Sequence and Emphasis in Automated Domain-Independent Discourse Generation

Martin Alberink Telematica Instituut
P.O. Box 589 NL-7500 AN Enschede The Netherlands
+31 53 4850485 Martin.Alberink@telin.nl

Lloyd Rutledge CWI
P.O. Box 94079 NL-1090 GB Amsterdam The Netherlands
+31 20 5924127 Lloyd.Rutledge@cwi.nl

Mettina Veenstra Telematica Instituut
P.O. Box 589 NL-7500 AN Enschede The Netherlands
+31 53 4850485 Mettina.Veenstra@telin.nl

ABSTRACT

For humans to gain comprehensive views of large amounts of repository contents, they need to have insight into the relations among information objects. It is a challenge to automatically generate presentations of repository contents, through, for example, search results, which reveal such relations to readers. Such presentations must reflect properties of information objects such that large sets of information objects appear as a coherent whole. An approach to this is generation of discourse structures that convey such properties of information objects in presentations. Semantic Web technology provides a conceptual basis for generation of discourse in Web-based information environments.
This paper describes automatic generation of sequence and emphasis in presentations of information objects. It shows generation of object sequences and emphasis in accordance with a user input of relevance of information attributes in our Topia architecture. The resulting presentations allow users to encounter information objects in decreasing order of relevance. This makes it easier to identify relevant information objects among many others, as well as to observe their relations with the other information objects.

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, Standardization.

Keywords
Discourse, Narrative, Coherence, Semantics, Sequence, Order, Emphasis, Hypermedia, Cluster, Semantic Web, RDF.

1. INTRODUCTION

Search engines on the Web typically generate presentations of retrieval results as plain lists of links to information objects, possibly sorted according to relevance or hyperlink connectivity. Such presentations do not easily allow users to assess sets of retrieved information objects as a whole, since this requires that users inspect the retrieved objects one by one. This inhibits users from readily identifying the information objects that are relevant for their information need. Structuring sets of repository contents in coherent presentations, taking preferences of individual users into account, would supposedly facilitate users orientation in presentations of large amounts of information objects. Semantics of electronic content on the Web encoded in Semantic Web technology [7] provide a basis for deriving relations among information objects in Web-based information environments. Such relations, when included, could add to coherence in presentations of information objects.
Our focus is on automated generation of coherent presentations of database contents in order to allow users to find their way in large information collections. We aim at enhancing coherence in presentations of sets of information objects by transforming semantics encoded in RDF into constructs of common discourse that are meaningful for human users. Well-known discourse, such as narrative, conveys relations among information objects in addition to the information itself, such as by means of story lines, sequences, emphasis and focalisation [3]. Automatic production of such rich discourse typically produced by human authors remains elusive. Instead, we aim at generating simple but commonly encountered discourse constructs that can be based on attributes and relations, a typed of semantics supported by the Semantic Web. This papers focus is on automated generation of two such basic discourse constructs in presentations of information objects: sequence and emphasis on information objects in presentations. Sequences show interrelations among a set of objects in a semantic dimension, such as time, place or causality. Emphasising objects conveys a distinct property of such objects with respect to the other elements or a special relation with the other elements. Our presumption is that the resulting presentations of retrieval results enable users to assess the contents and relevance of information objects faster and with less effort compared with the common lists of search results. This, we hypothesize, assists users in deciding on navigation and exploration directions while traversing the information space, in order to help users grasp the contents of information repositories and discover what they find relevant or useful [11].
Section 2 of this paper discusses the approach in this paper in relation to other research. Section 3 describes the Topia (Topic-based Interaction with Archives) project [18] that produced the results described in this paper. Automated generation of sequence and emphasis

as discourse constructs in web environments is described in sections 4 and 5 respectively. Section 6 explains involvement of a user statement of relevance of relations in the automated generation of sequence and emphasis in hierarchical presentations of search results. This section shows that the resulting presentations direct readers to the information relevant for them in the search result, while preserving directions to the other retrieved objects. Section 7 shows that the resulting presentations are capable of structuring retrieval results in different perspectives. Sections 8 describes future work on the topic in this paper and section 9 wraps up this paper with a summary and conclusions from this work.

2. RELATED WORK

A number of research projects discussed in this section focus on automated discourse generation in presentations of content stored in digital systems. Their approaches differ in three senses: the balance between human-specified and computer-inferred semantics for discourse generation, the types of discourse constructs that transformation of semantics results in, and the way of presenting discourse structures by conveying relations among information objects. Sections 2.1 through 2.4 position related work in the range from almost completely human specified discourse structures to discourse with high-level human specification only. They also discuss the position of sequence and emphasis in these different approaches.

2.1 Fixed discourse frameworks

Underlying frameworks of automatically generated discourse structures are in the range from nearly fixed to largely flexible. At the one extreme are nearly complete and rigid presentation structures that only leave room for objects to be inserted. Presentations of retrieval results of search engines fall in this category. Such presentations typically contain hyperlinks to retrieved items in straight lists. Application of sequence and emphasis to lists of retrieval results can convey relations among information objects and relations between information objects and information needs of users.

2.2 Template-based discourse

Templates that specify discourse structures are a step provide more flexible discourse frameworks than lists of retrieval results. Gaps in such templates allow insertion of information to fill in the contents of the story. Computer-generated sequence and emphasis in such presentations are bound to the information in each gap. It is important that the information filling the gaps is coherent in its connection with the template. The Artequakt project has a template-based approach focusing on discourse of textual biographies using narrative templates [1]. Sequence and emphasis in textual information is contained in the text itself. For text that originates from natural language generators, the coherence of the text as well as its connection with the template are important sequence criteria.

2.3 Semantics-based discourse

Geurts approach focuses on generation of discourse based on domain knowledge, straight from semantic information [10]. Discourse of specific types can be generated, such as biographies and curricula vitae. Such discourse requires semantics-based sequences and emphasis, since it should be in accordance with their usual contents and structure. Automatically generated sequences in such presentations should be in accordance with user expectations in order to make such presentations coherent.

2.4 Semantics-driven discourse

At the other extreme along the line of discourse framework flexibility is discourse without human-specification of the discourse structure. Such discourse results from characteristics of semantics that abstract from the meaning of the semantics itself but are based on the occurrence of the semantic relations only. Our Topia architecture generates presentations with such discourse structures and applies sequence and emphasis in the discourse. Sequence and emphasis are capable of adding a fraction of the semantics that human authors can generate. They are however universal across application domains. Further in this paper, section 6 explains how sequence and emphasis can be generated such that they are in accordance with relevance of objects for individual users. Section 7 shows that this principle can produce discourse that show information in different perspectives.
Section 3 explains how the Topia architecture derives relations as well as the discourse and presentation structures.

3. TOPIA ARCHITECTURE

The research described in this paper is part of the development of the process architecture of the Topia project. The Topia architecture automates generation of presentation structures of retrieval results with discourse constructs [18]. Figure 1 shows its four phases. The information objects in Topias repository are 740 artefacts from the art collection of the Rijksmuseum Amsterdam [17]. Attributes of the artefacts are encoded in about 64,000 RDF triples.

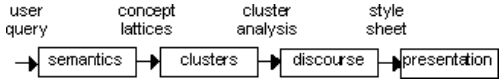

Figure 1. Topia architecture overview [18]

Users access the Topia repository by specifying queries. After retrieving a set of artefacts together with their attributes in the first stage, the second stage generates a concept lattice: a structure of clusters of information objects and the attributes they have in common in a

subsumption graph [9]. The third stage transforms clusters and subsumption relations in a concept lattice into a conceptual presentation with discourse constructs. The final stage specifies the layout, the presentation of recurrent themes and the interaction with users in an HTML or SMIL presentation, generated by an XSLT style sheet.

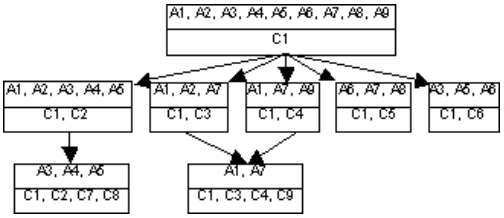

Figure 2. Cluster graph of concept lattice from Table 1 [18]

Table 1. Artefacts mapped against properties in a concept lattice for query on "water" [18]

(C1) "Water"	(C2) Genre: Water, ice and snow	(C3) Genre: Dutch landscapes	(C4) Genre: Field meadows	(C5) Genre: Buildings in landscapes	(C6) Artist: Jacob van Ruisdael	(C7) Genre: Tree forests	(C8) Genre: Riverscapes	(C9) Artist: Paul Joseph Constantin Gabriel
"A watercourse at Abcoude" (A1)	X	X	X					X
"Watercourse near 's-Graveland" (A2)	X	X						
"Mountainous landscape with waterfall" (A3)	X				X	X	X	
"A water mill" (A4)	X					X	X	
"Landscape with waterfall" (A5)	X				X	X	X	
"Water mill" (A6)				X	X			
"Windmill on a polder waterway, known as 'In the month of July'" (A7)		X	X	X				X
"A waterside ruin in Italy" (A8)				X				
"The battle of Waterloo, 18 june 1815" (A9)			X					
Concept Size	**5**	**3**	**3**	**3**	**3**	**3**	**3**	**2**

The concept lattices generated in the second stage of the Topia architecture not only contain all individual artefacts in a retrieval result, but also all clusters of artefacts in a retrieval result that have one or more attributes in common. Each cluster of artefacts together with their common set of attributes is a concept in the concept lattice. Concept lattices subsume concepts under other concepts that contain their smallest supersets of artefacts, in a directed graph [9]. As an illustration, Table 1 shows the retrieval result of the query specifying the string "water" in the title of artefacts. The rows are the titles of the retrieved artefacts and the columns are attributes of one or more of the retrieved artefacts. The crosses in the table indicate the occurrence of the corresponding attribute for the object concerned. Figure 2 partly shows the concept lattice that results for this retrieval result, generated by the Topia architecture. The concepts are the pairs of adjacent bars with the objects printed in the upper bar and the attributes in the lower. The set of common attributes expresses what the relation is among the set of objects in each concept. For example, Figure 2 shows that artefacts A3, A5 and A6 have C1 and C6 as common attributes.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <sections size="8">
- <section>
<title>material is "Oil on canvas"</title>
- <sections size="4">
- <section>
<title>artist is "Jan Willem Pieneman",
genre is "Battles, Group, Group portraits,
Historical scenes",
theme is "Struggle and Strife",
title is "The Battle of Waterloo, 18 June 1815"
and year is "1824"</title>
- <sections size="1">
- <section>
<title>The Battle of Waterloo, 18 June 1815</title>
<date>1824</date>
<artist>Jan Willem Pieneman</artist>
<image>SK/Org/SK-A-1115.org.jpg</image>
</section>
</sections>
</section>
- <section>
<title>theme is "Netherlands and the Water"
and title is "Landscape with Waterfall"</title>
- <sections size="1">
<section />
</sections>
</section>
- <section>
<title>place is "Den Haag"</title>
- <sections size="2">
<section />
</sections>
</section>
- <section>
<title>title is "Mountainous Landscape
with Waterfall"</title>
- <sections size="1">
<section />
</sections>
</section>
</sections>
</section>
- <section>
<title>genre is "Water, ice and snow"</title>
+ <sections size="3">
</section>
- <section>
<title>place is "Amsterdam"</title>
+ <sections size="2">
</section>
- <section>
<title>genre is "Buildings in landscapes"</title>
+ <sections size="2">
</section>
- <section>
<title>artist is "Jacob van Ruisdael"</title>
+ <sections size="3">
</section>
- <section>
<title>material is "Oil on panel"</title>
+ <sections size="3">
</section>
- <section>
<title>genre is "Dutch landscapes"</title>
+ <sections size="2">
</section>
- <section>
<title>genre is "Fields, meadows"</title>
+ <sections size="2">
</section>
</sections>
```

Figure 3. Conceptual presentation generated by the Topia demo

Subsumption edges imply a relation between the clusters of artefacts and attributes in the concepts they connect: traversing subsumption edges in an upward direction leads to a more general concept, since such a concept has more objects and fewer attributes than the one traversed from. Likewise, traversing subsumption edges in a downward direction leads to a more specific concept.

The third stage of the Topia architecture generates hierarchical conceptual presentations by flattening the directed acyclic graph structure of concept lattices. Hierarchically organised structures are commonly used backbone structures, such as in books subdivided in chapters, sections and paragraphs, to facilitate orientation by human readers. The conceptual presentations specify the clusters of information objects and relations among the objects by means of the common attributes. Figure 3 shows the conceptual presentation of the retrieval result of the query “water”, while Figure 4 shows the presentation of the concept lattice on the screen.

Figure 4. Presentation generated by the Topia demo

Sections 4 and 5 focus on automated generation of sequence and emphasis respectively from universal aspects of semantic annotations. The sections also describe the support that web standards offer.

4. SEQUENCE

Sequences of objects in presentations convey to readers an order of objects along a certain dimension. Consequently, readers of presentations expect sequences of objects to be meaningful so that they have to be in accordance with logical and, if possible, useful sequence criteria. This section discusses organisation of sequences of objects in presentations at the four phases in the presentation generation process, namely semantics, discourse, presentation structure and style.

4.1 Semantics

Semantics imply domain-independent sequences of concepts in information repositories in multiple ways. First, sequences follow directly from explicitly ordered sets of objects. The RDF recommendation contains a <seq> class for explicitly ordered collections, while RDFs <bag> class supports an unordered collection of objects and RDFs <alt> class supports collections of objects that are equivalent in some sense. Second, sequences follow from the occurrence of a relation between subsequent objects, such as chains of objects with identical relations. Semantics encoded in RDF triples allow derivation of such sequences by examining the attribute-object pairs of subjects. Third, sequences follow from numeric characteristics, such as weights indicating relevance of objects. Semantics encoded in RDF allow identification of numeric quantities, since the data type specification in XML schemas reveals whether items are of a numeric type. Fourth, inference rules allow derivation of relations between objects that order objects as sequences, such as for generation of rich

narrative sequences. The first three sequence criteria are universal, since they are independent of the semantics themselves of the attribute instances.
Section 4.2 discusses sequence as a discourse construct for relating information objects in presentations.

4.2 Discourse

Many common ways of presenting a set of elements imply a notion of sequence [19]. It is important that sequence criteria make sense to users. For sequences of objects to be comprehensible and meaningful for humans, they should be arranged according to similar characteristics. Complexity, specificity, and causal relations between subsequent objects are general sequence criteria that need specification. The specification determines their semantics. Sequences of objects can result from mapping characteristics of these objects to other characteristics that are sequence criteria, for example through inference rules. As an example, events can be related to their period if periods are expressed as numbers, such as years, which allows a chronological sequence of events. The meaning of the resulting sequences depends on the sequence criterion, so that it is possible that sequences are not useful for readers.
Hierarchical presentation structures, such as Topias conceptual presentation structures, contain subsumption of multiple concepts under other concepts. A depth-first traversal of a hierarchical object structure is a sequence of specificity with specialisation and generalisation steps to lead readers through hierarchies in a comprehensible way. Concepts subsumed under a certain concept have equal position in the hierarchy. In the Topia conceptual presentations, the sequence of concepts subsumed under a concept is according to the relevance criterion explained in section 6. The sequences of artefacts within concepts are according to year of creation, a manually chosen numeric criterion [18]. However, the relevance criterion is as well applicable to individual artefacts in a concept.
Common objects of subsequent clusters can be the transition from the one cluster to the next. Maximisation of this type of transition for a given set of clusters can be a criterion for sequence of clusters. For this it should be possible to arrange the presentation ordering of the contacts of the two clusters such that the common object is the last item in the first cluster and the first item of the next. Such transitions save presentation space since the common objects are presented once for both clusters. It acts as a conceptual transition, a segue, between the clusters. Segues improve the aesthetics of the presentation and help convey the relation between the groups.

4.3 Presentation structure

In earlier work, we propose presentation structures in hypermedia for the sequence nucleus type in Rhetorical Structure Theory [15] conveying sequence. These presentation structures are bookshelf order (the order of stacked bookshelves), temporal order and next-buttons for navigating to the next node in a sequence [19]. In this earlier work, we also suggest presentation structures that contain hardly any notion of sequence in a set of objects compared to common presentation structures, namely random arrangements of objects, patchworks and grid structures. Scattering objects by these structures reduces an implied notion of sequence. Such presentations devoid of an implication of sequence avoid the risk of presenting information in sequences that have no meaning to users. Alternatively, criteria that presumably help users assess the content of presentations can be a basis for sequences of objects. This section discusses presentation structure devices for representation of sequence in three aspects of hypermedia presentations: space, time and link structure.

4.3.1 Space

Conveyance of sequences in space requires that the objects in the sequence be positioned in space with respect to each other, such that usual reading directions of humans imply the sequence. The relative distance between subsequent objects conveys the relative distance in the attributes or relations that are the basis for the sequence. Two-dimensional media can express two-dimensional sequences by putting objects in tables for conveying sequences according to two criteria. CSS has properties for supporting the positioning of objects required for the above-mentioned structures.

4.3.2 Time

Time-based presentations suggest a sequence running from the beginning of the time series to the end. This inherent sequence in time-based presentations strongly implies sequence. The sequence can however be adjusted by flashbacks and flash-forwards, resulting in presentation sequences that can disturb the expected ordering. Time-based presentations can convey development of real events in time. In addition, time-based presentations can convey ordering of space, such as in guiding tours [20]. Time-varied transitions convey the relative distance between subsequent objects, as well as the beginnings and ends of sequences that are put in concatenation. SMIL-enabled web-based presentations can contain the above-mentioned features in progressions in time.

4.3.3 Links

Sequences in navigation structures guide users through one or more paths of nodes in sequences specified by the navigation structure. Links in such navigation paths can represent relations between subsequent objects, spatial relations or separations of nodes applying to different events in time [20]. The hyperlink construct in the HTML standard supports these techniques.

4.4 Style

Sequences are typically presented in lists of ordered items. The CSS property list-style-type conveys sequence, or lack thereof, to the user. Most of its values prescribe numeric systems, typically numbers that precede the display of the element's children. The numeric system values correspond with the <ol> element in HTML, specifying an ordered list. These numeric systems emphasise that the displayed items fall in a sequence. The remaining values, such as disc and circle, correspond instead with the <ul> element, specifying an unordered list. They potentially communicate that the list is not necessarily a sequence.

5. EMPHASIS

Emphasis on objects in presentations indicates to readers that such objects have one or more properties, such as relevance, that distinguish the objects from other objects. Consequently, viewers of presentations expect emphasised objects to be worthy of note in some sense.
This section discusses derivation, from semantics, domain-independent distinguishing features, which are expressible in presentations by using emphasis. The discussion concerns the four phases in the presentation generation process, namely semantics, discourse, presentation structure and style.

5.1 Semantics

Semantics imply domain-independent distinguishing features of concepts in information repositories in multiple ways. First, distinguishing features follow directly from annotations that explicitly express that concepts are distinct with respect to other concepts, or distinct for specific users. Second, distinguishing features follow implicitly for concepts with attributes or a combination of attributes that not many other concepts have. Similarly, distinguishing features follow for concepts with attributes or a combination of attributes that are relevant for specific users. These latter cases of implicit distinguishing features are universal, since they are independent of the semantics themselves of the attribute instances. Appropriate presentation of concepts with such distinguishing features is dependent on the degree and nature of the features that distinguishes such concepts.
Semantics encoded in RDF triples [13] allow derivation of distinguishing features of RDF subjects by examining the attribute-object pairs of subjects for particularity with respect to other RDF subjects.
Section 5.2 discusses emphasis as a discourse construct for relating information objects in presentations.

5.2 Discourse

Distinct discourse characteristics of information objects suggest distinguishing features of such objects, and emphasises such objects with respect to other information objects. Examples of distinct discourse characteristics are central or extreme representations of information objects or groups of information objects, additional discourse characteristics and annotations. Variations of intensity, position, distance or direction of information objects in presentations convey such distinct discourse characteristics, emphasising the objects concerned.
Regularity in discourse characteristics suggests a thread running through a presentation tying it together. Such regularity can be repetition of objects or specific types of objects, objects with consistently applied specific discourse characteristics, and rhythm, being a fixed structure of repetition. Such threads are conceived as prominent themes, express emphasis on the objects involved and thus allow focalisation of presentations. Broken regularity, such as absence of objects or discourse characteristics at some positions in otherwise regular structures, suggests exceptions.
Concepts in concept lattices are themes characterised by the attributes that the objects in a concept have in common. Regular discourse structures convey such themes. The subsumption structure of concept lattices is also a regular structure, since downward traversal invariantly results in specialisation and upward traversal in generalisation.
Concepts with many objects or attributes compared to other concepts are distinct concepts, as well as concepts with objects and attributes that are relevant for users. A relatively dense interconnection structure of concepts in concept lattices is also a distinguishing feature of the concepts involved. Putting such distinct objects at central or extreme positions in discourse structures emphasises such objects for users.

5.3 Presentation structure

Presentation structures specify the relations among objects in presentations while abstracting from the physical aspects of presentations. Presenting distinct objects in a way that is different in some sense from the presentation of other objects emphasises such objects. In hypermedia presentations, putting distinct objects at prominent positions, such as at a central or extreme position emphasises such objects, as well as association of additional objects such as text, images or symbols with such objects. Regular structures in one of the dimensions of hypermedia convey themes. This section discusses presentation structure devices for representation of emphasis in three aspects of hypermedia presentations: space, time and link structure.

5.3.1 Space

From a layout point of view, putting distinct objects at a central position, such as in the middle of the screen, or at extreme positions, such as on top of the screen, emphasises such objects. Alignment of objects along a spatial dimension conveys a stratification of emphasis on objects. Examples are distribution of objects in a lattice structure, indentation for indicating levels in hierarchical structures and the organisation of books, where titles of chapters are on top of pages and footnotes at the extreme bottom. In addition, alignment of objects conveys a regular structure of themes. In the hierarchical conceptual presentations generated by the Topia architecture, spatial grouping of branches conveys the fact that they are a theme, corresponding with a concept. Alignment of concepts and artefacts in the orientation bar conveys a stratification of emphasis that is related to the number of objects in concepts. CSS elements support positioning and alignment of objects in HTML presentations.

5.3.2 Time

Putting distinct objects at the beginning or end of a time sequence emphasises such objects. Increased presentation duration of objects in time-based presentations also emphasises objects. Variable pacing allows conveying a stratification of emphasis, typically by slowing down the pace proportionally to emphasis. Flashbacks and flash-forwards emphasise objects or events and allow repetition and regular

structures. Temporal grouping, rhythm and fixed-length pauses also convey regularity. Players or browsers that support SMIL enable web-based presentations with the above-mentioned features in progressions in time.

5.3.3 Links

Objects linked to from many places in navigation structures emphasises such objects. Such objects can be central nodes that act as home or start pages of, for example, web sites such as portals. Furthermore, objects that have links to other objects have emphasis with respect to objects without links to other objects. In addition, names of links can express emphasis since they can contain an identification or annotation of the link. These techniques are all supported by the hyperlink construct in the HTML standard.

5.4 Style

The classical type of technique for emphasising objects is highlighting them in order to give emphasised objects distinct presentation characteristics with respect to other objects. Techniques for highlighting are setting objects size, use of different fonts and colours, flashing objects, use of icons such as arrows and frames around objects. A feature such as frame size conveys the intensity of the emphasis, and colour possibly the type of emphasis.
Style features such as colour are applicable to individual objects and do not inherently constrain the presentation of other objects. Style features allow addressing individual objects for emphasising. However, possible unintended effects of a combination of style features in presentations should be avoided, while in addition style can affect presentation of information and its presentation structure [16]. As examples, background colours should not mask colours in the media items or conflict with them, and application of many different fonts may inhibit readability.

6. USER CONTROL

In presentations of hierarchical structures, a meaningful sequence of a set of concepts that are subsumed under a concept provides readers with a means of relating the subsumed concepts to each other according to the applied sequence criterion. If the sequence criterion is relevance of objects, readers reading the sequence from beginning to end encounter each of the concepts in the sequence before all other concepts that are less relevant. A difficulty is that it is hard to tell beforehand what makes concepts relevant for users. We base the sequence of concepts subsumed under a concept on relevance for users, and consider a number of criteria that are optional relevance criteria for individual users. These criteria are the portion of the retrieval result covered by the objects, the amount of information available about the objects, and the relevance of the available information for individual users. We now explain how these relevance criteria relate to characteristics of concepts.
The first relevance criterion mentioned, being the portion of the total number of retrieved objects in concepts, is proportional to the number of objects in concepts. Consequently, we consider the number of objects in concepts as a measure of the concepts relevance.
The second relevance criterion, being the amount of information available about the objects, is proportional to the number of attributes of concepts. Consequently, we consider the number of attributes of concepts as another measure for the concepts relevance.
The third relevance criterion, being the relevance of the available information for individual users, requires that a specification of the relevance of attributes for users be available. Since users goals vary, different users consider different attributes as relevant. A way of letting users specify relevance of attributes is by requesting an assignment of positive numbers to attributes as relevance weights, such that higher numbers correspond to higher relevance of attributes. Since higher numbers of the previously mentioned relevance criteria also correspond to higher levels of relevance, a measure of the total level of relevance of a concept that follows from the three individual relevance criteria can be calculated according to the following formula.

$$R_{concept} = N_{objects} \times \sum_{i=1}^{N_{attributes}} W_i$$

In this formula, Rconcept is the relevance of the concept, Nobjects is the number of objects in the concept, Nattributes is the number of attributes in the concept and Wi is the weight of attribute i in a set of Nattributes attributes.
Multiplying the number of objects with the sum of the weights assigned to the attribute types results in their having equal effect on the resulting concept relevance, without their having to be of equal order of magnitude. Adding the number of objects to the sum of weights results in their having equal effect on the outcome only if they are in the same order of magnitude. Since the order of the number of objects in concepts generally increases with the total number of objects in the database, this would entail a need to bring the weights into accordance with the number of objects in concepts.
A set of weight values containing the values zero and one only allows users to designate attributes as either relevant or irrelevant without further distinguishing between the relevance levels.
The formula shows the calculation of the concept relevance when all three relevance criteria mentioned are involved. Leaving some of the relevance criteria aside requires adjustment of the formula. Excluding the first relevance criterion, being the number of objects in concepts, implies that the factor Nobjects must be removed from the formula. Excluding the second relevance criterion, being the number of attributes in concepts, implies that an additional division by the number of attributes in the concept must follow calculation of the resulting Rconcept. Excluding the influence of differently valued weights implies that the number of attributes Nattributes replaces the summation factor.
A sequence of presentation of siblings in decreasing order of concept relevance in hierarchical conceptual presentations results in readers encountering siblings in decreasing order of relevance. Emphasising concepts with relevance levels that exceed a certain threshold level, such as zero, allows users to identify the objects in presentations with the specified relevance level at a glance.
The Topia architecture puts siblings in hierarchical conceptual presentations in a sequence according to relevance as explained in this section. With their query, users specify the level of relevance of the types of attributes that occur with the retrieved information objects. Figure 5 shows the specification form. The form shows the weights as well as a direction for users for applying the weights. Users

specify one of six levels of relevance for each of these attribute types, or tick the extreme left column for specifying attribute types that should not be included in the presentation.
Attributes in the Topia repository have a type and a value. Topia allows user specification of the relevance of only the attribute types that occur in the retrieval result. Conceptually, users could as well be allowed to specify weights of attribute values. However, attribute types typically have many attribute values, resulting in a large amount of attribute values that occur in retrieval results. Letting users specify relevance for all of these requires considerable efforts. RDF encoded databases allow automated extraction of the attribute types and values of retrieved objects.

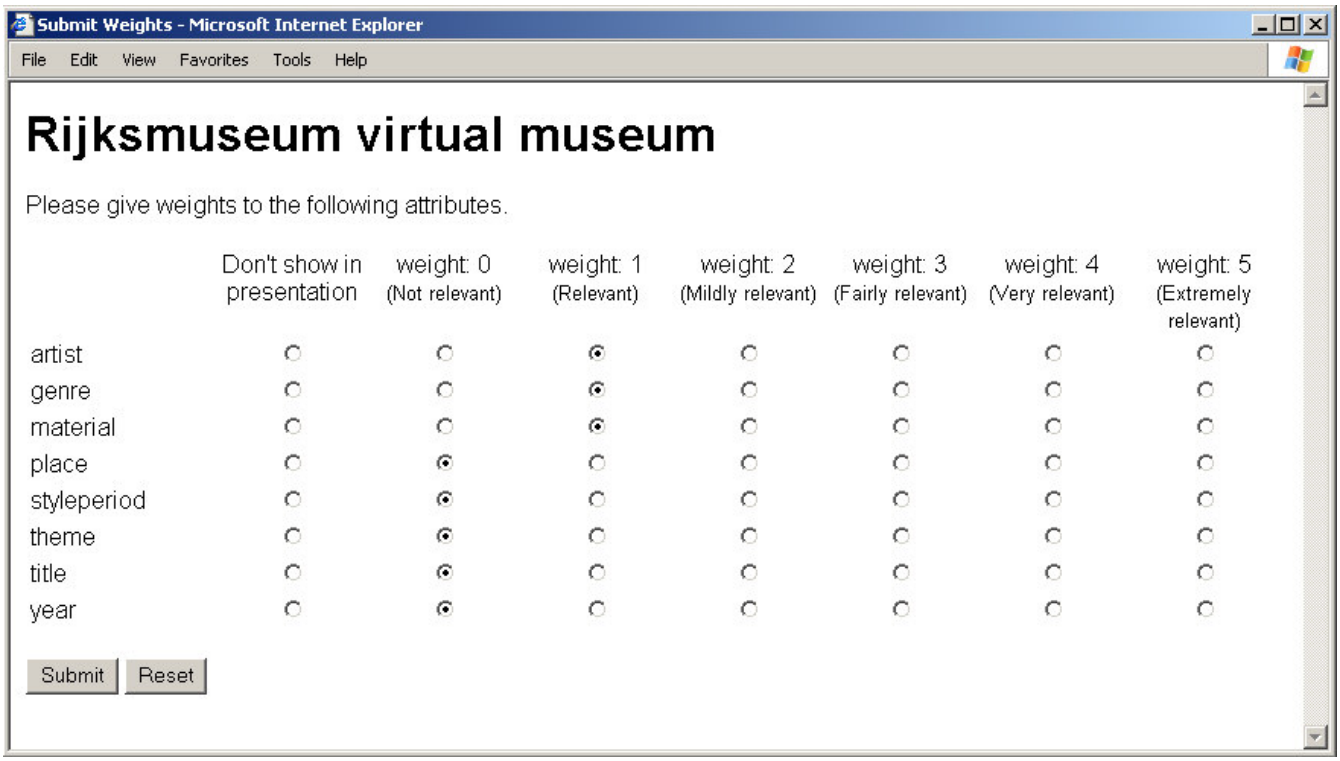

Figure 5. User specification of relevance of attribute types

To process the specified levels of relevance, they are assigned the integers from zero to five. Higher numbers in this range correspond to higher relevance levels, as shown in the table. The values of weights in the form are illustrative and not critical for a good performance of the sequence principle. In fact, users could be allowed to specify the weight values freely, allowing users to apply a weight distribution different from a set of successive integers. For calculating relevance of concepts, the Topia architecture applies the mentioned formula in order to involve all three stated relevance criteria.
A hierarchical list of concepts conveys the retrieval results, as described in section 3. Since people read from top to bottom, presenting the sequenced concepts from top to bottom requires a presentation device, so that at each hierarchical level, users encounter concepts in decreasing order of relevance. Emphasised concepts, with a relevance exceeding the threshold level, appear as blue links, while the non-emphasised are ghosted out.
Section 7 shows that sequences of siblings according to relevance of clusters for users as explained in this section allows focalisation of presentations to specific points of view.

7. DIRECTING DISCOURSE

Section 6 showed that sequence and emphasis in presentations position a set of information objects at a point in the story space. For users to obtain discourse that shows specific perspectives of sets of information objects requires an appropriate statement of relevance of attributes. This section shows how a statement of relevance of attributes results in discourse that give a corresponding perspective of a set of retrieved information objects. The discussion focuses on one of the relevance criteria only, being the attribute weights, since it is the only relevance criterion that relates to the contents of information objects.
Increasing the weights of specific attributes moves concepts with these attributes to the front of sequences they are part of, allowing users to encounter such concepts first. Consequently, in order to put discourse in perspective, attributes that are characteristic of the required perspective must have higher weights than others in order to give the corresponding clusters high relevance. To illustrate this with the Topia architecture, we consider a user who wants artefacts about the theme water and specifies a query “water” in the artefact title. Among useful perspectives for readers of the retrieval results are the perspective of the art domain on the one hand and the perspective of time and place on the other hand. Considering the attributes that occur in the retrieval result at the extreme left in Figure 5, the following weight configurations are in accordance with the two perspectives.
1. Perspective of art domain: attributes artist, genre and material have weight value 1, other attributes have weight value 0.
2. Perspective of time and place: attributes place and year of creation have weight value 1, other attributes have weight value 0.
Figure 6 shows a presentation in the art domain perspective resulting from the weight configuration stated in item 1. Concepts that have

attributes of type artist, genre or material appear above other concepts in the presentation sequence.
Figure 7 shows a presentation in the perspective of time and place resulting from the weight configuration stated in item 2. Concepts that have attributes of type place or year appear above other concepts in the presentation sequence.
In addition to users themselves, discourse domain experts can be involved in specifying the weight configuration of attributes for discourse with specific perspectives. Dynamic RDF encoded databases do not allow retrieval of an up-to-date set of attributes of information objects before the time of retrieval. Consequently, it is not known beforehand what attributes are available, which of the attributes relate to the required perspective and how they should be weighted to ensure a proper position of objects and attributes in the resulting discourse of the required type. A classification of attributes in the repository gives discourse domain experts a means for specifying the relevance of classes of attributes in presentations with specific perspectives.

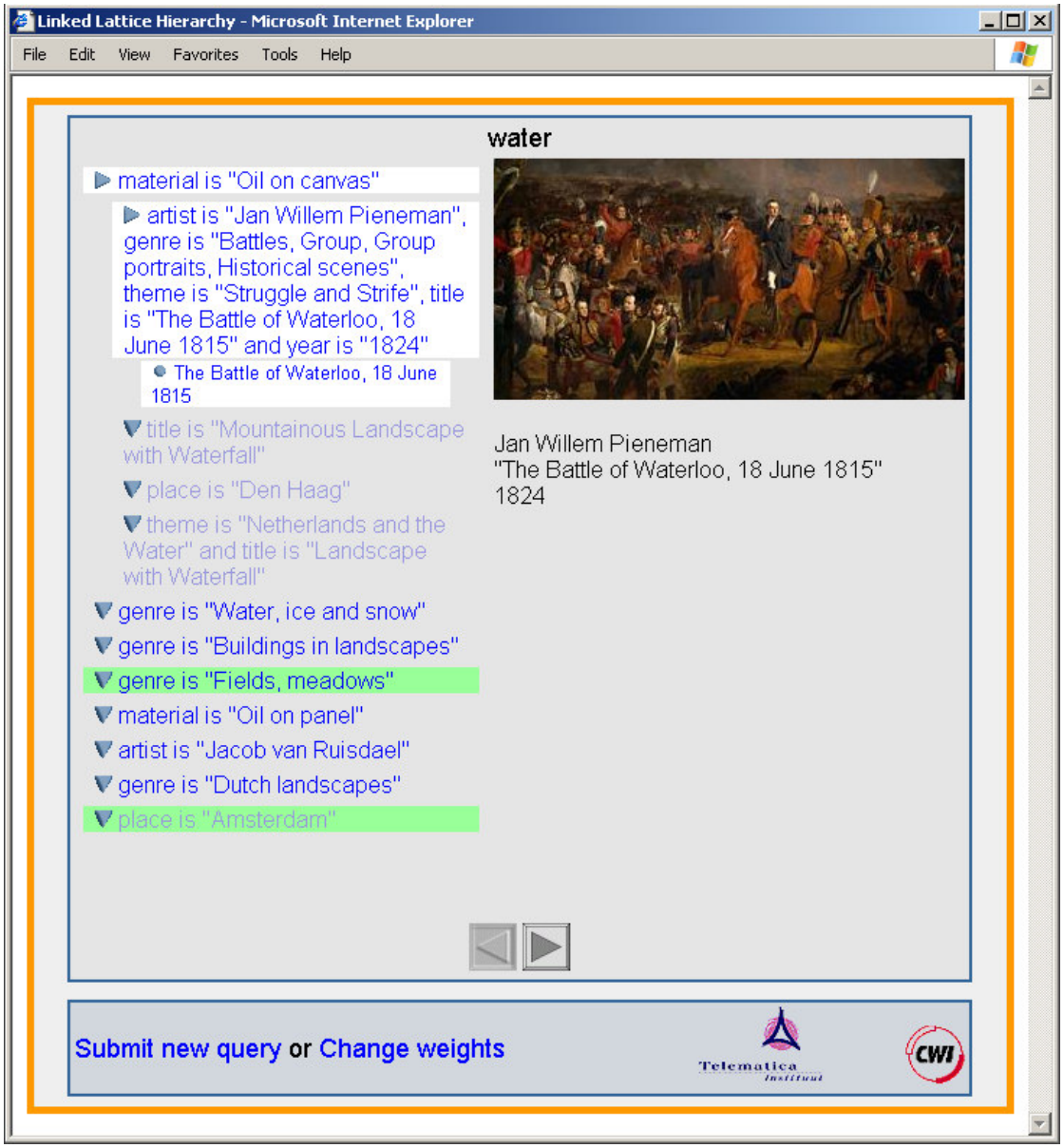

Figure 6. Discourse in perspective of art domain

Figure 7. Discourse in perspective of time and place

8. FUTURE WORK

The work presented in this paper bases automatically generated sequence and emphasis on the relative number of objects and attributes of concepts and on relevance of attribute types for individual users. Another domain-independent criterion for sequences and emphasis is the subsumption structure in concept lattices. The subsumption structure occurring in concept lattices depends on the occurrence and distribution of attributes among the retrieved information objects. Sequences of concepts can be based on their number of child concepts or parent concepts, while emphasis on concepts can be based on a high number of parent concepts or child concepts. Analysis of concept lattices reveals the presence of distinct structures such as central concepts or intensively interconnected clusters of concepts, which can be emphasised. Presenting such relevant and prominent characteristics of concept lattices by means of discourse constructs to convey patterns in the retrieval result will be a topic of future research.

Topias current implementation generates concept lattices based on exact match of attributes of information objects. Extension of the exact match criterion with measures based on proximity of attributes can potentially increase the number and quality of clusters. Clustering techniques exploiting proximity of attributes have found their application in data mining for partitioning sets of objects [5]. The type of clustering technique determines the properties of the resulting clusters and hence the type of coherence among objects in clusters. In order to let users experience the objects in the resulting clusters as semantically close, the required distance measure between attributes for clustering should be accordingly. In spite of the required tuning, density-based numeric clustering techniques take the distribution of numbers in the retrieved data set into account for generating clusters of objects with relatively small numeric distance between the objects. Such techniques can be particularly useful for clustering numeric properties, such as the year of creation of artefacts.

Vector space models of information objects in an attribute space have found common application to express similarities between information objects for information retrieval purposes [21]. Vector space models are a conceptual basis for clustering objects based on

non-numerical attributes and for calculating clusters similarity to user queries. Discourse constructs such as sequence and emphasis can express such cluster characteristics in presentations. Future work will extend the applied clustering techniques and focus on their presentation in discourse constructs.
Another application of sequences is for conveying themes as threads through concept lattices. Such themes can concern subsequent clusters of attributes that have a specific identical attribute, but that do not occur under the same concept in the concept lattice. The user statement of relevance of attributes can be extended to a user statement of themes to be presented as paths along subsequent clusters in presentations. We will focus on automatic generation of such themes by means of sequence and emphasis and possibly other discourse constructs.
RDF databases are flexible because of their support for integration and inference rules without having to redefine the database structure. Consequently, attributes that occur in retrieval results cannot be determined earlier than at time of retrieval. It will be interesting to think about development of semantic structures that let domain discourse experts specify generation of perspectives of presentations by means of discourse constructs, in the absence of an exact knowledge of the attributes that occur in retrieval results.

9. SUMMARY AND CONCLUSION

This paper focuses on the automated derivation of two discourse constructs, being sequence and emphasis, from semantic annotations. The results of this work are a continuation of the Topia project, which generates discourse structures from clustering of semantic annotations. Other approaches focus on human-authored narrative templates for specifying sequence and emphasis. We present requirements for automated domain-independent generation of sequence and emphasis in the four phases of our processing chain, being analysis of semantic annotations, clustering, discourse structure generation and hypermedia generation. We also present an overview of the support that web standards, including the Semantic Web standard, offer for this. Principles for discourse generation that are independent of specific domain semantics allow automatic generation of narrative presentations from the contents of multiple repositories in web environments, irrespective of their application field.
Domain-independent criteria for sequence and emphasis follow from two sources of information. First, such criteria can be derived from attributes of information objects. Hard-coded sequences, numerical attributes and chains of information objects with identical relations between subsequent objects are sequence criteria that can be derived automatically. The occurrence of relatively large clusters of information objects that have identical attributes is a criterion for emphasis, as well as occasional attributes of objects with respect to those of other objects. A second criterion for sequence and emphasis is relevance of information objects for individual users. We present a relevance criterion that takes both types of criteria into account. The latter, subjective, criterion is according to a user-specified expression of relevance of information objects, stated by assigning relevance weights to attribute types that occur in the metadata repository.
This paper demonstrates application of the presented relevance criterion in the Topia architecture, in order to generate sequenced and emphasised clusters of objects in presentations of artefacts from the Rijksmuseum Amsterdam collection. RDF encoded annotations allow derivation of the actual set of attributes that occur with the retrieved objects at time of retrieval. Finally, we show that the user statement of relevance is a basis for generating presentations that put the retrieval result in specific perspectives.

ACKNOWLEDGMENTS

Funding for work on this paper came from the Topia project of the Telematica Instituut and CWI. Lynda Hardman and Frank Nack of CWI provided many helpful comments for improvement. Stanislav Pokraev of Telematica Instituut helped clarify the discussion of XML and RDF technology for this work. We thank the Rijksmuseum Amsterdam for their permission to use their Websites database and media content. We also thank IBM for sponsoring the project.

REFERENCES

1. Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., and Shadbolt, N.R. Automatic Ontology-based Knowledge Extraction from Web Documents, IEEE Intelligent Systems, 18(1) (January-February 2003), 14-21.
2. André, E., The Generation of Multimedia Documents, in. Dale, R, Moisl, H. and Somers, H. (eds.), A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text, Marcel Dekker Inc., 2000, 305-327.
3. Bal, M. Narratology: introduction to the theory of narrative, second edition. University of Toronto Press, 1997.
4. Bateman, J., Kamps, T., Kleinz, J. and Reichenberger., K. Towards constructive text, diagram and layout generation for information presentation, Computational Linguistics 27(3), 2001, 409-449.
5. Berkhin, P. Survey of clustering data mining techniques, http://www.accrue.com/products/rp_cluster_review.pdf
6. De Bra, P. Pros and Cons of Adaptive Hypermedia in Web-based Education. Journal on CyberPsychology and Behavior, Vol. 3, No. 1, Mary Ann Lievert Inc., 2000, 71-77.
7. Decker, S., Melnik, S., van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M. and Horrocks, I. The Semantic Web: The roles of XML and RDF, IEEE Internet Computing, 15(3), 2000, 63-74.
8. Buckingham Shum, S., Uren, V., Li, G., Domingue, J. and Motta, E. Visualizing Internetworked Argumentation, In: Kirschner, P.A., Buckingham Shum, S.J. and Carr, C.S. (eds), Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, Springer-Verlag: London, 2003, 185-204.
9. Ganter, B., and Wille, R., Applied Lattice Theory: Formal Concept Analysis. Preprints http://wwwbib.mathematik.tudarmstadt.de/Math-Net/Preprints/Listen/pp97.html, 1997.
10. Geurts, J., Bocconi, S., van Ossenbruggen, J., and Hardman, L. Towards Ontology-driven Discourse: From Semantic Graphs to Multimedia Presentations, technical report INS-R0305, http://ftp.cwi.nl/CWIreports/INS/INS-R0305.pdf, 2003.
11. Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K. Finding the flow in web site search. Communications of the ACM, Vol. 45, No. 9, 2002, 42-49.

12. Kamps, T., Diagram Design : A Constructive Theory, Springer Verlag, 1999.
13. Lassila, O. and Swick, R.R. (eds), Resource Description Framework (RDF) Model and Syntax Specification. World Wide Web Consortium (W3C) Recommendation, February 22nd, 1999.
14. Little, S., Geurts, J. and Hunter, J., Dynamic Generation of Intelligent Multimedia Presentations through Semantic Inferencing. In: Proceedings of the Sixth European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002), Springer, September 2002, 158-189.
15. Mann, W., Mattheissen, C., and Thompson, S. Rhetorical Structure Theory and Text Analysis. Information Sciences Institute Research Report, ISI/RR-89-242, 1989.
16. Van Ossenbruggen, J. and Hardman, L. Smart Style on the Semantic Web. In: Semantic Web Workshop, WWW2002, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-55/ossenbruggen.pdf, 2002.
17. Rijksmuseum Amsterdam, Rijksmuseum Amsterdam Website. http://www.rijksmuseum.nl
18. 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. ACM Hypertext, 2003. (to appear)
19. Rutledge, L., Davis, J., Van Ossenbruggen, J. and Hardman, L. Inter-dimensional Hypermedia Communicative Devices for Rhetorical Structure. In: Proceedings of the International Conference on Multimedia Modeling 2000 (MMM00), Nagano, Japan, November 13-15, 2000, World Scientific, 89-105.
20. Rutledge, L., van Ossenbruggen, J., Hardman, L. and Bulterman, D. Structural Distinctions Between Hypermedia Storage and Presentations. In: Proceedings of ACM Multimedia (pages 145-150), ACM Press, 1998.
21. Wong, S., Raghavan, V. Vector Space Model of Information Retrieval: A Reevaluation. In: Rijsbergen, C.J. van (Hrsg.), Research and Development in Information Retrieval, Cambridge University Press, Cambridge, UK, 1984, 167-186.