

# Conceptual Analysis of Hypertext

Robert E. Kent<sup>1\*</sup> and Christian Neuss<sup>2</sup>

<sup>1</sup> Washington State University, Pullman, WA 99164, USA

<sup>2</sup> Technische Hochschule Darmstadt, 64289 Darmstadt, Germany

**Abstract.** In this chapter tools and techniques from the mathematical theory of formal concept analysis are applied to hypertext systems in general, and the World Wide Web in particular. Various processes for the conceptual structuring of hypertext are discussed: summarization, conceptual scaling, and the creation of conceptual links. Well-known interchange formats for summarizing networked information resources as resource meta-information are reviewed, and two new interchange formats originating from formal concept analysis are advocated. Also reviewed is conceptual scaling, which provides a principled approach to the faceted analysis techniques in library science classification. The important notion of conceptual linkage is introduced as a generalization of a hyperlink. The automatic hyperization of the content of legacy data is described, and the composite conceptual structuring with hypertext linkage is defined. For the conceptual empowerment of the Web user, a new technique called conceptual browsing is advocated. Conceptual browsing, which browses over conceptual links, is dual mode (extensional versus intensional) and dual scope (global versus local).

## 1 Conceptual Knowledge Systems

Using ideas from library science [2,5], hypertext systems [7], and formal concept analysis [1,3], tools are currently being developed [12,15,16] for the conceptual analysis of networked information resources in general, and the World Wide Web in particular. Networked information resources include (1) individual text files, (2) WAIS databases, and (3) starting points for hypertext webs.

Resources are best thought of, not as objects, but as conceptual classes (formal concepts). We offer a concept-oriented approach for the description and organization of networked information resources, which will facilitate their subsequent discovery and access. This should not be thought of as yet another object-oriented approach. Although objects generate their own classes, classes are not only more general but also include intensional information. By identifying concepts with classes, this can be regarded as a class-oriented approach — an approach that has been advocated recently by Terry Winograd in the IETF-URI working group discussion on library standards and URI, and supported by Ronald Daniel and Dirk Herr-Hoyman.

---

\* This research was funded by a grant from Intel Corporation.

Formal concept analysis [1,3] is a relatively new discipline arising out of the mathematical theory of lattices and the calculus of binary relations. It is closely related to the areas of knowledge representation in computer science and cognitive psychology. Formal concept analysis provides for the automatic classification of both knowledge and documents via representation of a user's faculty for interpretation as encoded in conceptual scales. Such conceptual scales correspond to the facets of synthetic classification schemes, such as Ranganathan's Colon classification scheme, in library science.

Formal concept analysis uses objects, attributes and conceptual classes as its basic constituents. Objects and attributes are connected through has-a incidence relationships, while conceptual classes are connected through is-a subtype relationships. Incidence is the most primitive notion in formal concept analysis. A *formal context* represents incidence by collecting together all of the relevant has-a relationships. It is a triple  $\langle G, M, I \rangle$  consisting of a set of objects  $G$  (Gegenstände, in German), a set of attributes  $M$  (Merkmale, in German), and a binary incidence relation  $I \subseteq G \times M$ , where  $gIm$  asserts that "object  $g$  has attribute  $m$ ." In many contexts appropriate for Web resources, the objects are document-like objects and the attributes are properties of those document-like objects which are of interest to the Web user.

A *conceptual class* or *formal concept* is the central notion in formal concept analysis. A formal concept consists of a collection of entities or objects exhibiting one or more common characteristics, traits or attributes. Conceptual classes are logically characterized by their extension and intension. The *extension* of a class is the aggregate of entities or objects which it includes or denotes. The *intension* of a class is the sum of its unique characteristics, traits or attributes, which, taken together, imply the concept signified by the conceptual class. In this paper conceptual classes are identified with the concept which they signify. The process of subordination of concepts and collocation of objects exhibits a natural order, proceeding top-down from the more generalized concepts with larger extension and smaller intension to the more specialized concepts with smaller extension and larger intension. This is-a relationship is a partial order called generalization-specialization. Concepts with this generalization-specialization ordering form a class hierarchy  $\mathcal{L} = \mathcal{L}\langle G, M, I \rangle$  called a *concept lattice*. Formal concept analysis uses formal concepts as its central notion and uses concept lattices as an approach to knowledge representation [1]. The use of conceptual classes as a conceptual structuring mechanism corresponds to the use of similarity clusters in information retrieval [7], although conceptual classes are based more on logical implication rather than a nearness notion. However, see the discussion about conceptual linkage below.

The enriching notion of a *conceptual knowledge system* from formal concept analysis [4,16] allows, not only the modeling of knowledge representation, but also the ability to do knowledge inferencing, knowledge acquisition, and knowledge communication. In a conceptual knowledge system there are three basic notions: objects, attributes, and conceptual views. These are connected through four basic relationships: an object has an attribute (incidence), an object belongs

to a conceptual view (instantiation), an attribute abstracts from a conceptual view (abstraction), and a conceptual view is a subordinate to another conceptual view (subtype). These notions and relationships partition the frame of a conceptual knowledge system as in Table 1. In a conceptual knowledge system we distinguish between (1) anonymous concepts which are automatically and implicitly generated from the four basic relationships and represent a form of conceptual resource discovery, and (2) named and explicitly specified concepts which we call conceptual views. Compare the distinct, but closely related, notion of a Nebula-style view [16].

**Table 1.** Conceptual Knowledge System Relationships

	Views	Attributes
Views	subtype	abstraction
Objects	instantiation	incidence

Table 2 represents a conceptual knowledge system within the conceptual universe  $\mathcal{D}$  of all documents in an information system and their properties (see Figure 2 in [9]). In addition to a set of document-like objects and attributes, it contains the five conceptual views {Object, Document, PostScript, Plan1, Plan2}. The crosses in the table of basic relationships in Table 2 are partitioned into the four parts described in Table 1: subtype, in the upper left; abstraction, in the upper right; instantiation, in the lower left; and incidence, in the lower right. The bottom panel of Table 2 is the line diagram of the lattice of conceptual classes, which represents the conceptual space for the document conceptual knowledge system.

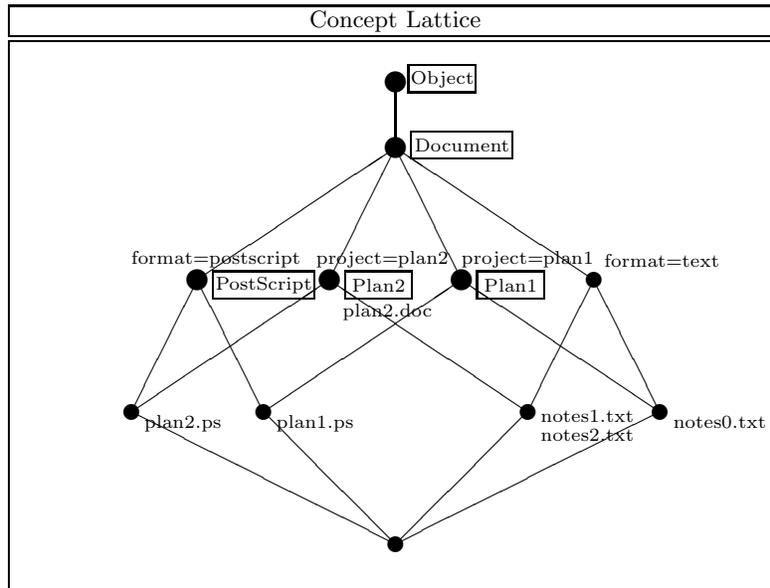
The representational mechanism of conceptual knowledge systems serves as a firm foundation for the basic paradigms of internet resource discovery and wide area information management systems: organization-navigation and search-retrieval [16]. The use of conceptual knowledge systems is a natural outgrowth of the original formal concept analysis approach for structuring and organizing the networked information resources in the World Wide Web [12].

## 2 Resource Meta-information

Due to the rapid growth of the World Wide Web, resource discovery has become a serious problem. Because of its decentralized architecture, the user experiences the Web as a large information repository without an underlying structure. The process of “surfing” pages by repeatedly following hyperlinks is the most popular use of the Web. It can however lead to the phenomenon of getting “lost in hyperspace.”

**Table 2.** Conceptual Knowledge System in the Document Universe

Views/Objects	Basic Relationships	Views/Attributes
1 Object		1 Object
2 Document	1 ×	2 Document
3 PostScript	2 × ×	3 PostScript
4 Plan1	3 × × ×	4 Plan1
5 Plan2	4 × × × ×	5 Plan2
6 plan1.ps	5 × × × × ×	6 project=plan1
7 plan2.ps	6 × × × × × ×	7 project=plan2
8 plan2.doc	7 × × × × × × ×	8 format=postscript
9 notes0.txt	8 × × × × × × ×	9 format=text
10 notes1.txt	9 × × × × × × ×	
11 notes2.txt	10 × × × × × × ×	
	11 × × × × × × ×	



From the very beginning, approaches have been made to organize information about networked information resources into catalogs and indexes. Index files were originally maintained manually. However, the rapid growth of the Web soon made necessary automatic methods for generating resource directories. Automatic tools called “robots”, “Web wanderers” or “spiders” soon evolved. These are programs which automatically connect to a remote server and recursively retrieve documents. Since Web robots often put heavy loads on Web servers, they have been controversial, and are sometimes disliked by server maintainers.

Web robots are trailing-edge technologies. The main problem with robots is that they are not true Web wanderers — the retrieval program does not transfer itself from the index site to the provider site, but instead it transfers in the reverse direction over the network all the potentially indexable documents. Since document repositories may contain hundreds of megabytes, the bandwidth requirements are enormous. Exacerbating this problem is the fact that current indexing tools gather independently, without sharing information with other indexers.

A partial answer to these problems are Networked Information Discovery and Retrieval (NIDR) systems such as Harvest [13]. A more complete answer will involve NIDR systems with conceptual structuring mechanisms[12] such as the WAVE<sup>3</sup> system (Web Analysis and Visualization Environment) which is being developed by following principles espoused in this chapter. NIDR systems are leading-edge technologies which reduce the load on information servers, reduce network traffic, and reduce index disk space requirements, principally by use of resource meta-information (also called metadata) — they extract meta-information at the provider site, sending this, and not the raw data, over the network. This section reviews various formats used by NIDR systems and library science for representing resource meta-information as bibliographic records [15].

**Uniform Resource Characteristics:** The on-going discussions concerning metadata in various internet engineering task force (IETF) working groups are centered around the following notions [11]. A Uniform Resource Locator (URL) is used for hyperlink markup in Web documents. Since a URL specifies a location and retrieval protocol of a given networked information resource, it is not a long-lived, stable reference. A Uniform Resource Name (URN) is used to identify a resource. It is long-lived and persistent, and uniquely names a networked information resource. A Uniform Resource Locator is used to locate an instance of a resource identified by an URN. A Uniform Resource Characteristic (URC) is used to represent URNs with associated meta-information. URCs are analogous to the bibliographic records of library science. URCs encode meta-information about network resources in the form of attribute-value pairs.

**IAFA Templates:** The internet anonymous ftp archives (IAFA) working group of the IETF has proposed a format for indexing information that can be used

---

<sup>3</sup> The first author is the principal investigator for an Intel funded project which is developing and assessing the WAVE system.

to describe various internet resources. The IAFA template specification [8] encodes pieces of meta-information. The IAFA templates are intended to be both human and machine readable. Archie servers support this format to provide information about items available for anonymous ftp. Work is currently underway for the construction of Uniform Resource Identifiers.

**Harvest Summary Object Interchange Format:** Harvest is a set of tools to gather, extract, and search relevant information across the internet [13]. It provides methods for distributed indexing, building topic specific indices, flexible search strategies, and replica systems. Harvest generates a content summary for each information object it gathers. These records are stored in a format called the Summary Object Interchange Format (SOIF). SOIF is based on a combination of the IAFA templates and `BIBTEX`.

**Bibliographic Records from Library Science:** In order to compare URCs, IAFA templates, and Harvest SOIFs with bibliographic description in library science, listed here are some attributes, which are classified according to the eight areas of the international standards for bibliographic description (ISBD) [5]: title and statement of responsibility (title, author); edition (version); material (or type of publication) specific details; publication, distribution, etc.; physical description (content-type, content-length, size, cost, etc.); series (time-to-live); notes (abstract); and standard number and terms of availability (uniform resource name, uniform resource locator).

Table 3 lists two generic interchange formats which can be used to specify faceted information in conceptually scaled networked information resources [15]. Such faceted information can occur in various interfaces in a resource discovery system. From a mathematical viewpoint, these two representations are equivalent to each other. Software exists for converting between the two forms.

The left side of Table 3 displays the Formal Context Interchange Format (FCIF). FCIF is oriented towards the formal contexts of formal concept analysis. FCIF represents order-theoretic formal contexts of networked information resources, consisting of two partially ordered sets, a poset of objects and a poset of single-valued attributes, and an order-preserving incidence matrix which represents the relationship between objects and attributes. The right side of Table 3 displays the Concept Lattice Interchange Format (CLIF). CLIF is oriented towards the concept lattices of formal concept analysis. CLIF provides a storage-optimal representation of order-theoretic lattices of conceptual classes for networked information resource meta-information, consisting of (the inverse relationships for) two generator monotonic functions, from the posets of objects and attributes to the lattice of conceptual classes, and a successor matrix which represents the subtype relationship between conceptual classes.

FCIF and CLIF subsume both the URCs of the IETF and the SOIFs of Harvest. The FCIF and CLIF interchange formats are more general mechanisms than either URCs or SOIFs, and allow for the specification of more complex conceptually structured systems of resources. Actually, as Figure 3 points out, both FCIF and CLIF are better thought to occur after conceptual scaling, whereas

**Table 3.** Interchange Formats for Faceted Resource Meta-information

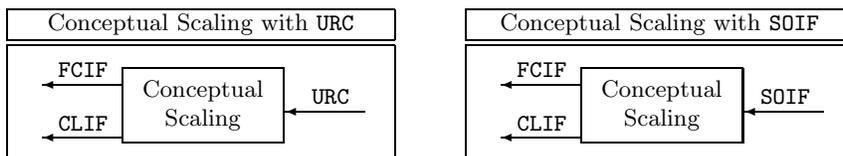
Formal Context Interchange Format	Concept Lattice Interchange Format
<p><b>TYPE</b>  <math>T</math></p> <p><b>OBJECT</b>  <math>O_1 \{ O_{1,1} O_{1,2} \cdots O_{1,o_1} \}</math>  <math>O_2 \{ O_{2,1} O_{2,2} \cdots O_{2,o_2} \}</math>  <math>\dots</math>  <math>O_n \{ O_{n,1} O_{n,2} \cdots O_{n,o_n} \}</math></p> <p><b>ATTRIBUTE</b>  <math>A_1 \{ A_{1,1} A_{1,2} \cdots A_{1,a_1} \}</math>  <math>A_2 \{ A_{2,1} A_{2,2} \cdots A_{2,a_2} \}</math>  <math>\dots</math>  <math>A_m \{ A_{m,1} A_{m,2} \cdots A_{m,a_m} \}</math></p> <p><b>INCIDENCE</b>  <math>O_1 \{ A_{1,1} A_{1,2} \cdots A_{1,i_1} \}</math>  <math>O_2 \{ A_{2,1} A_{2,2} \cdots A_{2,i_2} \}</math>  <math>\dots</math>  <math>O_n \{ A_{n,1} A_{n,2} \cdots A_{n,i_n} \}</math></p>	<p><b>TYPE</b>  <math>T</math></p> <p><b>GENERATOR: OBJECT</b>  <math>C_1 \{ O_{1,1} O_{1,2} \cdots O_{1,o_1} \}</math>  <math>C_2 \{ O_{2,1} O_{2,2} \cdots O_{2,o_2} \}</math>  <math>\dots</math>  <math>C_p \{ O_{p,1} O_{p,2} \cdots O_{p,o_p} \}</math></p> <p><b>GENERATOR: ATTRIBUTE</b>  <math>C_1 \{ A_{1,1} A_{1,2} \cdots A_{1,a_1} \}</math>  <math>C_2 \{ A_{2,1} A_{2,2} \cdots A_{2,a_2} \}</math>  <math>\dots</math>  <math>C_p \{ A_{p,1} A_{p,2} \cdots A_{p,a_p} \}</math></p> <p><b>SUCCESSOR</b>  <math>C_1 \{ C_{1,1} C_{1,2} \cdots C_{1,s_1} \}</math>  <math>C_2 \{ C_{2,1} C_{2,2} \cdots C_{2,s_2} \}</math>  <math>\dots</math>  <math>C_p \{ C_{p,1} C_{p,2} \cdots C_{p,s_p} \}</math></p>

- $O_i$  and  $O_{i,o}$  are object names (strings).
- $A_i$  and  $A_{j,a}$  are attributes *tag#value*, where # is =, ≤, etc.
- $C_k$  and  $C_{k,s}$  are indexes (natural numbers) of conceptual classes.
- $x_i$  and  $y_j$  are coordinates (natural numbers) of conceptual class nodes.

both URC and SOIF specify “raw meta-information” which exists before conceptual scaling [3].

From the philosophical viewpoint of formal concept analysis, conceptual scaling is an act of interpretation. It maps raw uninterpreted data, such as occurs in URC or SOIF, into the end-user’s conceptual scheme. URC and SOIF represent database entity relations, whereas FCIF represents has-a incidence relationships between objects and attributes and CLIF represents is-a subtype relationships between conceptual classes. These attributes are simple structured queries of the form tag#value, where # is any relational operator =, ≤, etc. The equality operator represents nominal scaling, whereas the inequality operator represents ordinal scaling [3]. Through conceptual scaling, which often is just nominal or ordinal scaling, we can compare FCIF and CLIF with URC and SOIF.

Fig. 1. Conceptual Scaling with Various Interchange Formats



### 3 Conceptual Linkage

The structuring primitive for the World Wide Web is the hyperlink. The essence of a hyperlink is a (possibly typed) binary association between two objects [6]. The semantics of a hyperlink is that the two connected objects have something in common — a property or a semantic category [7]. By extending ideas from the field of formal concept analysis [1], in this chapter we offer a principled approach for elevating the notion of a hyperlink from objects to conceptual classes (concepts). These new extended linkage structures, which preserve the idea that things are linked through shared attributes, are called *conceptual links*. The notion of conceptual links derived here can be compared to similar notions in hypertext systems [7], which are intuitive but not principled. The crisp notion of a conceptual link is represented here by the richer, graded notion of *conceptual linkage*. Conceptual linkage is a fuzzy relationship. It gives a measure of similarity and implication between concepts. Conceptual linkage can be reduced to conceptual links by a crispification operation. The crucial idea of conceptual linkage is derived from an extended theory of conceptual knowledge systems.

Conceptual linkage can be used as the structuring primitive for the conceptual organization of the knowledge implicit in the World Wide Web. There are important parallels between conceptual knowledge systems and hypertext systems. In particular, conceptual links are analogous to Web hyperlinks. Actually,

this is more than an analogy, since objects (or their abstracted synoptic surrogates in the form of metadata objects) generate conceptual classes. This impels us to make the following observations.

- *Networked resources are concepts (conceptual classes).*
- *Conceptual linkage extends and enriches Web hyperlinks.*
- *Conceptual space customizes and makes coherent Web hyperspace.*

There are two modes for conceptual linkage: extensional and intensional. Since these are dual notions in lattices, we only discuss the extensional mode here. In the *extensional mode* of conceptual linkage, concepts are regarded as attributes and are represented by their extent. Any two concepts in extensional mode are linked by the objects which they share, the objects common to their extents. The more linking objects there are, the closer are those concepts and the stronger is the conceptual linkage. This closeness can be measured by the cardinality of the set of linking objects<sup>4</sup>.

The *extensional similarity* measure  $\sigma_{\bullet} : \mathcal{L} \times \mathcal{L} \rightarrow \aleph = \{0, 1, \dots\}$  is a measure of the similarity of any two concepts according to their common extent cardinality. It is the composite  $(\text{meet}) \circ (\text{extent}) \circ (\text{cardinality})$ , and is defined by the formulae

$$\sigma_{\bullet}(k_0, k_1) = \|\text{extent}(k_0 \wedge k_1)\| = \|\text{extent}(k_0) \cap \text{extent}(k_1)\|$$

for any two concepts  $k_0, k_1$  in a concept lattice  $\mathcal{L}$ . The bounds on this measure are  $0 \leq \sigma_{\bullet}(k_0, k_1) \leq \min\{\|\text{extent}(k_0)\|, \|\text{extent}(k_1)\|\}$ . The extensional similarity between two concepts is a rough (fuzzy?) measure of their similarity. The closer the concepts, the larger the extensional similarity, up to a maximum size of the extent cardinality of either. When this upper bound is reached, one concept is below the other in the concept lattice

$$k_0 \leq k_1 \quad \text{iff} \quad \|\text{extent}(k_0)\| = \sigma_{\bullet}(k_0, k_1).$$

The more dissimilar the concepts, the smaller the extensional similarity, with lower bound 0. When this bound is reached, the two concepts have nothing extensionally in common. For browsing over the conceptual space of a conceptual knowledge system, we take a state space approach where we regard concepts as conceptual states and browsing as state transition. Since extensional similarity is symmetric  $\sigma_{\bullet}(k_0, k_1) = \sigma_{\bullet}(k_1, k_0)$ , it does not accurately represent the notion of conceptual state and conceptual state transition, because it ignores the

---

<sup>4</sup> For any cardinality, we count only a kind of atomic concept called an irreducible concept: join irreducible for object concepts and meet irreducible for attribute concepts. In a concept lattice, an object (that is, an object concept) is join irreducible when it cannot be decomposed as the join of two other objects, and an attribute (attribute concept) is meet irreducible when it cannot be decomposed as the meet of two other attributes. For atomicity to be realizable, we assume that formal concept analysis optimization processes of purification and reduction have been carried out. Purification fuses objects which generate the same concept. With respect to this conceptual knowledge system, these objects are indiscernible and equivalent. Purification does the same for attributes. Reduction converts objects and attributes which are not irreducible into conceptual views.

asymmetric nature of the current state: we are at state  $k_0$ , we are not at state  $k_1$  (although we may want to transition there). The notion of “current state” is well represented by extensional linkage.

*Extensional linkage*  $\lambda_{\bullet} : \mathcal{L} \times \mathcal{L} \rightarrow [0, 1]$ , which ranges between 0 and 1, is a fuzzy measure of the implication between concepts. This asymmetric measure of linkage or implication, which is defined as the ratio of the sizes of extents

$$\lambda_{\bullet}(k_0, k_1) = \frac{\|\text{extent}(k_0 \wedge k_1)\|}{\|\text{extent}(k_0)\|} = \frac{\|\text{extent}(k_0) \cap \text{extent}(k_1)\|}{\|\text{extent}(k_0)\|} = \frac{\sigma_{\bullet}(k_0, k_1)}{\|\text{extent}(k_0)\|},$$

measures the implication “ $k_0$  implies  $k_1$ ”. Extensional linkage can be informally interpreted as a measure of relevance:  $\lambda_{\bullet}(k_0, -)$  measures the strength of connection, transitional strength, or relevance, from conceptual state  $k_0$  to other conceptual states. Extensional linkage can be formally interpreted as the probability of  $k_1$  conditioned on  $k_0$ ; that is, the conditional probability  $p(k_1|k_0)$ . The maximum measure of linkage or implication represents a strict, full, or Boolean measure of linkage or implication “ $k_0$  strictly implies  $k_1$ ”. This occurs at the concept lattice order

$$k_0 \leq k_1 \quad \text{iff} \quad \lambda_{\bullet}(k_0, k_1) = 1.$$

So, conceptual linkage subsumes the hierarchical linkage of the lattice order of concepts. Extensional linkage  $\lambda_{\bullet}$  can be represented by a square matrix of real numbers in the interval  $[0, 1]$ , whose dimension is the cardinality of the set of conceptual classes in the lattice of the conceptual knowledge system.

## 4 Conceptual Neighborhood

The lattice of concepts in a conceptual knowledge system is intuitively regarded as an environment or conceptual space. There are two dual senses or modes for the idea of a “local neighborhood” of a concept within its conceptual space. These two senses of neighborhood are closely bound up with the two modes of conceptual linkage. The *extensional neighborhood*  $\mathcal{N}_{\bullet}(k)$  of a “seed” concept  $k$  regards the concept as an attribute: it fuses the intent of the concept as a collective attribute and distributes the extent downward over a local neighborhood lattice. Precisely defined, the conceptual knowledge system of the extensional neighborhood is the restriction of the global conceptual knowledge system to the extent of the concept — all objects not in the extent are ignored. In terms of conceptual structure, for any conceptual state  $k$  the local extensional neighborhood concept lattice  $\mathcal{N}_{\bullet}(k)$  is the restriction of the global lattice  $\mathcal{L}$  by means of the *meet restriction* operation  $k \wedge (\cdot) : \mathcal{L} \rightarrow \mathcal{N}_{\bullet}(k)$  is right adjoint right inverse to a monotonic map  $\mathcal{N}_{\bullet}(k) \rightarrow \mathcal{L}$  which embeds the extensional neighborhood lattice into the global lattice. This means that meet restriction is meet-preserving since it is right adjoint, and surjective since it is right inverse.

The size of the extensional neighborhood depends upon the genericity of the seed concept. The extensional neighborhood of the top concept is very large,

the entire global conceptual knowledge system. The extensional neighborhood of the bottom concept is very small, having only one concept. Since the extent is usually much smaller than the entire set of objects of the global conceptual knowledge system, the concept neighborhood notion gives a drastic reduction in the size of the conceptual space. The collection of all attributes which label the “root” node (top concept) is the intent of the seed concept. At the opposite pole, any attribute which labels the bottom node is extensionally disjoint from the seed concept in the global lattice (except for any “solution objects” — objects which satisfy all properties). We can loosely regard the extensional neighborhood lattice line diagram to be a hierarchy labeled by the extent of  $k$ . These extensional objects are distributed over this local neighborhood lattice by means “distinguishing attributes”. By definition, these attributes are not in the intent of  $k$ . This observation forms the basis for local browsing in the extensional mode via intensional difference.

Between any two concepts  $k_0$  and  $k_1$  in a concept lattice  $\mathcal{L}$  is the *intensional difference*  $\partial^\bullet(k_0, k_1) = \text{intent}(k_1) \setminus \text{intent}(k_0)$ , an asymmetric measure which records those attributes of  $k_1$  that are not attributes of  $k_0$ . Elements in  $\partial(k_0, k_1)$  are attributes which “distinguish”  $k_1$  from  $k_0$ . The intensional difference  $\partial^\bullet: \mathcal{L} \times \mathcal{L}^{\text{op}} \rightarrow \wp M = \langle \wp M, \supseteq, \cup, \emptyset \rangle$  is a generalized metric or distance function, which satisfies the zero law  $\emptyset \supseteq \partial^\bullet(k, k)$  and the triangle law  $\partial^\bullet(k_0, k_1) \cup \partial^\bullet(k_1, k_2) \supseteq \partial^\bullet(k_0, k_2)$ . All lattice order information is contained in the intensional difference, since

$$k_1 \leq k_0 \quad \text{iff} \quad \partial^\bullet(k_0, k_1) = \emptyset.$$

The intensional difference is the basis for the idea of a dictionary definition. A word (thought of as an object concept) is defined by restricting or specializing a superordinate (more generic) concept by means of a collection of distinguishing properties: a concept  $k_1$  is-a concept  $k_0$  which satisfies all attributes  $m$  in the intensional difference  $\partial^\bullet(k_0, k_1)$ . For example, “a tree is a plant which is woody, perennial and has a main stem.” Here “tree” is the concept being defined (definiendum), “plant” is the superordinate concept, and “woody”, “perennial”, and “main stem” are in the intensional difference. In the same fashion, in a conceptual lattice, we can then think of the collection of differentiating attributes as representing the difference between a defined concept and the superordinate concept.

The *intensional difference* measure  $\delta^\bullet: \mathcal{L} \times \mathcal{L}^{\text{op}} \rightarrow \aleph = \langle \aleph, \geq, +, 0 \rangle$  is also a generalized metric, which satisfies the zero law  $0 \geq \delta^\bullet(k, k)$  and the triangle law  $\delta^\bullet(k_0, k_1) + \delta^\bullet(k_1, k_2) \geq \delta^\bullet(k_0, k_2)$ . It is a measure of the difference between any two concepts according to their intensional difference cardinality. The intensional difference measure is defined by the formulae

$$\delta^\bullet(k_0, k_1) = \|\partial^\bullet(k_0, k_1)\| = \|\text{intent}(k_1) \setminus \text{intent}(k_0)\|$$

for any two concepts  $k_0, k_1$  in a concept lattice  $\mathcal{L}$ . Again, all lattice order information is contained in the intensional difference measure, since

$$k_1 \leq k_0 \quad \text{iff} \quad \delta^\bullet(k_0, k_1) = 0.$$

The minimum measure 0 occurs when concept  $k_1$  is at or below concept  $k_0$  in the main lattice. This occurs when no attribute distinguishes concept  $k_1$  from concept  $k_0$ , although there might be an attribute which distinguishes concept  $k_0$  from concept  $k_1$ . The intensional difference measure counts the number of distinct distinguishing attributes. It measures how distinguished  $k_1$  is from  $k_0$ .

A *ranked order*  $\langle \mathcal{X}, \rho \rangle$  consists of a partially ordered set  $\mathcal{X} = \langle X, \leq \rangle$  and an monotonic map  $\rho: \mathcal{X} \rightarrow \mathbb{N} = \{0, 1, \dots\}$  to the natural numbers called a *ranking*. Ranked orders can be displayed by inverse image  $\rho^{-1}(n) = \{x \in X \mid \rho(x) = n\}$ , either directly  $(\rho^{-1}(0), \rho^{-1}(1), \dots, \rho^{-1}(\max))$  or in reverse order  $(\rho^{-1}(\max), \rho^{-1}(\max-1), \dots, \rho^{-1}(0))$ . Ranked orders are used here as reduced representations for concept lattices. They are most useful for browsing via the local conceptual neighborhoods, in either the extensional mode where we browse over the views and attributes of the global lattice, or the intensional mode where we browse over the views and objects. Table 4 displays the extensional mode rankings for the conceptual view “Plan1”. The upper panel displays the extensional similarity ranking at conceptual state “Plan1”, a reduced representation for the global document conceptual space displayed in Table 2. Here concepts “Document” and “Object” have merged in the ranking with concept “Plan1”, whereas the opposite ranking pole shows that concept “Plan2” is extensionally disjoint from concept “Plan1”. This ranking displays all of the irreducible conceptual views and attribute concepts in the document universe  $\mathcal{D}$ . The lower panel displays the intensional difference ranking of concept “Plan1”, a reduced representation for the local document neighborhood of “Plan1”. This ranking displays only the extent of concept “Plan1”.

**Table 4.** Extensional Mode Rankings for the Conceptual View “Plan1”

<b>Global Scope</b>	{	Extensional Similarity Ranking $\sigma_{\bullet}(\text{Plan1}, -)$
		3    {[Object], [Document], [Plan1, project=plan1]}
		2    {}
		1    {[PostScript, format=postscript], [format=text]}
		0    {[Plan2, project=plan2]}
<b>Local Scope</b>	{	Intensional Difference Ranking $\delta^{\bullet}(\text{Plan1}, -)$
		0    {[Plan1]}
		1    {[plan1.ps], [notes0.txt]}

*Conceptual browsing* is browsing over conceptual linkage. It is dual mode (extensional versus the intensional) and dual scope (global versus local). Extensional and intensional mode are temporally disjoint, whereas global scope is antecedent to local scope both logically and temporally: choose a mode; first browse globally in that mode and then browse locally in the same mode. Theoretically, conceptual browsing ranges over all concepts, with concepts being represented by internal indexes. Practically, conceptual browsing ranges only over named concepts: objects, attributes, and conceptual views. In extensional mode we browse over concepts by restriction to their extents. In intensional mode we do just the lattice dual — we browse over concepts by restriction to their intents. Browsing in the global scope means browsing over the global concept lattice, whereas browsing in a local scope means browsing over a local neighborhood concept lattice. Conceptual browsing is summarized in Table 5.

If a concept lattice is regarded as a form of database structure, then conceptual browsing can be used for database access, as in information retrieval [7]. In this approach conceptual linkage is used for processing queries. A query in intensional mode involves only the attributes of the formal context under consideration. By definition, an *intensional query* is a subset of attributes. It can be identified with a new temporary “goal query” object which has been added to the formal context. The goal query object is regarded to be the current conceptual state for browsing in intensional mode. Then, intensional linkage ranking is a vector of similarity coefficients, each coefficient measuring the closeness of a concept to the goal query. Either conceptual views and objects can be display as a ranking, or those conceptual views and objects can be returned whose similarity coefficient is above a given threshold. By duality, an *extensional query* is a subset of objects. The query is identified with a new temporary “goal query” attribute, which is regarded to be the current conceptual state for browsing in extensional mode. The objects in the query are regarded as prototypes. Extensional queries correspond to a prototype representation for categories (conceptual classes). Issuing an extensional query results in returning similarity measures between conceptual classes and the collective prototype of the query’s objects.

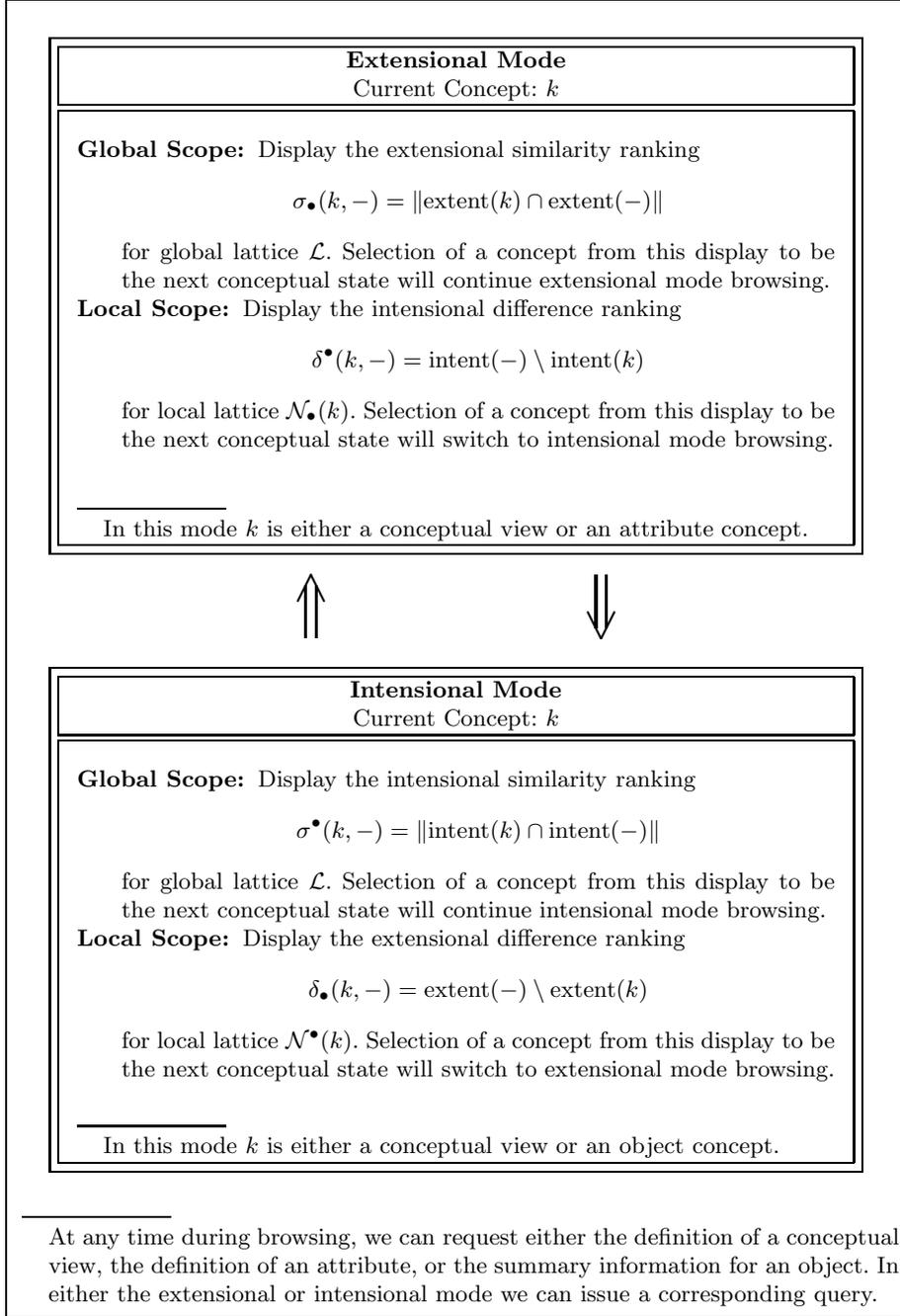
## 5 Conceptualization Processes

The intuitive idea behind hypertext is “semantic connection” [6]. Currently, hyperlink creation is done manually at document creation time [7]. There are two problems with this manual approach:

- The document creator (writer, publisher) may inadvertently omit some important and meaningful semantic connections.
- Legacy data (pre-HTML documents) needs enormous manual effort in order to convert to hypertextual form.

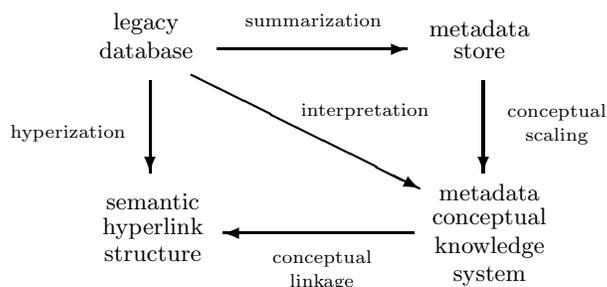
Figure 2 gives a high-level description of processes involved in the conceptual organization and representation of the information in legacy databases. The interpretation process [14] is a composite of summarization followed by conceptual

**Table 5.** The Process of Conceptual Neighborhood Browsing



scaling [3,12],  $(\text{interpretation}) = (\text{summarization}) \circ (\text{conceptual scaling})$ . Summarization is the abstraction and construction of metadata objects from actual data. The gathering component of the Harvest system [13] is a good example of summarization. Conceptual scaling, also called relational data filtration, is a user-oriented process for customizing and building a faceted representation of information based upon user interest profiles, etc. Here a user may refer to either a single individual, a small group of individuals, or even a whole community. Conceptual scaling uses type-structured standing queries, known as conceptual scales [3], alerts, continuous queries, or SDI (selective dissemination of information) [2].

**Fig. 2.** Conceptual Interpretation and Database Hyperization

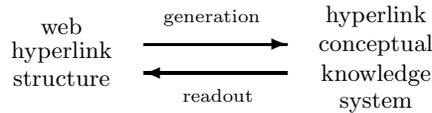


The hyperization process is a process of automatic web archiving. As such, it answers the concerns expressed above about manual web creation. Actually, hyperization could represent either the batch process of web archiving or the interactive process of web guidance during client browsing. Hyperization is a composite of interpretation followed by conceptual linkage,  $(\text{hyperization}) = (\text{interpretation}) \circ (\text{conceptual linkage})$ . As depicted in Figure 2, conceptual linkage involves the automatic creation of *crisp* web hyperlink structure by a reduction process of crispification. There is, however, information loss in just the creation of crisp web hyperlinks. In this sense, it is better to remain at the higher level of the conceptual knowledge system, rather than reducing to web hyperlink structure. At the higher level of the conceptual knowledge system, conceptual linkage richly expresses conceptual structure and semantic content.

Figure 3 describes the equivalence between the hyperlink structure of the Web and its representation as a conceptual knowledge system. In the application of the conceptual knowledge system model to Web hyperlinkage, both objects and attributes are Web objects (HTML documents, images, etc.). There are two dual interpretations for hyperlink incidence: (cross-referential) one Web object has a second Web object as an attribute when the first points to the second; and (hierarchical, such as gopher-space) the opposite incidence [6,7]. The

web-cks equivalence in Figure 3 is mediated through the inverse passages of concept generation and incidence readout:  $(\text{generation}) \circ (\text{readout}) \equiv (\text{identity})$  and  $(\text{readout}) \circ (\text{generation}) \equiv (\text{identity})$ . These inverse passages comprise the standard process diagram from formal concept analysis, here applied to the incidence relationships of Web hyperlinks. The process of concept generation results in a conceptual representation for Web hyperlinkage.

**Fig. 3.** Conceptual Structuring of Hypertext Incidence

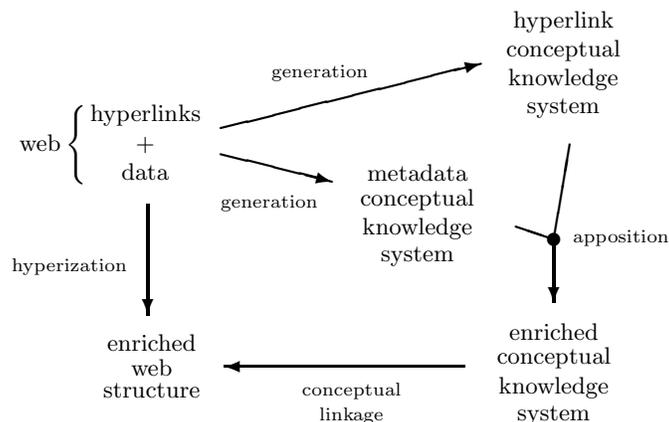


The information in the Web can be split into two distinct aspects: hyperlinkage and document content [6,7]. Distinct processes applied to these two aspects generate distinct conceptual representations. The semantic constraint between hyperlinkage and document content can be applied later during the process of conceptual scaling and data filtration. Substitution of the non-linked content of a web for the legacy database component in Figure 2 describes a process for the conceptual representation of Web content. In Figure 4 we describe an enriched process which combines the hyperlink conceptual representation of Figure 3 with the metadata conceptual representation of Figure 2. This enriched combining process uses a standard combinator from formal concept analysis called apposition. At the incidence matrix level of a conceptual knowledge system, apposition is a summing process, whereas at the concept lattice level it is a constrained producing process. The same comments that we made above about crisp conceptual linking are true here also: it is better to do conceptual linkage at the enriched conceptual knowledge system level — here there is no loss of information and a richer conceptual expression.

## 6 Summary, Implementation, and Future Work

This chapter has discussed two approaches for making use of automatic classification techniques: web archive construction and user navigational guidance (conceptual browsing). Automatic classification provides the foundation for the automatic generation of local web hypertextual structure based upon summarized and conceptually scaled information about objects. Manual specification of conceptual connectivity, such as for ordinary hyperlinks and conceptual views, can automatically be incorporated. Automatic classification also provides the foundation for guidance and analysis during user browsing and concept link traversal via the World Wide Web over any community's information space. In summary, formal concept analysis is a principled foundation for classification,

**Fig. 4.** Enriched Web Construction



organization, and indexing in NIDR systems. By using ideas from formal concept analysis, Web hyperlinks can be elaborated into Web conceptual links, and Web hyperspace can be coherently organized as Web conceptual space.

By the time of publication, many of the ideas discussed in this chapter will have been implemented in the NIDR system called WAVE which was mentioned above. These ideas include formal contexts, concept lattices, conceptual knowledge systems, conceptual optimization, conceptual linkage, and conceptual browsing. Development of the WAVE system will provide a preliminary answer to the research question: “What is the appropriate architecture for a digital library?” It will be demonstrated in the distributed context of the World Wide Web, by using both the technique of automatic classification and the notion of a conceptual knowledge system, that the WAVE system provides the kernel architecture for a digital library. A critical measure of success for the WAVE system will be the ability to *understand* a user’s intentions. The understanding of intentions is a very deep research question, and as Dennis Reinhardt has pointed out to the first author (private communication), machine understanding will not surpass human capability in this area during the course of this research. However, the WAVE system could augment human understanding with the ability to *express* a user’s intentions. This sense of “understanding” a user’s intentions will be a critical factor in the success of the WAVE system. Other measures, such as how customizable, how adaptable, or how flexible the system is for the user, are subordinate strategies which will aid the ability to express the user’s intentions.

Both on-going and future work can be discussed in terms of three processes for the conceptualization of networked information resources, as diagrammed in Figure 2: summarization, conceptual scaling, and conceptual linkage. The first process, summarization, has been implemented as the front component of a

NIDR system, where meta-information is extracted. An important example of a summarization processor is the gatherer component of the Harvest system. The third process, conceptual linkage, is now being implemented as the first phase (funded) of the WAVE system development. This phase, called *WAVEGuide*, will replace the broker indexing component of the Harvest system, extending broker capabilities by adding dynamic and customizable knowledge organization techniques. *WAVEGuide* will be used for interactive information analysis and browsing guidance during exploratory search by client Web browsers over a community's information space. The second process, conceptual scaling, will next year be implemented as the second phase of WAVE system development. This phase, called *WAVEForm*, will represent the process of faceted analysis which occurs in library science classification. It also corresponds to the design of user interest profiles in current awareness services [2].

## References

1. Wille, R.: Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. Ordered Sets, I. Rival (ed.), Reidel, Dordrecht-Boston (1982) 445–470
2. Rowley, J.: Organising Knowledge: An Introduction to Information Retrieval. Gower, Aldershot, Hants, England (1987)
3. Ganter, B., Wille, R.: Conceptual Scaling. Applications of Combinatorics and Graph Theory in the Biological and Social Sciences, Springer, New York (1989), F. Roberts (ed.), 139–167
4. Wille, R.: Concept Lattices and Conceptual Knowledge Systems. Computers and Mathematics with Applications, vol. 23, 493–522 (1992)
5. Wynar, B., Taylor, A.: Introduction to Cataloging and Classification, 8th ed. Libraries Unlimited, Englewood, Colorado (1992)
6. Berners-Lee, T., Cailliau R., Groff J., Pollerman, B.: World-Wide Web: The Information Universe. CERN, Geneva, Switzerland, (1992)
7. Wilson, E.: Hypertext Libraries: The Automatic Production of Hypertext Documents. Research in Humanities Computing, S. Hockey and N. Ide (eds.), 232-246 (1994)
8. Deutsch, P., Emtage, A.: Publishing Information on the Internet with Anonymous FTP. Bunyip Information Systems Inc., (May 1994)
9. Bowman, M., Dharap, C., Baruah, M., Camargo, B., Potti, S.: A File System for Information Management. Proceedings of the Conference on Intelligent Information Management Systems, (June 1994)
10. Bowman, M., Danzig, P., Hardy, D., Manber, U., Schwartz, M.: Harvest: A Scalable, Customizable Discovery and Access System. technical report CU-CS-732-94, University of Colorado, (July 1994)
11. Mealling, M.: Encoding and Use of Uniform Resource Characteristics. Internet Engineering Task Force (IETF), Internet draft document draft-ietf-uri-urc-spec-00.txt, (July 1994)
12. Kent, R.E., Neuss, C.: Creating a 3D Web Analysis and Visualization Environment. Electronic Proceedings of the Second International World Wide Web Conference (WWW'94), Mosaic and the Web, (October 1994)
13. Bowman, M., Danzig, P., Hardy, D., Manber, U., Schwartz, M.: The Harvest Information Discovery and Access System. Electronic Proceedings of the Second

International World Wide Web Conference (WWW'94), Mosaic and the Web, (October 1994)

14. Kent, R.E.: Enriched Interpretation. Proceedings of the Third International Workshop on Rough Sets and Soft Computing (RSSC'94), (November 1994)
15. Neuss, C., Kent, R.E.: Conceptual Analysis of Resource Meta-information. Electronic Proceedings of the Third International World Wide Web Conference (WWW'95), (April 1995)
16. Kent, R.E., Bowman, M.: Digital Libraries, Conceptual Knowledge Systems, and the Nebula Interface. technical report, Transarc Corporation, Pittsburgh (April 1995) submitted for publication