the eleventh ACM conference on Hypertext and hypermedia (HT00) (San Antonio, Texas, USA, May 30 - June 3, 2000), 143-152.
- [8] Rijksmuseum Amsterdam, Rijksmuseum Amsterdam Website. <http://www.rijksmuseum.nl/>.
- [9] Rutledge, L., Alberink, M., Brussee, R., Pokraev, S., van Dieten, W., and Veenstra, M. Finding the Story — Broader Applicability of Semantics and Discourse for Hypermedia Generation. In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia (HT03) (Nottingham, UK, August 26-30, 2003), 67-76.
- [10] Rutledge, L., van Ossenbruggen, J., and Hardman, L. Structuring and Presenting Annotated Media Repositories, CWI Technical Report INS-E0402 (February 2004).
- [11] Vdovjak, R., Frasinicar, F., Houben, G.-J., and Barna, P. Engineering Semantic Web Information Systems in Hera. Journal of Web Engineering, 2, 1-2 (2003), Rinton Press, 3-26.